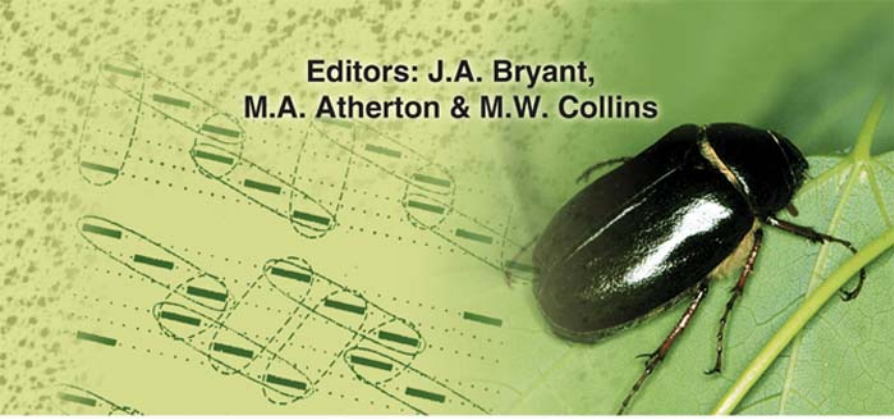
Editors: J.A. Bryant,
M.A. Atherton & M.W. Collins



Design and Information in Biology

From Molecules to Systems

WIT PRESS

Design and Information in Biology

From Molecules to Systems

WIT PRESS

WIT Press publishes leading books in Science and Technology.
Visit our website for the current list of titles.
www.witpress.com

WIT eLibrary

Home of the Transactions of the Wessex Institute, the WIT electronic-library provides the international scientific community with immediate and permanent access to individual papers presented at WIT conferences. Visit the eLibrary at
http://library.witpress.com

Design and Nature

Objectives

Our understanding of the modern world is largely based on an ever increasing volume of scientific knowledge. Engineering designers have at their disposal a vast array of relationships for materials, mechanisms and control, and these laws have been painstakingly assembled by observation of nature. As space activity accustoms us to cosmic scales, and as medicine and biology to the molecular scale of genetics, we have also become more aware of the rich diversity of the structural world around us.

The parallels between human design and nature has inspired many geniuses through history, in engineering, mathematics and other subjects. Much more recently there has been significant research related to design and invention. Even so, current developments in design engineering, and the huge increase in biological knowledge, together with the virtual revolution in computer power and numerical modelling, have all made possible more comprehensive studies of nature. It is these developments which have led to the establishment of this international book series.

Its rationale rests upon the universality of scientific laws in both nature and human design, and on their common material basis. Our organic and inorganic worlds have common energy requirements, which are of great theoretical significance in interpreting our environment.

Individual books in the series cover topics in depth such as mathematics in nature, evolution, natural selection, vision and acoustic systems, robotics, shape in nature, biomimetics, creativity and others. While being rigorous in their approach, the books are structured to appeal to specialist and non-specialist alike.

Series Editor

J.A. Bryant
Dept. of Biological Sciences
University of Exeter
Exeter, EX4 4QG
UK

M.A. Atherton
School of Engineering & Design
Brunel University
Uxbridge
UK

M.W. Collins
School of Engineering & Design
Brunel University
Uxbridge
UK

This page intentionally left blank

Design and Information in Biology

From Molecules to Systems

Editors

J. A. Bryant
University of Exeter, UK

M. A. Atherton
Brunel University, UK

M. W. Collins
Brunel University, UK

WIT PRESS Southampton, Boston

J. A. Bryant
University of Exeter, UK

M. A. Atherton
Brunel University, UK

M. W. Collins
Brunel University, UK

Published by

WIT Press
Ashurst Lodge, Ashurst, Southampton, SO40 7AA, UK
Tel: 44 (0) 238 029 3223; Fax: 44 (0) 238 029 2853
E-Mail: witpress@witpress.com
http://www.witpress.com

For USA, Canada and Mexico

WIT Press
25 Bridge Street, Billerica, MA 01821, USA
Tel: 978 667 5841; Fax: 978 667 7582
E-Mail: infousa@witpress.com
http://www.witpress.com

British Library Cataloguing-in-Publication Data

A Catalogue record for this book is available
from the British Library

ISBN: 978-1-85312-853-0
SET ISBN: 978-1-85312-854-7
ISSN: 1478-0585

Library of Congress Catalog Card Number: 2002111318

Printed in Great Britain by Athenaeum Press Ltd., Gateshead

Dedication

Dr. Mark J. Randall
born 6th April 1971
died 21st September 2000

This book is dedicated to the memory of Mark J. Randall, the author of the chapter entitled 'Insect Observations and Hexapod Designs'.

Shortly before his untimely death, Dr. Randall enthusiastically agreed to provide a chapter on walking mechanisms, based on nature, at a Sussex summer garden party in 2000 where he met one of the editors. Our plans for a book series in Design & Nature were in their early stages at this time but the standard of Mark's work was so high that we sought to reproduce a chapter from his own book in lieu of the one he was unable to complete. Mark worked on artificial intelligence applied to how robots might 'learn' to traverse difficult terrain, and he was particularly interested in the de-mining of war zones, a concern consistent with his Christian faith.

Mark was born in Pontypool, Wales. He studied at Imperial College, London, gaining a 2.1 honours degree in Theoretical Physics, followed by a PGCE at Magdalen College, Oxford. He obtained his PhD from the University of the West of England in 1999 where he was a Senior Lecturer. He has left behind his wife Emma and three young daughters, Deborah, Anna and Ellena. Emma has kindly allowed us to reproduce the material of the chapter.

We would like, finally, to thank Professor Carlos Brebbia at the Wessex Institute of Technology for suggesting this dedication.

This page intentionally left blank

Contents

Chapter 3

Chapter 4

Chapter 11

Chapter 12

Chapter 13

Design and Nature – Introduction to the Series

Michael W. Collins
Brunel University, UK.

Prologue

almost a miracle [Cecil Lewis, 1, p. 126]

almost miraculously [Stuart Kauffman, 2, p. 25]

'It was a beautiful evening' wrote Cecil Lewis [1] of the day in 1917 when he took a new SE5 on a test flight. 'At ten thousand feet the view was immense, England quartered on its northern perimeter at twenty two thousand feet, Kent was below me …… for a second the amazing adventure of flight overwhelmed me. Nothing between me and oblivion but a pair of light linen-covered wings and the roar of a 200-hp engine! …… It was a triumph of human intelligence and skill – almost a miracle' (See Plate 1).

Cecil Lewis was only 19 years old at the time, having left the English public school Oundle, in order to join the Royal Flying Corps in the First World War.

Almost 40 years later, in happier times than those of Cecil Lewis, as another ex-schoolboy I was 'filling in time' with a Student Apprenticeship before going to University. My very first job was as 'D.O. Librarian' in an aeronautical engineering drawing office. The circumstances may have been prosaic, but one feature always intrigued me. At the apex of the very large pyramid, at the top of every document distribution list, was the Chief Designer. Of course, I never met him or even saw him, but to me his title expressed the fount of authority, intelligence and creativity, the *producer* of 'almost miracles' for the 1950's.

'Almost miracles' mean different things to different people. Another 40 years brings us to a new millennium, to Stuart Kauffman [2] writing in 2000. Kauffman, a highly regarded American biologist 'is a founding member of the Santa Fe Institute, the leading centre for the emerging sciences of complexity' [2, cover blurb]. In discussing DNA symmetry and replication, he says [2, page 25]: "It seems to most biologists that this beautiful double helix aperiodic structure is almost miraculously pre-fitted by chemistry and God for the task of being the master molecule of life. If so, then the origin of life must be based on some form of a double-stranded aperiodic solid" (see Plate II). Yes, Kauffman is in the heady business of studying life starting 'from non-life here, or on Mars'.

We have reflected Cecil Lewis's and Kauffman's near miracles in Plates I and II. In the case of Cecil Lewis he was still in the first flush of man's ability to fly. Not for him the necessity of filing a flight plan. Like the *natural* fliers, the birds, he could move at will in three-dimensional space.

However, even then, he could far out-fly them, whether in speed or in height. Stuart Kauffman, however, moves about in multi-dimensional space. *His* is a 'fitness landscape in … thirteen-dimensional parameter space' [2, p. 70].

We do need, at the same time, our sense of wonder to be well informed. Of course, Cecil Lewis's near miracle has been totally replaced, by Jet Propulsion, by travel to the moon, and now by planetary exploration. In the same way, while the thirteen-dimensional space of Kauffman may well impress many of his readers, and the eleven-dimensional space of Stephen Hawking [3] was obviously expected to impress the average UK Daily Telegraph readers, in engineering terms this is a standard practice. Two of the Series Editors [MAA and MWC] started to consider [4] the problem of *visualisation* of complex data. This included reference to the optimisation of nuclear power station design [the UK Magnox system], which used a contour-tracking procedure focusing on 30 major parameters out of about 100 parameters in total [Russ Lewis, 5].

We conclude this prologue with the realisation that nature, nature's laws and the use of nature's laws in human design all have the capacity to enlighten, inform and inspire us. This series will have achieved its end if it demonstrates only a small part of that capacity.

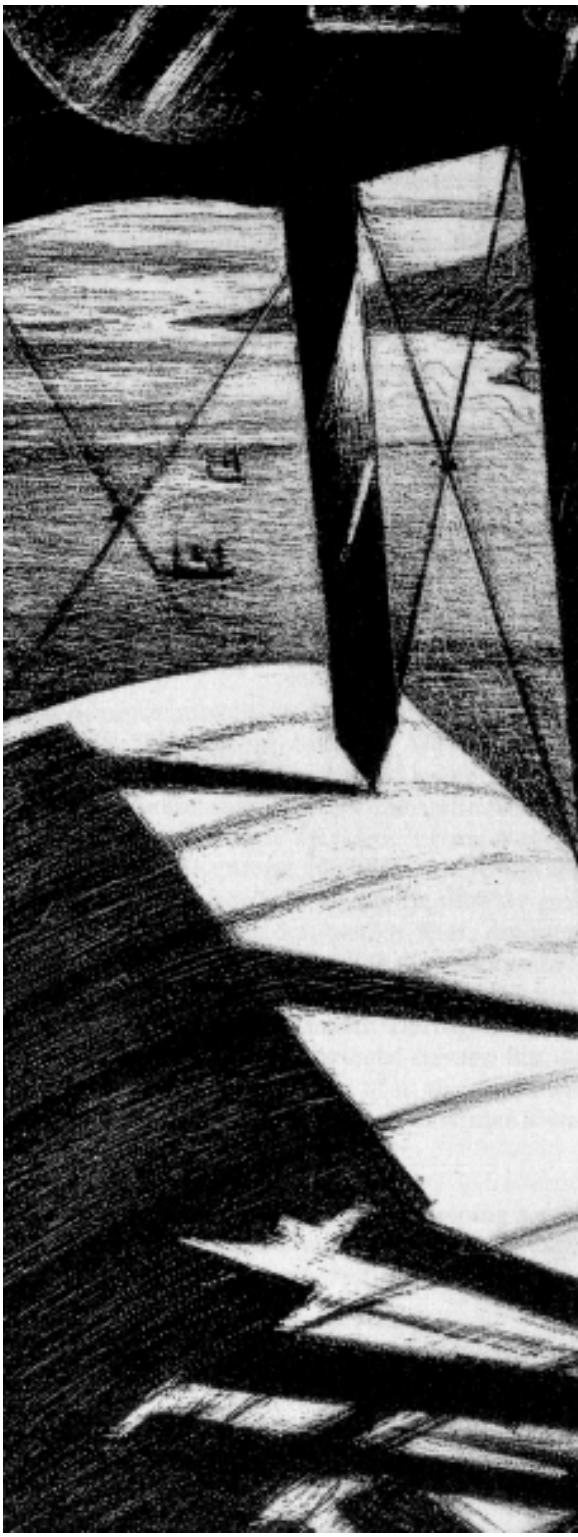


Plate I: 'It was a triumph of human intelligence and skill - almost a miracle'. 'View from an aeroplane' [1, p182-183] (Reproduced by permission of the Victoria and Albert Museum Picture Library).

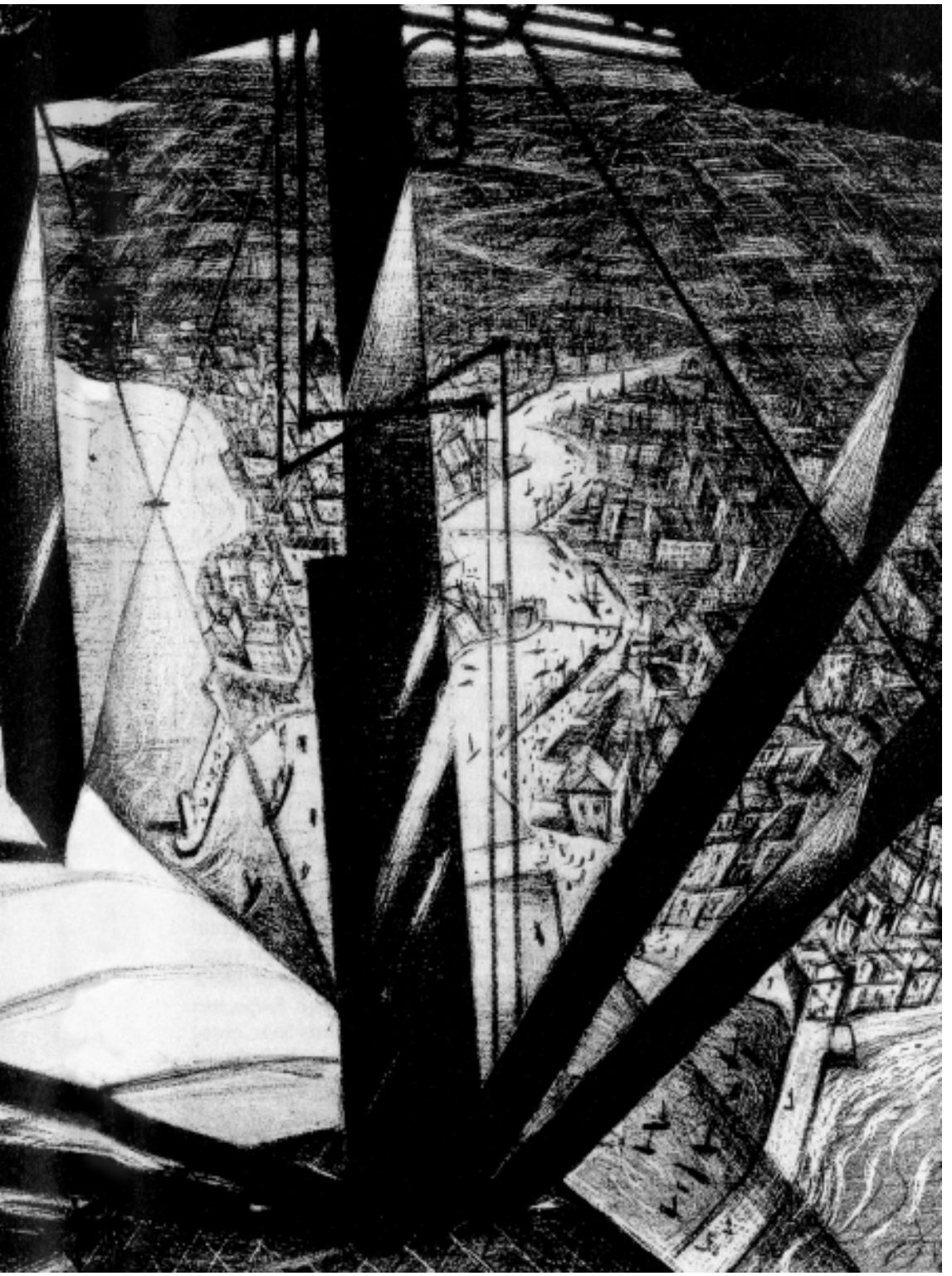


Plate II: 'This beautiful double helix.....structure is almost miraculously pre-fitted....'

DNA is a double helix

Plate II: Watson and Crick proposed that DNA is a double-helical molecule. The helical bands represent the two sugar-phosphate chains, with pairs of bases forming horizontal connections between the chains. The two chains run in opposite directions; biochemists can now pinpoint the position of every atom in a DNA macromolecule.

(Reproduced by permission from *Life. Volume 1 the Cell and Heredity*. 4th Edition by W.K. Purves, G.H. Orians and H.C. Heller, p246, Sinauer Associates, W.H. Freeman & Co.)

Research Field of John Bryant, Series Co-Editor

John Bryant's research is mainly concerned with the initiation of DNA replication - the very start of the process of copying the genome. It is far more complicated than we envisaged even ten years ago.....indeed it has a beautiful and almost awe-inspiring complexity. Each stage is tightly regulated so as to ensure that the cell only duplicates its genome at appropriate times. As we understand more about these control mechanisms we can only wonder at, and about, the evolutionary processes through which they developed.

Nature and engineering

The beavers have practised civil engineering since they became a species
[Eric Laithwaite, 6, p. 231]

Intellectually, the engineer and the artist are not far apart [Michael French, 7, p. 179]

The subject area of our series has great public interest and popularity, if we take the increasing number of publications as evidence. But this needs clarifying. Like Eric Laithwaite having to make a choice at Grammar School [6, p. xi] we might be forgiven for supposing that *our* subject is either biology or physics. On thinking more carefully, we could define our subject as the commonality of the laws of physics, in the natural [biological] and man-made [engineered] worlds. This is nearer the truth.

In the event, Eric Laithwaite chose physics. He went on to become a noted engineer and inventor, being awarded, in 1966, the Royal Society S.G. Brown Medal for Invention. So, for *him*, the beavers were engineers, not scientists.

In the same way, Michael French compares biologists, not with physicists, but with engineers and architects [7, pp. 1–2]. His book, like Laithwaites', is engineering-oriented – 'about design for function, and invention' [7, p. xvii].

So, despite so many of the recent publications being by biologists and physicists, we have chosen two engineers to start our Introduction. In fact, their approach represents a relatively new exploitation of the laws of physics, and materials science, as used in the biological design of living organisms. This points us in the direction of 'biomimetics' which is a recent concept involving the application of biological materials in engineered systems [p. xvii, Vol. 4 of this Series].

Laithwaite and French raise other issues. The first is noticeable by its absence. Those readers whose discipline is chemistry or chemical engineering might wonder if the subject has been 'air-brushed out'. Of course not – if no chemistry, then there is no DNA, no design in nature. We have already quoted Kauffman in this regard.

The next, lightly touched on by French [7, p. 235], as also by Kauffman [2, p. 24], is the question of what is meant by 'beauty'. While French strictly connects it to function in design, we will connect it to art in general, and find it is an integral part of our overall study. As French implies, the engineer and the artist are good friends.

The final issue is the question of mathematics. Whereas French [7, p. 291] rather pejoratively quotes from Bondi that 'mathematicians are not particularly good at thinking … good rather at avoiding thought', Laithwaite is obviously fascinated by the whole thing. For instance, he, like me, is highly intrigued [who isn't] by the identity.

$$e^{i\pi} = -1$$

In the same vein, he deals in some detail with the topics of 'ideal shape' in the form of the golden section [6, pp. 199–202] of Fibonacci numbers, and of helices in plants. He points out that the logarithmic spiral [6, pp. 201–202] retains its shape with growth, coinciding with French's reference to gnomonic shell growth [7, p. 273].

So then, two engineers have introduced our discussion. We now have to explain what *we* mean by the word '*design*'.

Design in the mainstream

The buttercups and the locomotive show evidence of design
[Michael French, 7, p. 1]

I am, in fact, not so much concerned with origins or reasons as with relations or resemblances
[Theodore Cook, 8, p. 4]

We can best describe our use of the word 'design' by the acronym *wysiwyg* – what you see is what you get. Our ambition is to explore fully the richness of the 'design of the buttercup' and the comparison of the designs of nature and engineering, all in the same spirit of Michael French. We shall avoid all issues like 'despite there being 'evidence of design' we do not believe …' on the one hand, or 'because there is evidence of design, we therefore believe …' on the other. The point has been put more elegantly by Theodore Cook, as long ago as 1914 [8, above].

So, we do not, as does Richard Dawkins, use the expression *designoid*. In 'Climbing Mount Improbable' [9, p. 4], he addresses this very point. 'Designoid objects look designed, so much so that some people – probably, alas, most people – think that they are designed'. So he uses designoid because of his antipathy to theism – 'no sane creator … would have conceived on his drawing board' he says on p. 121 [9]. In our use of the word design, however, we retain Richard Dawkins' friendship, with his pitcher plant giving 'every appearance of being excellently well designed' [9, p. 9], and his approbation of engineers – 'often the people best qualified to analyse how animal and plant bodies work' [9, p. 16].

However, in using the word design, neither do we mean *conscious design*, *intelligent design*, *[intelligent] design or [ ] design* … merely design.

Typical use of these explanations is given as follows, with the understanding that 'conscious design' is rather an archaic description:

i.	Conscious design	[Cook, 8, p. 4], [Ruse, 10, p. 44]
ii.	Intelligent design	[Miller, 11, p. 92], [Ruse, 10, p. 120]
iii.	[Intelligent] design	[Miller, 11, pp. 93, 126], [Ruse, 10, p121]
iv.	[ ] design	[Behe, 12, p209], [Dembski, 13, Title]

The last-mentioned author, William Dembski has, sadly, suffered for his beliefs, as explained in 'The Lynching of Bill Dembski' [14]. Nevertheless, in fairness, Dembski separates out the ideas of 'design' and 'designer', as this extended quote makes clear:

> 'Thus, in practice, to infer design is typically to end up with a 'designer' in the classical sense. *Nevertheless, it is useful to separate* [MWC's italics] design from the theories of intelligence and intelligent agency' [13, p. 36].

While the use of the word 'design' here may not be coincidental with that of Dembski, yet the act of separation is crucial, and consistent with the rationale for this Series. By using *wysiwyg* we are trying to retain the friendship of both Dawkins and Dembski and, further, to retain and parallel their common enthusiasm and commitment. In the Series, then, we seek to stay in the mainstream of all aspects of design in the natural and man-made worlds, stressing commonality rather than controversy and reconciliation of differences rather than their sharpening. In that spirit, where necessary, current controversies will be openly discussed and separate issues carefully identified.

Even this brief discussion has shown that the concept of 'design' is both subtle and wide-ranging in its connotations. We now address three specific aspects which are sometimes ignored or even avoided, namely, *mathematics*, *thermodynamics* and *history*.

Mathematics

We like to think mathematics was discovered, not invented
Prof. Tim Pedley, verbal, Salford, 1998

The universe appears to have been designed by a pure mathematician
[James Jeans, 15, p. 137]
quoted in [Paul Davies, 16, p. 202]

Now while the commonality of scientific laws in the natural world is generally accepted, the fact that the world is also mathematically *oriented* is less well understood. Of course, the concept of mathematics being somehow 'built in' to nature's structure is highly significant in terms of our rationale – nature's designs being parallel to man-made designs. Paul Davies expressed this concept in various telling phrases. In 'The Mind of God' we read '… all known fundamental laws are found to be mathematical in form' [16, p. 84]. 'To the scientist, mathematics … is also, astonishingly, the language of nature itself' [p. 16, 93], and as the heading for Figure 10 [p. 109] 'The laws of physics and computable mathematics may form a unique closed cycle of existence'.

In fairness, it should be added, as does Davies, that this approach is not universally accepted, and mathematicians have 'two broadly opposed schools of thought' [16, p. 141]. In the chapter on mathematics in nature' in *this* Volume the issue is dealt with more fully. However, the point we make here is that the overall detailed study of mathematics is essential for our rationale, which cannot be done in more general single-authored books. Paul Davies himself [16, p. 16] starts the reader with 'no previous knowledge of mathematics or physics is necessary'. Philip Ball, in his beautiful exposition of pattern formation in nature, likewise, restricts the mathematical treatment – 'I will not need to use in this book' (he says [17, p. 14]) 'any more mathematics than can be expressed in words rather than in abstruse equations'. Despite this restriction, however, Ball eulogizes mathematics – 'the natural language of pattern and form is mathematics … mathematics has its own very profound beauty … mathematics is perfectly able to produce and describe structures of immense complexity and subtlety ' [17, pp. 10–11].

The conclusion is straightforward – mathematics is an essential part of the design 'spectrum'.

Thermodynamics

The second law, like the first, is an expression of the observed behaviour of finite systems
[Gordon Rogers and Yon Mayhew, 18, p. 809]

Thus the second law is a statistical law in statistical mechanics
[Stuart Kauffman, 2, p. 86]

In seeking to understand thermodynamics there is not so much an obstacle to be surmounted, as ditches to be avoided. This is because thermodynamics uses concepts in common English use like 'energy', 'work', 'heat' and 'temperature', and because the First Law is an expression of the well-accepted 'conservation of energy' principle. However, these concepts are very closely defined in thermodynamics, and it is essential to understand their definitions. When we reach the Second Law, the problem is all too clear. What does entropy *really* mean? Why do different statements of

the Law look completely different? So an 'amateur' understanding of thermodynamics can lead to an absence of appreciation of the Zeroth Law [to do with equilibrium and temperature] an erroneous confidence in First Law issues, and greater or lesser confusion regarding the Second Law! These are ditches indeed.

The other key aspect of thermodynamics is that it is part of the warp and weft of our industrial society. It was through the French engineer Carnot's brilliant perceptions, leading to the Second Law, the procedures for optimising work-producing heat engine design became clear. The same Law, with its stated necessity for heat rejection and reversibility, was the explanation of what otherwise looked like rather low heat engine efficiencies. In fact, essentially as a consequence of the Second Law, best practice power station efficiencies were of the order of 30% over a long period of time. As a major consequence of the enormous consumption of fossil fuels [coal and oil for example] in those power stations, and including internal combustion engines, carbon dioxide concentration in the atmosphere has increased dramatically. Over the two centuries 1800–2000 the increase has been some 28%, with approximately half that figure occurring since 1960. This is shown by Fig. 3.3 of John Houghton [19, p. 31]. Such is a major part of the background to the Greenhouse effect.

Carnot perceived that a crucial factor in achieving higher efficiencies was for the heating source to be at the *highest possible temperature*, which led in its turn to the definition of the Absolute Temperature Scale by the British engineer, Lord Kelvin.

It was then the German physicist Clausius who defined entropy – 'a new physical quantity as fundamental and universal as energy' [Kondepudi and Prigogine, 20, p. 78]. It was not just the heat that was important, but the *heat modified by the absolute temperature*, the entropy, that was needed. As a consequence, quantitatively low values of entropy are 'good', and perhaps this has led to conceptual difficulties. Similarly, entropy increases are caused by the individual processes in the heat engine operation [irreversibilities]. Finally, the Austrian physicist Boltzmann developed a theory of molecular statistics and entropy, leading to the association of entropy with *disorder* [20, p. xii]. Altogether then non-scientific [and even scientific and engineering] readers might be forgiven for viewing entropy as a sort of 'spanner in the thermodynamic works' – to be kept as low as possible.

Now it is not fully appreciated that the Laws of Thermodynamics are *empirical* – so [write Rogers and Mayhew] 'the Second Law, like the First, is an expression of … observed behaviour'. That empirical prevalence extends to the statistical mechanics interpretation – 'the macro state is a collection of microstates … the Second Law can be reformulated in its famous statistical mechanics incarnation' [Kauffman, 2, p. 86]. Post World War II, Shannon's information theory, has caused entropy to be associated formally with information. 'The conclusion we are led to' [Paul Davies, 21, p. 39] 'is that the universe came stocked with information, or negative entropy, from the word go'. Incidentally, our 'forgiven' readers might feel well justified by the expression negative entropy!

So much for the classical past of thermodynamics. Davies's quote points us to a new look at the subject. *What we are now seeing is an almost overwhelming desire to systematise the application of thermodynamics to biology.*

... vast amounts of entropy can be gained through the gravitational contraction of diffuse gas into stars ... we are still living off this store of low entropy [Roger Penrose, 22, p. 417].

... far from equilibrium states can lose their stability and evolve to one of the many states available to the system ... we see a probabilistic Nature that generates new organised structure, a Nature that can create life itself [Dilip Kondepudi and Ilya Prigogine, 20, p. 409].

The sequence of the application of thermodynamics to biology can be traced back to Erwin Schrödinger's lectures given at Trinity College, Dublin, Ireland, at the height of the Second World

War, currently published as 'What is Life?' [23a, 23b]. In the chapter 'Order, Disorder and Entropy' Schrödinger postulates the following sequence: that living matter avoids the decay to equilibrium [or maximum entropy] by feeding on negative entropy from the environment, that is by 'continually sucking orderliness from its environment', and that the plants which form the ultimate source of this orderliness, themselves 'have the most powerful supply of negative entropy in the sunlight' [23a, pp. 67–75].

To take things further, we turn from the more readily available Reference 23a, to 23b, where Roger Penrose's original Foreword has evolved into a substantial Introduction. This latter Introduction is an important source in itself as it takes up Schrodinger's postulation of the sun's negentropic effect. Using Penrose's own words, [23b, p. xx]: the Sun is not just an energy source, but … a very hot spot in an otherwise dark sky … the energy comes to us in a low-entropy form … and we return it all in a high entropy form to the cold background sky. Where does this entropy imbalance come from? … the Sun has condensed from a previous uniform distribution of materials by gravitational contraction. We trace this uniformity … to the Big Bang … the extraordinary uniformity of the Big Bang … is ultimately responsible for the entropy imbalance that gives us our Second Law of Thermodynamics and upon which all life depends. So, too, we repeat Davies [21, p. 39]' as 'a kind of converse to chaos theory'.

> *I regard the concept of 'gnergy' as one of the most important results of my theoretical investigations in biology over the past two decades*
>
> [Sungchal Ji, 25, p. 152]

> *Such a law could be my hoped-for fourth law of thermodynamics for open self-constructing systems*
>
> [Stuart Kauffman, 2, p. 84]

We pass rapidly on to Sungchal Ji, with the proposed concept of 'gnergy' encompassing both energy and information, and to Kauffman with his hoped-for Fourth Law of Thermodynamics. At least they cannot be accused of lack of ambition! Ji's rather beautiful graphical interpretation of the evolutions of density and information since the Big Bang [25, p. 156] is reproduced here, as Plate III. [In doing so, however, it may be noticed that Ji's zero initial information density is hardly consistent with Davies' initial stock of information. This point will be addressed in the chapter on thermodynamics in Volume 2 of the Series]. Eric Chaisson's more concise research paper approach [26] should be noted, as it elegantly combines and quantifies some of the key issues raised by both Ji and Kauffman. It forms a nice introduction to the subject area.

We are about to bury our thermodynamics 'bone'. However, it must be appreciated that other 'dogs' still prefer non-thermodynamics 'bones', for example Stephen Boyden [27] and Ken Wilber [28]. The latter's ambitious 'A Theory of Everything' is sub-titled 'An Integral Vision for Business, Politics and Spirituality'. In his Note to the Reader he makes what to him is a conclusive remark about the 'second law of thermodynamics telling us that in the real world disorder always increases. Yet simple observation tells us that in the real world, life creates order everywhere: the universe is winding up, not down' [28, p. x]. For readers who, like me, cannot put their 'bones' down, this statement cannot be allowed to rest, and represents another issue for Thermodynamics in Volume 2. However, my comment is not meant to be pejorative. Ken Wilber seeks, as do so many writing in this subject area, a mastery almost painful to appreciate!

The final point here is the most interesting of all, namely that the origin of life remains a question. 'How this happened we don't know' said Stephen Hawking recently [29, p. 161]. Somewhat differently, Ilya Prigogine some ten years ago [30, p. 24] – 'we are still far from a detailed

explanation of the origins of life, notwithstanding we begin to see the type of science which is necessary … mechanisms which lead from the laws of chemistry to "information". However, where there's a bone there's a dog, [if the reader will forgive this final use of the metaphor] and in this case our dog is Michael Conrad. Conrad's essential thesis contrasts with that of Schrödinger [31, p. 178], and is rather that – to quote the Abstract [31, p. 177] – 'the non-linear self-organising dynamics of biological systems are *inherent* [my italics] in any … theory … requirements of both quantum mechanics and general relativity'. Conrad's line [ten of the twenty four references in 31 are by himself as sole author] is termed the fluctuon model, and is particularly interesting in relating to 'nanobiological phenomena and that might be detected through nanobiological techniques'. Stuart Kauffman [2, Chapter 10] similarly surveys quantum mechanics and general relativity, but more in the nature of questioning than Conrad's tighter theorising ….

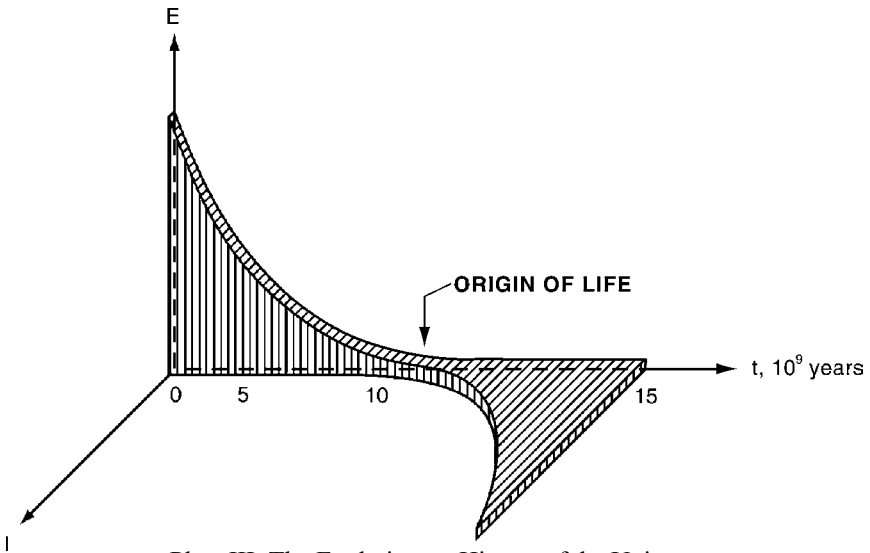


Plate III: The Evolutionary History of the Universe.

In this graph, E and I represent energy and information densities. t is time, on an approximately logarithmic scale, with the origin an estimated 15 billion years ago at the Big Bang. The substantial increase in I occurs following biological 'emergence of the first self-replicating systems … about 3 billion years ago' [after Sungchal Ji].

History

'This concept of an ideal, perfect form behind the messy particulars of reality is one that is generally attributed to Plato'

[Philip Ball, 17, p. 11]

'Leonardo da Vinci was my childhood hero, and he remains one of the few great geniuses of history'

[Michael White, 32, p. xi]

'The only scientific book I read that summer was Charles Darwin's 'The Origin of Species' ... but we do still read Darwin'

[Kenneth Miller, 11, p. 6]

Having sought to show that a fuller understanding of 'Design in Nature' needs to be both mathematically and thermodynamically oriented, we will now point out its historical aspects. We will focus on thee principal characters – Plato, Leonardo da Vinci and Darwin. It is not so easy to give reasons for the choice of these three, but I believe that they represent timeless flashes of genius. Somewhat unexpectedly, they can be viewed in the context of *engineering design*.

So, Plato is associated with one of the key aspects of design, namely form [cf quote by Ball above]. Leonardo epitomizes the ideal of the engineering designer, namely a 'universal man' at home in any branch of knowledge and able to conceptualise almost limitlessly. So we read [33, p. 488] … 'Italian painter, sculptor, architect, engineer and scientist … of immensely inventive and enquiring mind, studying aspects of the natural world from anatomy to aerodynamics'. Finally, Darwin can be associated with the idea of progress and adaptation with time [namely, evolution]. It is difficult to overemphasise this, the point being made explicitly in the *titles* of two recent books. Michael French's [7] title is 'Invention and Evolution'. Design in Nature and Engineering'. Similarly, Norman Crowe on architecture [34] 'Nature and the idea of a man-made world. An Investigation into the Evolutionary Roots of Form and Order in the Built Environment'. These are but two examples. We mentioned flashes of genius. These flashes also possess mathematical connotations. ...*Plato esteemed the science of numbers highly* … [David Smith, 35, p. 89].

In that Plato postulated transcendent [non-earthly] form, he must have been close in approach to the multi-dimensional character of the studies we have already discussed in our Prologue. Platonism *per se* is dealt with at some length by Roger Penrose [22, pp. 146–151] whose 'sympathies lie strongly with Platonistic view that mathematics truth is absolute, external and eternal …'. Paul Davies [16, p. 145] carries Penrose's sympathies forward as … 'Many physicists share his Platonic vision of mathematics'.

'Norway builds Da Vinci's 500 year-old bridge ... it conformed with the laws of mathematics'

[Roger Boyes, 36]

Turning to Leonardo, Michael White freely admits his hero's deficiency in this area. And yet, despite Leonardo's being 'barely competent' [32, p. 152] in mathematics and reliant on Pacioli ['he gained a good deal from Pacioli [32, p. 153]], he designed better than he knew. So we have the Norwegian artist Veljorn Sand, who was the persistent catalyst [it took him 5 years] to secure funding for Leonardo's design, paying Leonardo two compliments, firstly to do with his genius ['when you work with geniuses, you work with eternal forms that never go out of fashion'] and secondly his *implicit* mathematical ability [… 'the design was of lasting beauty because it conformed with the laws of mathematics and geometry … the Mona Lisa of bridges']. To round off Leonardo's relationship with mathematics, he was nothing if not ambitious, and is on record himself as having a very deep commitment. Is it a case of an initial shortcoming being more than subsequently compensated for? So Sherwin Nuland [a surgeon] gives this different picture of Leonardo … 'for Leonardo, mathematics was the ultimate key to the understanding of the nature he scrutinised so carefully … to all of science, including the biology of man' [37, p. 53]. Nuland quotes Leonardo as 'no human investigation can be termed true knowledge if it does not proceed to mathematical demonstration'.

Darwin and Mathematics

> *Inside the sanctum sanctorum they got things done … to Stokes this was 'flimsy to the last degree' … But Huxley pulled off the coup … It was published intact'*
>
> [Adrian Desmond, 38, p. 42]

> *'… Kelvin got very few calculations wrong … here he understandably failed to include the contribution of the heat of radioactivity'*
>
> [Dennis Weaire, 39, p. 61]

> *'Darwin's view of persistent co-evolution remains by and large unconnected with our fundamental physics, even though the evolution of the biosphere is manifestly a physical process. Physicists cannot escape the problem … We will search for constructive laws true of any biosphere. We will found a general biology. And we will be spellbound'*
>
> [Stuart Kauffman, 2, pp. 245, 269]

Finally, Darwin and mathematics. 'The Origin of Species' [41] is essentially, in engineering terms, and experimental report writ large, unaccompanied by mathematical theory. So we have an amusing account as to why Eric Laithwaite chose physics rather than biology. 'Physics seems to be mostly sums, biology mostly essays … my best friend is going to do biology, so I can keep asking him about it and keep in touch that way. That does it … I'll do physics' [6, pp. xi–xii]. Eric Laithwaite's schoolboy choice was a personal reflection of an extremely sharp division in the Royal Society regarding the application of Darwin's work. In fact, Desmond's quote above relates not to a publication of Darwin himself, but an ms submitted on Huxley's suggestion by Kovaleski. The real point here is that the Royal Society's conservative Physical Secretary, George Gabriel Stokes' [38, p. 41] opposed the Kovaleski acceptance because it would make 'speculative Darwinism as axiomatic as Newton's laws' and compromise the rock-like status of knowledge' [38, p. 42]. Now GGS lost, and if Desmond's comment is fair, GGS was spectacularly wrong since Darwin *is* roughly on a par of acceptance with Newton's laws. Not only so, but GGS's close friend Kelvin managed to miscalculate the age of the Earth [second quote above], a scientific *cause celebre* of the time.

GGS is given 'a bad book' by Desmond. In fact, he was an extraordinarily talented and productive physical mathematician and Stokes Summer Schools are run in Ireland, organised by Alastair Wood of Dublin City University [who wrote the parallel section on GGS [39] to that of Kelvin]. I declare a personal interest here. I have an immense regard and affection for Stokes, having worked for decades on numerical studies of convective heat transfer using the Navier-Stokes equations. In fact, Stokes spoke better than he knew, in making an outright comparison [having renamed the word 'speculative'] of Darwinistic [biology] with Newtonian [Physics]. That 1873 assessment was repeated in out anecdotal comment of Laithwaite around 1940, and repeated more tellingly by Kauffman in 2000. Here we remind ourselves that Kauffman is a biologist himself.

Digressing, Darwin was not the only experimentalist to have problems with the Royal Society. Joule [James Prescott Joule 1818–1889] the near-genius who worked assiduously on the equivalence of various forms of energy – notably heat and work – suffered the indignity of having only abstracts of submitted papers published by the Royal Society, on two occasions [J.G. Crowther, 41, pp. 189, 204]. He was young, very young, so despite the setbacks he was still only 32 years old when finally elected to the RS [41, p. 214].

Our final Darwin-related character is Kelvin who, despite the age-of-the-earth *faux pas*, has almost ethereal status of having proposed the Absolute Temperature Scale. In a subsequent volume in this series it is intended to focus on the contributions of [the two Scotsmen] James Clerk Maxwell and Kelvin to thermodynamics, and how this now relates to present day biology –

information, complexity and the genome for example. The latter is epitomised by the recent work of Jeffrey Wicken, the full title of a major publication speaking for itself – 'Evolution, Thermodynamics and Information. Extending the Darwinian programme' [43]. So do the titles of some 17 Journal publications that he references [43, p. 233] for example 'A thermodynamic theory of evolution' in 1980 [44].

In all this, out quiet participant is Darwin himself. Part of his genius, I believe, was his caution, and he let his data collection speak for itself. No mathematics *there*, but an immense sub-surface, iceberg-like, volume of mathematics *underneath*, shown for its worth, as the genome unfolds, and interpreted in terms of information, complexity and Shannon entropy by those such as Kauffman and Wicken.

History summarised

So our three examples of Plato, Leonardo da Vinci and Darwin, have been given a brief introduction. Rather improbably, their genius has been introduced in terms of *engineering design* and *mathematical significance*. Above all, their genius was, and is, timeless. How else could Plato's views on form and mathematics be regarded as relevant two and a half *millennia* later? How else could Leonardo's bridge design be accepted half a millennium later? How else could Darwin's conclusions stand the test of exhaustive and sometimes hostile assessment, lasting for almost a century and a half?

A further aspect of this timelessness, which will be merely stated rather than discussed, is that the Renaissance [epitomised by Leonardo] had as one of its sources the rediscovery of the Greek texts … 'the finding of ancient manuscripts that gave the intellectuals of the Renaissance direct access to classical thought …' [32, p. 39]. So Michael White gives as Appendix 11: 'Leonardo and his place in the History of Science' [32, pp. 339–342], a chronological sequence running from Pythagoras through to Newton ….

Epilogue

Miraculous harmony at Epidaurus [Henri Stierlin, 45, p. 168]

At the commencement of the Prologue to this Introduction, two 'almost miracles' were described. We conclude with a final example going back to 330 BC – to the absolute end of Greek classicism [45, p. 227]. 'Miraculous harmony at Epidaurus' is how Henri Stierlin describes the wonderfully preserved Greek 'theatre set into the hill of Epidaurus' [44, pp. 168–169] – see Plate IV. There are three distinct aspects to this piece of architecture by Polyclitus the Younger. The design has a mathematical basis - including what is now termed the Golden Section and the Fibonacci sequence. Secondly, the harmony spoken of by Stierlin is a consequence of the theatre's 'symmetry' - a subtle technical quality originating in Greek ideas of form. Lastly, the combination of what we now call 'the built environment' with its natural environment has a timeless aesthetic attractiveness. In fact, Plate IV is reproduced not from the reference we have discussed but a Greek Tourist Organisation advertisment.

In concluding, our introduction has covered an almost impossible range of disciplines, but it is only such a range that can possibly do justice to the theme of design in nature. If 'we' is broadened to comprise editors, contributors and publishers, we want to share our sense of inspiration of design in the natural world and man-made worlds that our three authors of near miracles, Cecil Lewis, Stuart Kauffman and Henri Stierlin have epitomised.

Plate IV: 'Miraculous harmony at Epidaurus'.

(See page xiv of Optimisation Mechanics in Nature): 'Around the orchestra, the shell-like theatre set into the hill fans out like a radial structure, whose concentric rows of seating are all focused on the stage where the dramatic action would unfold. With its diameter of 120m., the theatre of Epidaurus is one of the finest semi-circular buildings of Antiquity. Its design, the work of Polyclitus the Younger, according to Pausanias, dates from the end of the fourth century B.C. It is based on a series of mathematical principles and proportions, such as the Golden Section and the so-called Fibonacci Sequence. Its harmony is thus the result of a symmetria in the real sense of the term' [45, p168].
(Reproduced by permission of the Greek National Tourism Organisation).

References

[1] Lewis, C., *Sagittarius Rising*, 3rd Edition, The Folio Society: London, 1998.
[2] Kauffman, S.A., *Investigations*, Oxford, 2000.
[3] Hawking, S., *Why we need 11 dimensions*. Highlighted paragraph in 'I believe in a 'brane' new world', extract from Ref. 29. Daily Telegraph, p. 20, 31st October 2000.
[4] Atherton, M.A., Piva, S., Barrozi, G.S. & Collins, M.W., Enhanced visualization of complex thermo fluid data: horizontal combined convection cases. *Proc. 18th National Conference on Heat Transfer*, Eds. A. Nero, G. Dubini & F. Ingoli, UIT [Italian Union of Thermo fluid dynamics], pp. 243–257, 2000.
[5] Lewis, R.T.V., *Reactor Performance and Optimization*. English Electric Company [now Marconi] Internal Document, 1960.
[6] Laithwaite, E., *An Inventor in the Garden of Eden*. Cambridge, 1994.
[7] French, M., *Invention and Evolution. Design in Nature and Engineering*. 2nd Edition, Cambridge, 1994.
[8] Cook, T.A., *The Curves of Life*. Reproduced from original Constable edition, 1914, Dover, 1979.
[9] Dawkins, R., *Climbing Mount Improbable*. Penguin, 1996.
[10] Ruse, M., *Can a Darwinian be a Christian*. Cambridge, 2001.
[11] Miller, K.R., *Finding Darwin's God*. Cliff Street Books [Harper Collins], 1999.
[12] Behe, M.J., *Darwin's Black Box*. Touchstone [Simon & Schuster], 1998.
[13] Dembski, W.A., *The Design Inference*. Cambridge, 1998.
[14] Heeren, F., *The Lynching of Bill Dembski*, The American Spectator, November 2000.
[15] Jeans, J., *The Mysterious Universe*, Cambridge, 1931.
[16] Davies, P., *The Mind of God*, Penguin, 1993.
[17] Ball, P., *The Self-Made Tapestry*, Oxford, 1999.
[18] Rogers, G. & Mayhew, Y., *Engineering Thermodynamics, Work and Heat Transfer*, 4th Edition, Prentice Hall, 1992.
[19] Houghton, J., *Global Warming*, Lion, 1994.
[20] Kondepudi, D. & Prigogine, I., *Modern Thermodynamics*, Wiley, 1998.
[21] Davies, P., *The Fifth Miracle*, Penguin, 1999.
[22] Penrose, R., *The Emperor's New Mind*, Oxford, 1989/1999.
[23a] Schrödinger, E., *What is Life?* with *Mind and Matter and Autobiographic Sketches*, and a Foreword by R. Penrose, Canto Edition, Cambridge, 1992.
[23b] Schrödinger, E., *What is Life?* and an Introduction by R. Penrose, The Folio Society: London, 2000.
[Note: these are quite distinct publications. The key section *What is Life?* is type-set differently and the page numbers do not correspond.]
[24] Stewart, I., *Does God Play Dice?* 2nd Edition, Penguin, 1997.
[25] Ji, S., *Biocybernetics: A Machine Theory of Biology*, Chapter 1 in: *Molecular Theories of Cell Life and Death*, Ed. S. Ji, Rutgers, 1991.
[26] Chaisson, E., The cosmic environment for the growth of complexity, *Biosystems*, **46**, pp. 13–19, 1998.
[27] Boyden, S., *Western civilization in biological perspective*, Oxford, 1987.
[28] Wilber, K., *A Theory of Everything*, Gateway: Dublin, 2001.
[29] Hawking, S., *The Universe in a Nutshell*, Bantam Press, 2001.

[30] Prigogine, I., *Schrödinger and the Riddle of Life*, Chapter 2 in: *Molecular Theories of Cell Life and Death*, Ed. S. Ji, Rutgers, 1991.
[31] Conrad, M., Origin of life and the underlying physics of the universe, *Biosystems*, **42**, pp. 117–190, 1997.
[32] White, M., *Leonardo*, Little Brown & Co.: London, 2000.
[33] *The Complete Family Encyclopaedia*, Fraser Stewart Book Wholesale Ltd., Helicon Publishing: London, 1992.
[34] Crowe, N., *Nature and the Idea of a Man-Made World*, MIT Press: Cambridge MA, USA & London, UK, 1995.
[35] Smith, D., *History of Mathematics*, Volume 1, First published 1923, Dover Edition, New York, 1958.
[36] Boyes, R., *Norway builds Da Vinci's 500-year-old bridge*, The Times [UK Newspaper], London, 1November 2001.
[37] Nuland, S., *Leonardo da Vinci*, Weidenfield & Nicolson, London, 2000.
[38] Desmond, A., *Huxley Evolution's High Priest*, Michael Joseph: London, 200.
[39] Weaire, D., *William Thomson [Lord Kelvin] 1824–1907*, Chapter 8 in: *Creators of Mathematics: the Irish Connection*, Ed. K. Houston, University College, Dublin Press: Ireland, 2000.
[40] Darwin, C., *The Origin of Species*, Wordsworth Classics Edition, Ware, Herefordshire, UK, 2000.
[41] Crowther, J.G., *The British Scientists of the Nineteenth Century*, Volume 1, Allen Lane/ Penguin, Pelican Books, 1940.
[42] Wood, A., *George Gabriel Stokes 1819–1903*, Chapter 5 in: *Creators of Mathematics: the Irish Connection*, Ed. K. Houston, University College, Dublin Press: Ireland, 2000.
[43] Wicken, J.S., *Evolution, Thermodynamics and Information*, Oxford University Press 1987.
[43a] Wicken, J.S., A thermodynamic theory of evolution, *J. Theor. Biol.*, **87**, pp. 9–23, 1980.
[45] Stierlin, H., *Greece from Mycenae to the Parthenon*, Series on Architecture and Design by TASCHEN, Editor-in-Chief A. Taschen, Taschen: Cologne, Germany, 2001.

Preface

The structure of DNA gave to the concept of the gene a physical and chemical meaning by which all its properties can be interpreted. Most important, DNA - right there in the physical facts of its structure - is both autocatalytic and heterocatalytic. That is, genes have the dual function, to dictate the construction of more DNA, identical to themselves, and to dictate the construction of proteins very different from themselves.
Max Perutz, Nobel Laureate, quoted by Horace Freeland Judson [1]

...deoxyribonucleic acid turned out to be a substance of elegance, even beauty. Structure and those dual functions are united in DNA with such ingenious parsimony that one smiles with the delight of perceiving it.
Horace Freeland Judson [1]

This, the second of the two volumes providing the 'holistic introduction' to the whole 'Design and Nature' series, focuses initially on DNA as a starting point for the consideration of the evolution of information and complexity in the natural world. The significance of the dual functions of DNA and the complexity of biomolecules, emphatically revealed by the results from the Human Genome Project, are considered both in the context of the way living things work and in relation to the origin of life.

Despite the obvious complexity, DNA function, cellular behaviour in general and the overall life and activities of living organisms all occur within the parameters imposed by the laws of nature that govern the workings of the universe. This is true at all levels of biological complexity from each individual chemical reaction through specific activities of organisms such as walking or flight to the functioning of whole ecosystems. It is equally true of the activities of humankind: we work within the laws of nature, not just as biological organisms but also when we are acting as engineers, inventors or trying to mimic specific aspects of life such as self-replication or autonomy. Further, consideration of particular fields of human endeavour as living systems may help to improve those systems as we analyse the parallels between the activities of human society and the wider world of biology, especially in relation to design and information.

We have thus covered design and information in biology at several different levels and from a number of perspectives. The book is not a comprehensive coverage of the subject - indeed a comprehensive coverage in one book or even in several would be impossible. Rather, as stated in

---

[1] Freeland, H.F. (1996) *The Eighth Day of Creation: Makers of the Revolution in Biology*, 2nd edition. Cold Spring Harbor Laboratory Press, Plainview, NY.

the Preface to Volume 1, each chapter is a personal flash of illumination from the author or authors and it is very much hoped that these chapters will indeed illuminate for our readers particular facets of the subject with which they are as yet unfamiliar.

We are very grateful to the authors for so willingly agreeing to contribute to this volume, which, for several of them, is unlike any other book to which they have previously contributed. It is their scholarly and thoughtful writing that has enabled us to achieve the flow from molecules to systems that we envisioned when the volume was conceived. We are also indebted to our friends and colleagues at WIT press who have supported and encouraged us in this work and who have looked after the production process so efficiently.

John A. Bryant, University of Exeter, United Kingdom
Mark Atherton, Brunel University, United Kingdom
Michael W. Collins, Brunel University, United Kingdom

Chapter 1

Introduction: Part I – Design and information in biological systems

J. Bryant
School of Biosciences, University of Exeter, Exeter, UK.

Abstract

The term 'design' in biology usually refers to fitness for purpose: is the particular structure or mechanism effective in carrying out its function? This must be seen in the light of evolution by natural selection. Organisms which function better in a given environment than their close relatives and therefore have a greater reproductive success will become more abundant at the expense of less successful individuals. The millions of species of living organisms, from the most simple to the most complex, now present on earth have arisen by this process. Natural selection can only work if a particular advantage exhibited by particular individuals is heritable, i.e. embedded in the 'genetic machinery'. That genetic machinery is based on DNA, whose structure is elegantly fitted to its two functions: (a) carrying the genetic information that regulates the development, growth and functioning of the organism; (b) passing on that information to subsequent generations. Genetic information is carried in the order of deoxyribonucleotides in a DNA molecule. The double helical structure of DNA, coupled with the way that the bases within the deoxyribonucleotides pair specifically between the two chains of the double helix, provides a template mechanism for passing on the information. DNA's role as the regulator of minute-by-minute function is achieved by the copying of specific tracts of DNA (genes) into mRNA and the translation of the code in mRNA to make proteins. In evolution, it is the subtle changes in the order of deoxyribonucleotides (mutations) that generate the heritable variation based on which natural selection works. We are ignorant as to how these mechanisms originated, but it is thought that early in the development of life, the genetic material was RNA, which in addition to carrying genetic information could also mediate a limited range of the functions now performed by proteins.

1 Design, function and elegance

When design is spoken of in the day-to-day world we can discern two separate strands of meaning. The first of these is *fitness for purpose*. Does the object perform effectively the functions that it is meant for? Will a bridge support appropriate loads under all weather conditions? Will a toaster

warm and brown the bread evenly and without burning? Will a raincoat actually prevent the wearer from getting wet? It will be obvious that if any such objects fail to fulfil their function, then we can talk about a design fault and thus, used in this sense, design and function are inexorably linked. However, in day-to-day life, design also incorporates the rather more elusive term 'style' and thus possesses an aesthetic element. A bridge may carry traffic across a river but may also be an eyesore. A raincoat may keep the rain off but may do nothing to enhance the appearance of the wearer. These examples also show us that ideas of style may be transient: ideas of what is acceptable style change. In architecture, for example, the stark modernity of many mid-20th century buildings, where beauty was equated with function, is now largely regarded as ugly and obtrusive. And, of course, the extreme of this transience is seen in fashion where firstly function may be sacrificed for style and secondly what is regarded as 'stylish' one year may be rejected the next. Further, ideas of style may also be personal: what someone likes, another may dislike intensely as seen, for example, in the reactions to the 'ziggurat' student residences at the University of East Anglia (a prize-winning design by Sir Denys Lasdun in the late 1960s; Fig. 1a) or to the starkly modernist style of many of the buildings on the east campus of the University of Illinois at Chicago (Fig. 1b).

Although we have separated the two elements of fitness for purpose and style, they may in fact be linked. When an engineer solves a design problem in a particularly neat way or when an IT expert writes a new program that replaces an older, more cumbersome and less user-friendly version, then we often speak of elegance. Thus our aesthetic sense appreciates the cleverness of the way that a problem has been solved.

This element of elegance – fitness for purpose achieved by a neat and economical mechanism – leads to specific consideration of design in biological systems. In biology, the major implication of the term 'design' is fitness for purpose. The particular structure or system performs the necessary tasks within the life of a particular living organism. However, we also recognise that the structures and systems that have evolved in response to particular problems are often very elegant, indeed to the biologist they may be beautiful. (This emphasises a point that the author has made elsewhere [1, 2]: science is neither value-free nor free from the scientist's own particular set of values. It is entirely legitimate to speak of beauty.) In the author's own field [3], i.e. replication of the genetic material (DNA), the prevention of unscheduled replication is achieved by an array of interacting, subtly regulated and very effective mechanisms providing a system of control which, even after many years of work on the subject, I still find beautiful. Further, these elements of neatness and economy in relation to fitness for purpose have been admired not only by biologists but also by those designing objects for use in human society. Engineers, for example, have turned frequently to nature and some of the results of that are described in this and other volumes in the series. However, for the time being we must return to the specifically biological aspects.

2 Evolution and design

When we observe a particular structure or system in biology we are observing the current results of an ongoing evolutionary process. The diversity of living organisms and their occupation of particular ecological niches has arisen by the process of change driven by a set of mechanisms known as natural selection. In essence, this means that in a given population, the individuals that are more fitted for the particular environment will be more successful than those that are less fitted. In general, success here means reproductive success: individuals that produce more offspring will eventually come to dominate the population, even if the differential is very small. It is specifically in this context that the word 'fit' must be understood. The term 'survival of the fittest'

(a)

(b)

Figure 1: (a) Student residences at the University of East Anglia, Norwich, UK. Downloaded by permission from http://www.uea.ac.uk/slideshow_pages/slideshow1.html. (b) The east campus of the University of Illinois at Chicago, IL, USA. Downloaded by permission from http://www.uic.edu/depts/oar/virtualtour/slideshow/index.html.

has provoked hostile reactions and has been extensively misapplied. Indeed, the term was never used by Darwin but was coined by Herbert Spencer [in his book *Principles of Biology* (1864)] in what was probably a misunderstanding of what natural selection actually involves. Nevertheless, the term is now used frequently in the context of evolution where it means fitness for existence in a particular environment and does not refer directly to the 'health' of the organism. A perfectly

healthy organism may be less fitted for an environment and thus have less reproductive success than another equally healthy organism.

We need to note that the process of evolution has led to the existence of living organisms exhibiting a huge range of levels of complexity, from simple single-celled micro-organisms such as bacteria to very complex multicellular organisms such as mammals. This is not the place to become involved in a discussion of the evolution of complexity. However, we do need to consider design in the light of complexity.

Design is fitness for purpose: the structure or system performs the necessary function and that structure has evolved to be fit for purpose under the pressure of natural selection. A particular individual in which a particular function is performed more efficiently may enjoy a reproductive advantage and thus the 'improved' version that enables more efficient performance of function will be favoured. Some structures or systems occur almost universally across the vast spectrum of living organisms, performing at similar levels of efficiency (i.e. exhibiting similar fitness for purpose) across that range while other systems are developed to very different levels of complexity and efficiency in different types of organism. Thus, for example, the compound eye of the insect is a less efficient visual organ than the mammalian eye. However, within the level of complexity exhibited by the insects, the mechanisms involved in the development and function of the mammalian eye are not possible. But no one can say that the insects are 'unsuccessful': there are more species of insects than of any other group of animals. The complexity of design of many features may thus be related to the complexity reached by a particular group during evolution.

Finally, we need to note that there are many very elegant examples of evolution that involve two or, sometimes, more types of organism. Some of this involves organisms living together in close association; this is called symbiosis. In many cases, the symbiosis is of benefit to both organisms, in which case we use the term 'mutualism'. Usually, in mutualistic relationships, there is close cellular relationship between the two organisms; for example, mycorrhizal fungi form sheaths around the roots of their host plants. In some instances, the tissues of the two organisms are so intimately associated that the combination of the two looks like a completely new species, e.g. when a fungus combines with either a photosynthetic blue–green bacterium or a photosynthetic alga to form a lichen. Interestingly, the study of DNA sequences and gene organisation suggests that lichen-style symbioses have evolved independently at least five times during the long history of life on earth [4]. All such close associations imply a history of co-evolution as the two organisms adapted to become more and more mutually interdependent.

However, co-evolution is not confined to these close relationships: there are also instances in which there is no intermixing or intimate contact at the cell or tissue level. Two examples out of many will suffice. First, many plants are insect-pollinated. In visiting flowers to obtain nectar and/or pollen, the insect in question transfers some pollen to other flowers of the same species thereby bringing about cross-fertilisation. Amongst the orchids, several groups attract pollinating insects because the flowers strongly resemble the females of the pollinating species. In attempting to mate with the flower, the male picks up pollen which will be transferred to the next flower at the next unsuccessful copulation attempt. The insect receives no reward: there is no nectar, the pollen is unavailable to the insect (being contained in structures called pollinia) and, of course, sex is not at all satisfactory. It is hardly surprising that this is called 'deceit pollination'. The second example also concerns plants and insects. Many wild plants accumulate in their leaves chemical compounds that are toxic to predatory insects. The biochemical pathway that leads to the synthesis of these compounds is fit for the purpose of deterring predators. However, some insects have evolved a defence against the plant defences either by the possession of detoxification mechanisms, or, more sophisticated still, by accumulating the plant toxins so that the insects themselves become toxic in turn to their potential predators.

3 Evolution and information

In the discussion of evolution, the concept was introduced that subtle differences between otherwise very similar individuals could lead to differences in reproductive success. However, in order for this difference in reproductive success to be maintained, it must be heritable. Having a selective advantage in one generation but not in the next will not lead to evolution. Now, although natural selection acts on the phenotype (i.e. on the features of the actual living organism resulting from an interaction between genes and the environment), long-term heritable differences must ultimately be based in the genes. It is the genes that carry heritable information. It is therefore helpful, at this point in this introductory chapter, to briefly discuss gene function.

In all cellular life, genes are made of DNA. (There are some classes of virus in which the genetic material is RNA. However, although we may regard this as a throwback to a time in early evolution when the prevalent genetic material may have been RNA, we cannot regard present-day viruses as primitive life forms because they rely on cellular organisms for their own multiplication.) Indeed, the discovery in 1944 that DNA is the almost universal genetic material [5] was the real turning point in biology and led to intense efforts to understand the structure of this vital molecule. These efforts led, of course, to the discovery in 1953 of the double helix [6, 7]. This structure is 'design in nature' at its most elegant. The genetic material needs to fulfil two major functions. First, it must carry the genetic information. Further, that genetic information must be in a form that is readily usable in the life of the organism. Secondly, the genetic information must be passed on faithfully from generation to generation, i.e. must be heritable.

How then does DNA carry genetic information? Although DNA molecules are large, their chemical structure is relatively simple: each single strand of DNA is a chain of nucleotides (to be more specific, deoxyribonucleotides, which are often called, incorrectly, bases; however, it is the components of the deoxyribonucleotides known as bases that provide the differences between the deoxyribonucleotides and whose interactions maintain the double helix.). There are only four types of these, and one of the puzzles raised by the discovery that DNA is the genetic material was how such a structure could contain the information needed for life. Of course, the answer is now well known. Although there are only four deoxyribonucleotides, the length of DNA molecules means that the linear array of these four building blocks can be immense. Further, there are no constraints as to which deoxyribonucleotide is next to another: they can occur in any order. This immediately raises the possibility that a specific linear array of deoxyribonucleotides can be a coded source of information just as a linear array of the 26 letters of the alphabet can be a coded source of information (provided we can read the language in which those 26 letters are deployed). The function of the code in DNA is discussed below. In the meantime, we focus on the second function of the genetic material, passing on information from generation to generation.

In order to be a faithful transmitter of hereditary information from generation to generation, the coded information must be accurately reproduced and this is ensured by the structure of DNA, the double helix. We have already noted that there are no constraints in placing deoxyribonucleotides next to each other along a single DNA strand. However, there are constraints on the deoxyribonucleotides that can exist opposite each other in the two strands of the double helix. This constraint lies in the properties of the bases (as already noted, these are the variable moieties within deoxyribonucleotides) and thus from this point the discussion is framed in terms of bases. The constraint on what base occurs opposite another in the two strands of the double helix is a result of specific base pairing: adenine can pair only with thymine and cytosine can pair only with guanine. The pairing depends on the ability to form hydrogen bonds from particular positions in each base molecule (Fig. 2). Adenine and thymine form two hydrogen bonds between each other, cytosine and guanine three. This means that the position of a particular base in one strand defines

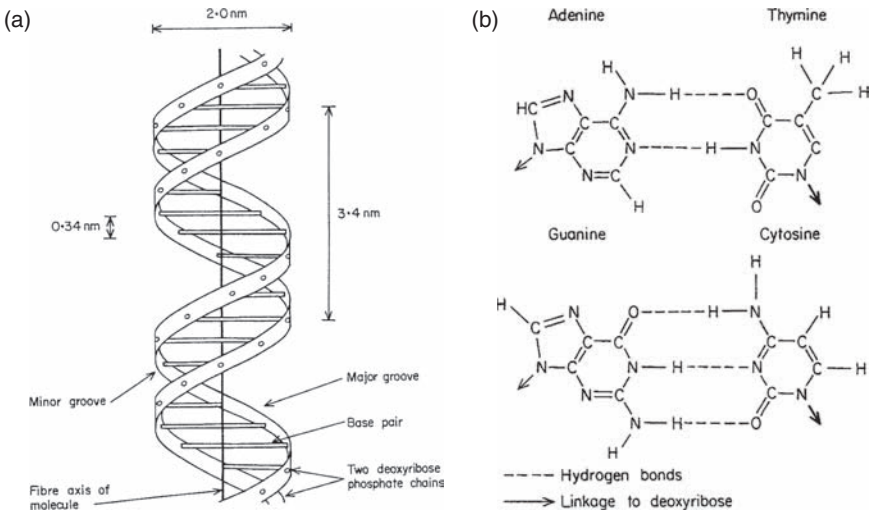


Figure 2: (a) The double helical structure of DNA. (b) Pairing between the bases of DNA. Both figures reproduced by permission from Bryant, J.A. (ed.), *Molecular Aspects of Gene Expression in Plants*, Academic Press: London and New York, 1976.

what base will be opposite it in the other. It can be seen straightaway that this structure preserves the genetic information. When the DNA is replicated, the two strands of the helix separate and each is able to act as a template for the formation of a new strand; the order of bases in the new strand being defined by the order of bases in the template strand. The enzymes (protein molecules that act as biocatalysts) that mediate DNA replication (enzymes mediate all cellular chemical transformations) are thus guided and indeed constrained by the order of bases in the template strands. The pre-existing strands really do provide the information to build the new strands. Thus two double helices now exist where before there was one; in this sense – the specification of the base sequence in the new strand by the base sequence in the old strand – DNA is a self-replicating molecule (but see Section 5).

It is thus apparent that the faithful copying of the genetic material (and, as described in Section 4 below, its function in providing the information for the ongoing life of the organism) is based on a truly elegant structure. But we need to highlight two more features that further illustrate the beauty of this molecule. When Watson and Crick first hit on the idea of a double helix defined by specific base-pairing they modelled both strands of the helix as being in the same orientation – both the same way up. However, the bases did not fit together perfectly in their pairs; the double helix was distorted, under strain and therefore likely to be less thermodynamically stable than it should be. However, if one strand is turned upside down in relation to the other, i.e. if the two strands are *anti-parallel*, then specific base pairing can occur along the full length of the double helix. There is no distortion or strain and the molecule is in the most stable configuration thermodynamically (the configuration of least energy). This inspired guess or wonderful intuition on the part of Watson and Crick was readily confirmed by direct experimentation. The second feature concerns the forces that hold the helix together, namely hydrogen bonds between the two strands and Van der Waal's interactions up and down the molecule. These interactions are strong enough to stabilise the helix but weak enough to allow ready separation of the strands when needed, e.g. for replication. This is a truly elegant molecule with its functions based beautifully in

its properties. Stuart Kauffman, a founder-member of the notable Santa Fe Institute states that the chemistry of DNA, 'this beautiful double helix aperiodic structure' fits it 'almost miraculously … for the task of being the master-molecule of life' [8].

4 Using the information

4.1 Introduction

In order for the coding information in DNA to be used by the cell, it must first be copied or transcribed and then the information in the copy must be translated. The products of translation are proteins, which are the working molecules of cells, carrying out functions from catalytic (as with enzymes) to structural. Thus the code in DNA provides the information for the synthesis of proteins. In this section, the basic mechanisms involved in this process are described (see Fig. 3). In the next chapter, the more complex aspects are discussed in greater detail.

4.2 Transcription

It has already been noted that the formation of specific base pairs is the mechanism by which the genetic information is faithfully copied. It is the same base-pairing mechanism that enables the coding regions of DNA, the genes, to be transcribed into working copies. These working copies are built not with deoxyribonucleotides but with ribonucleotides (Fig. 4). They are thus RNA molecules and are known as messenger RNA (mRNA). Each mRNA molecule is transcribed from one strand of a coding region (gene) of DNA and because of the specific base-pairing involved in transcription, it is complementary to the DNA strand from which it is transcribed.

As with DNA replication, transcription is carried out by enzymes, and a question that immediately arises is how do the transcription enzymes 'know' where to start and stop along the length of a DNA strand that contains many genes. Part of the answer lies in the structure of DNA itself. In addition to the genes which provide information that is transcribed into RNA molecules, the sequence of bases provides other types of information, including sequences 'upstream' of genes

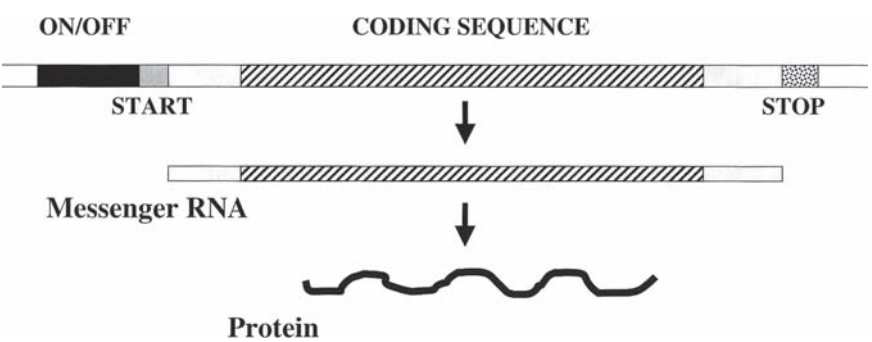


Figure 3: Diagram illustrating the basic mechanisms involved in gene function. 'Upstream' of the gene is a DNA sequence called the promoter which is involved in turning the gene on and off. A gene which is 'on' is copied into a complementary mRNA molecule. The code contained within the sequence of the mRNA is translated, enabling the cell to form the particular protein encoded in the RNA.

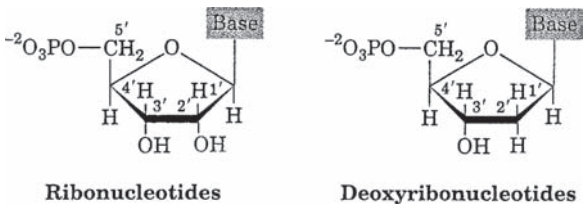


Figure 4: General structures of ribonucleotides (a) and deoxyribonucleotides (b). The two general structures differ only in the absence of oxygen on position 2′ in the sugar moiety. Reproduced by permission from Voet, D., Voet, J.G. & Pratt, C.W., *Fundamentals of Biochemistry*, John Wiley and Sons: New York, 1998.

that enable the transcription enzymes (RNA polymerases) to bind to the DNA. These upstream recognition sequences are known as promoters; immediately 'downstream' of a promoter is a base sequence which marks the point at which transcription starts. Further, at the other end of a gene there is a sequence of bases that marks the point at which the RNA polymerase stops transcribing the gene. Thus, DNA contains, in addition to the sequences that code for proteins, sequences that enable the process of transcription (and, as described in the next chapter, replication) to take place. This is similar in some respects to a computer program that contains within it not only the digital information involved in the program itself but also the information that enables the program to be run. This embedding of different types of function into the base sequence of DNA has been described as the 'many-sidedness' of DNA [9].

4.3 Translation

In the linear array of ribonucleotides in a particular mRNA molecule, there is encoded a recipe to build a particular protein from its individual building blocks, the amino acids. The decoding of this recipe is known as translation. The code is read in groups of three nucleotides, (these triplets are known as codons) each of which specifies the addition of a particular amino acid to the growing protein chain (Fig. 5). Specific base-pairing is also involved in this process: each amino acid is brought into position by a specific carrier molecule called transfer RNA (tRNA). An example will help to clarify this. The three-letter codon that specifies the amino acid methionine is AUG (ribonucleotides containing the bases adenine, uracil and guanine); the tRNA that carries methionine contains the anti-codon sequence UAC (ribonucleotides containing the bases uracil, adenine and cytosine), which is complementary to and therefore can base pair with AUG. It will not have escaped the reader's notice that each tRNA molecule must carry the amino acid that is specified by a particular triplet of bases, otherwise the translation mechanism would not work. This linkage of particular amino acids with particular tRNA species is of course carried out by enzymes (enzymes whose catalytic activity embodies this dual specificity).

The participation of tRNA in the translation of the code during protein synthesis shows that there are other types of RNA in addition to mRNA. These include RNA molecules known as ribosomal RNA (rRNA) which, together with an assemblage of proteins, make up a subcellular organelle called the ribosome. These particles are intimately involved in translation, providing the specific locations at which the mRNA codons base pair temporarily with the tRNA anti-codons during protein synthesis (Fig. 5). Like mRNA, both tRNA and rRNA are transcribed from genes. However, unlike mRNA, these other types of RNA do not embody a code that specifies synthesis of a particular protein. Instead, they participate in the actual mechanism of protein synthesis.

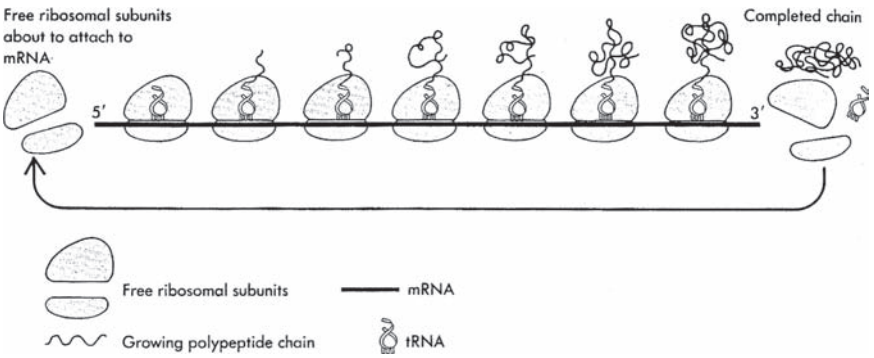


Figure 5: Diagram of the process of protein synthesis – translation of the code in mRNA. Ribosomes engage with the mRNA and move along it (from left to right in this diagram). The amino acids – building blocks for the protein – are brought to the ribosome by tRNA molecules. Each type of tRNA recognises the particular three-base sequence (codon) that specifies the amino acid that the tRNA is carrying. Reproduced by permission from Mathews, C.K. & van Holde, K.E., *Biochemistry*, 2nd edn, Benjamin/Cummins: Menlo Park, CA, 1996.

In the next chapter, we discuss further aspects of DNA replication, gene transcription and protein synthesis, including the roles of yet more types of RNA. In the meantime, we consider the relationship between these basic processes and evolution.

4.4 Genetic information and evolution

It has been emphasised already that natural selection works on the phenotype, the actual living individual, and that the process favours the individuals that are best fitted to a particular environment. This implies that individuals within an interbreeding population (or species) differ from each other. Some of those differences may not be heritable, e.g. those that are caused by direct nutritional or environmental effects on the growth or development of the individual. Other differences are however heritable because they are based in the genes and it is the selection of such differences that is the basic mechanism of evolution. How do genetic differences arise between individuals? Although the process of DNA replication is extremely accurate, it happens on very rare occasions that the wrong deoxyribonucleotide is inserted into the growing DNA strand. This causes a mutation, simply meaning a change (despite the sinister overtones that many people mistakenly attach to the word). Some of these changes may be deleterious or even lethal, others may confer an immediate advantage but most are completely neutral. Further, many of these neutral changes may not cause, in a given environment, any visible differences and are therefore hidden. Thus in the human species, *Homo sapiens*, any two individuals are likely to differ in about one in every thousand base pairs but many of those do not produce discernible differences at the level of the phenotype. So, within a particular species there will be a lot of hidden genetic variation between individuals.

At this point, we need to note that generation of genetic diversity simply by mutation of existing genes would be limited if the number of genes was small: there simply would not be enough spare genetic capacity to carry much genetic change. However, it is clear that evolution involves other types of genetic change in addition to mutation. These include the acquisition of extra DNA sequences so that the amount of DNA in more complex organisms is greater than in bacteria.

This is discussed more completely in the next chapter, but the main points to be made here are, first, that acquiring more DNA in itself provides a genetic difference from the cell or organism without the extra DNA and, secondly, it provides more capacity for carrying genetic mutation, as discussed by Brown [10].

We also need to note that in organisms which reproduce by the coming together of specialised sex cells or gametes, the reproductive process itself will set up new mixes of genes. This is firstly because of a mechanism called recombination which functions in many organisms during the formation of the gametes. Secondly, it is because in organisms where outbreeding is obligatory, two different versions of the organism's genome, one from the female parent and the other from the male parent, provide a new combination of genetic variants in the offspring.

As already noted, some new mutations may confer an immediate advantage in a given environment. However, it is more likely that a change in the environment (using the term in the broadest sense) will lead to a selection of variants that had hitherto been neutral and perhaps hidden. This highlights one of the commonest misconceptions about natural selection and evolution, namely that adaptation to changed environments is thought to have to wait for new mutations. In fact, natural selection generally acts on existing genetic variety. Only in times of rapid change is the generation of genetic variety likely to occur too slowly to allow adaptation. Thus, at present, there is concern that some organisms may not adapt quickly enough to the new conditions imposed by global warming and may therefore be heading for extinction.

5 Information and the origin of life

In the preceding section, the basic mechanisms involved in the minute-by-minute use and in the inheritance of genetic information were discussed. But how did all this start? The minimum requirement for life is an ability to self-replicate and thus the double helical DNA with its built-in template system ensuring faithful replication appears to satisfy that requirement. Furthermore, its ability to carry coded information that is conserved in replication suggests that it is indeed the master molecule of life. However, when the situation is examined more closely, it is obvious that this view, although very reasonable in the light of life as we know it, is more difficult to sustain when the origin of life is considered. The problem may be simply put as follows. The DNA codes for proteins but proteins are needed both to read the code and to replicate it. Without DNA there are no proteins but without proteins there is no replication of DNA and no mechanism for making proteins. We are thus in a loop with no obvious way out. Some have suggested that proteins may have been the original molecules of life but they are not self-replicating molecules. Currently, the most widely accepted view is that the original molecule of life was not DNA but RNA [10]. This view is based on the observations that today there are certain types of RNA molecules that are self-replicating and, further, some types of RNA have limited catalytic (equivalent to enzymic) activity, i.e. activity which in the past may have enabled RNA molecules to self-replicate without the need for proteins. Even if the 'RNA world' hypothesis is accepted (and, of course, there is no *direct* way of disproving or proving it), it is still a long way from a self-replicating RNA molecule to the even the simplest of single-celled organisms that are living today, and it is to the organisms of today that we now return.

6 Wider aspects of information transfer

As we have seen, at the heart of life itself, the DNA contains coded information, information that regulates the minute-by-minute biochemical activity of each cell and which is inherited from

generation to generation. Without this coded information, life as we know it would not exist. However, all living organisms, from the simplest single-celled microbe to the most complex mammal, are also dependent on other forms of information transfer. This happens at various levels, including cell to cell (both in single-celled and multicellular organisms), environment to organism, organ to organ (within a complex organism) and organism to organism. The information transfer is often based on chemical signalling but in particular instances may involve other mechanisms such as light perception or the transmission of electrical and electrochemical signals. Further, the effects of the transferred signals/information may be at any level from the core information molecule, DNA itself, to the whole organism (and, in some instances, to populations). A comprehensive treatment of this is outside the scope of this chapter (but will be the subject of a later volume in this series). Here we simply need to note that living organisms as we know them cannot function without a range information transfer and signal transduction mechanisms.

References

[1] Bryant, J. & Searle, J., *Life in Our Hands*, Inter-Varsity Press: Leicester, 2004.
[2] Bryant, J., Baggott la Velle, L. & Searle, J., *Introduction to Bioethics*, John Wiley and Sons: Chichester, 2005.
[3] Bryant, J.A., Moore, K. & Aves, S.J., Origins and complexes: the initiation of DNA replication. *Journal of Experimental Botany*, **52**, pp. 193–202, 2001.
[4] Raven, P.H., Evert, R.F. & Eichhorn, S.E., *Biology of Plants*, 7th edn, Freeman: New York, 2005.
[5] Avery, O.T., Macleod, C.M. & McCarty, M., Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Induction of transformation by a desoxyribonucleic acid fraction isolated from *Pneumococcus* type III. *Journal of Experimental Medicine*, **79**, pp. 137–158, 1944.
[6] Watson, J.D. & Crick, F.H.C., A structure for deoxyribose nucleic acid. *Nature*, **171**, pp. 737–738, 1953.
[7] Watson, J.D. & Crick, F.H.C., Genetical implications of the structure of deoxyribonucleic acid. *Nature*, **171**, pp. 964–967, 1953.
[8] Kaufman, S.A., *Investigations*, Oxford University Press: New York and Oxford, 2000.
[9] Baisnée, P.-F., Baldi, P., Brunak, S. & Pedersen, A.G., Flexibility of the genetic code with respect to DNA structure. *Bioinformatics*, **17**, pp. 237–248, 2001.
[10] Brown, T.A., *Genomes*, 2nd edn, Bios: Oxford, 2002.

This page intentionally left blank

Chapter 2

Introduction: Part II – Genomes, genes and proteins

J. Bryant
School of Biosciences, University of Exeter, Exeter, UK.

Abstract

Changes in DNA provide the variation based on which natural selection works, but increases in the amounts of DNA have also been involved in the evolution of more complex organisms. However, many complex organisms possess much more DNA than is required for their genetic needs; the function of most of the excess DNA is not known. What is known is that in addition to carrying the genetic code, DNA contains within it signals involved in regulating its function. For replication (copying), DNA is organised into units called replicons which contain specific start points (origins). Copying genes into mRNA is regulated by promoters, specific sequences of DNA adjacent to genes. Many genes are active for most of the time but there are also many that, through regulatory mechanisms that operate on their promoters, are responsive to specific cues or signals. In eukaryotic organisms there are further complexities in mRNA synthesis: the initial transcript copied from the gene contains sequences that interrupt the coding sequence. These are known as introns and are removed so that the pieces (exons) of the coding sequence may be joined together. In some genes, there are alternative sites for this splicing process so that one gene may code for more than one protein. Further, there is evidence for movement of introns or exons between genes, generating more variety in evolution. The final product of most genes is a protein; the sequence of amino acids in a protein, coded by the nucleotide sequence in mRNA, determines its shape and hence its function. We cannot yet predict shape from sequence but we are beginning to understand the processes involved in the correct folding of proteins. Finally, amongst the variations in the basic genetic mechanisms, many, perhaps all, organisms can target specific mRNAs for destruction via the synthesis of inhibitory RNA molecules.

1 Introduction

In the previous chapter it was shown that life on earth is centred round a 'self-replicating' molecule called DNA which both carries the genetic information from generation to generation and makes

available the information in order to direct, in a biochemical sense, the minute-by-minute activities of each cell. The genetic information is organised as genes, specific linear arrays of deoxyribonucleotides within very long DNA molecules. The sum total of an organism's genetic material, whether we are considering a species or an individual within a species, is called its genome and it is to a discussion of the genome that we now turn.

2 Genome evolution

As was shown in Chapter 1, variations occur in the sequence of individual genes and, further, different combinations of variants exist within a population or species. This genetic variation provides the basis for natural selection and hence evolution. But genomes themselves are also subject to evolution. Consider the figures in Table 1.

Firstly, we note a key division in the classification of living organisms between prokaryotes and eukaryotes. Prokaryotes (bacteria) are simple, usually single-celled organisms in which there are no subcellular organelles. Eukaryotes, which range from single-celled organisms such as yeasts right up to highly complex multicellular organisms, are those organisms in which cells do possess organelles (membrane-bound compartments within cells in which are located the structures and proteins related to particular functions). It is clear that there is some relation between genome size and the complexity of the organism. Prokaryotes have less DNA than unicellular eukaryotes, and amongst the eukaryotes in general, complex organisms have more DNA than simple organisms. This implies that not only can variations be introduced into the DNA sequence during replication

Table 1: Genome sizes (based on compilations in [1–3]).

Species	Taxonomic group	Genome size (base pairs of DNA)
Mycoplasma pneumoniae	Bacteria	1.0×10^5
Escherichia coli	Bacteria	4.2×10^6
Saccharomyces cerevisiae, yeast	Fungi	1.3×10^7
Caenorhabditis elegans	Nematodes (roundworms)	9.7×10^7
Drosophila melanogaster, fruit fly	Insects	1.8×10^8
Fugu rubripes, puffer fish	Fish	3.7×10^8
Xenopus laevis, clawed frog	Amphibia	3.1×10^9
Necturus lewisi, mud-puppy	Amphibia	1.2×10^{11}
Homo sapiens, human	Mammals	3.2×10^9
Arabidopsis thaliana, Thale cress	Plants	1.2×10^8
Vicia sativa, field bean	Plants	3.9×10^9
Vicia faba, broad bean	Plants	2.8×10^{10}
Fritillaria assyriaca, fritillary	Plants	1.2×10^{11}

but also, as mentioned in Chapter 1, on occasions, increases in the amount of DNA may occur. Detailed discussion of this lies outside the scope of this chapter; readers who are interested in the mechanisms that lead to increases in genome size will find a clear account in [1]. All that is needed here is to state that analysis of DNA sequences within and between genomes, combined with extensive observation of DNA replication mechanisms, confirms that genomes can increase in size. With more DNA, there is more 'raw material' for the generation of new genetic variation and hence of new genes.

Intuitively, we might expect this. We could not envisage a vertebrate animal or, for that matter, a plant, functioning with the same number of genes as a bacterium. It seems obvious that the very a complex organism needs more genes than a very simple one. In general, the intuitive expectation is correct but when the detail is examined, the picture is not quite so clear. The first point to make here concerns the actual amounts of DNA. How much DNA does an organism actually need to fulfill its genetic requirements? Certainly not as much as many of the organisms listed in Table 1 actually possess. Indeed, with the exception of the simplest organisms, the DNA amounts are far in excess of the genetic needs, even when allowance has been made for the complexity of the genes themselves and for the DNA sequences involved in regulating genes (such as the promoters mentioned in the previous chapter). This excess of DNA over the likely coding requirements in a genome has been confirmed very clearly for those plants and animals whose genomes have been sequenced. This is known as the C-value paradox (the amount of DNA in the single copy of a genome is the C-value). But it is actually even more paradoxical. Within some groups, the C-value varies little between species, as is seen in birds and mammals. In other groups, e.g. amphibia and plants, there is very extensive variation amongst the species in the group. Why does one plant species need thousands of times more DNA than another? Why does a particular plant or amphibian species need so much more DNA than a mammal? It is a paradox indeed. However, two major points are clear. The first is that, within a group, e.g. flowering plants, many of the larger genomes have arisen by a doubling during evolution of the whole genome. This phenomenon can be induced experimentally in the laboratory, but very occasionally scientists are fortunate enough to come across such an event in nature. This happened, for example, when a genome doubling occurred in a population of an infertile hybrid between two species of *Spartina*, cord-grass (a salt-marsh plant). The genome doubling, which restored fertility, effectively created a new species, *Spartina townsendii*, which has spread extensively round the British coasts from its original location in Southampton Water. Secondly, leaving aside genome doubling, much of the variation in genome size is based not on genes but on sequences that have no direct coding function. Furthermore, many of these non-coding sequences are repeated within genomes; for some sequences this amounts to millions of repeats. So, although the number of potential genes has increased by gene duplication and genetic variation, as more complex organisms have evolved, much of the variation between organisms has been caused by extensive amplification of non-coding sequences. This is beautifully illustrated by comparing two species of bean between which there is a seven-fold difference in DNA amounts (Table 1). These closely related species have very similar numbers of genes; the difference between their genomes is almost entirely based on non-coding repetitive DNA sequences.

The excess of DNA over the apparent coding needs is, in many multicellular organisms, vast, amounting to several orders of magnitude. What is its function? It is not uncommon to hear the excess DNA referred to as 'junk DNA'. However, this is a misleading term. The excess DNA is very unlikely to be functionless 'rubbish'. The synthesis of DNA is, in terms of biochemical resources, a very costly process which makes it improbable that it is undertaken for no reason. The location of a particular type of repeated sequence known as satellite DNA at the centromeres and telomeres suggests a structural function: the telomeres are the ends of chromosomes at which

a special biochemical mechanism exists to prevent chromosomes 'fraying' and the centromeres are the structures by which chromosomes are pulled apart during cell division. The involvement of particular DNA sequences in the structural features of chromosomes is another example of the 'many-sidedness' of this remarkable molecule. However, for much of the excess DNA in a genome, there is still no clear idea as to its function: some of the mysteries of DNA still remain unsolved.

In organisms possessing the larger genomes, the actual lengths of the DNA molecules are amazing. The DNA of eukaryotic organisms is of course distributed amongst several chromosomes (e.g. 23 pairs of chromosomes in humans), but even so, the individual DNA molecules, one per chromosome, in many multicellular organisms are very long. Indeed, in some plants, some of the DNA molecules are up to a metre long. These are then extraordinary molecules, very long but also very thin: the diameter of a DNA molecule is about 2 nm. A helpful way to consider this is to imagine that the nucleus, the cell compartment or organelle in which the chromosomes are located in eukaryotic organisms, is a ball of 10 cm diameter. On this scale, a DNA molecule of 1 m in length would be magnified to a length of 10 km but would only be 2 mm in diameter.

Even when we shrink the nucleus and the DNA molecules back to their normal sizes, the topological problems imposed by the need to pack very long molecules into the nucleus are huge. This packaging is achieved firstly by coiling the DNA round complexes of proteins called histones. The DNA–histone complex is called chromatin and this is further coiled to form a 30 nm diameter fibre, which, during the process of cell division, becomes very compacted and thus the chromosomes become visible with the light microscope [1]. Those of us who work in this field are amazed that, with this amount of coiling, the enzymes that mediate transcription (see Chapter 1) and replication of DNA can actually locate the correct places along the DNA molecules to start these processes.

3 Organisation of DNA for replication

The basic process of DNA replication, the copying of the genetic information, was described in Chapter 1. However, there is much more to this process than the basic biochemical mechanism. The overall process must be regulated and organised within the context of cell division. The regulatory mechanisms must ensure that DNA is replicated only at the right time within the life of a cell. Description of these mechanisms lies outside the scope of this chapter. Indeed, it would require several chapters to provide a comprehensive description. Here we simply note that regulation involves a network of different sets of mechanisms that operate at different levels, from those that involve direct action on the DNA up to those that link replication with developmental or even environmental cues.

However, one feature of replication that is especially relevant to a consideration of design is the organisation of the DNA molecule for replication. The process of replication does not start just anywhere but at specific sites within the molecule. These sites are known as origins of replication. In the small circular DNA molecules of bacteria, there is one origin of replication and the process proceeds outwards in both directions from the origin (Fig. 1a and b). (Note that this description does not cover the complexity that is inherent in the process that is caused by the two strands of the DNA double helix being in opposite orientations (i.e. are anti-parallel).)

The DNA molecules of bacteria are small and the process of copying an 'average' bacterial genome is complete in a matter of minutes. However, the genomes of eukaryotic organisms, and especially those of multicellular eukaryotes, are very different. Firstly, as already noted,

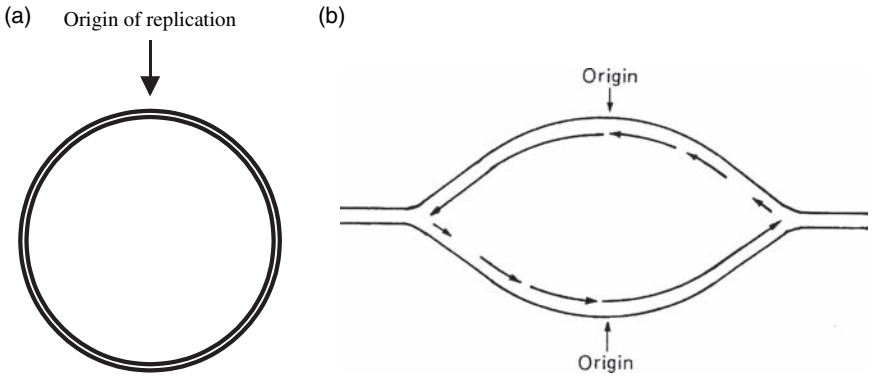


Figure 1: (a) Bacterial DNA molecules are replicated from a single origin of replication. (b) Diagram of the replicative process proceeding from the origin. Reproduced by permission from Bryant, J.A., *Biochemical Mechanisms Involved in Plant Growth Regulation*, eds. C.J. Smith, J. Gallan, D. Chiatante & G. Zocchi, Oxford University Press: Oxford, pp. 139–153, 1994.

the amounts of DNA possessed by eukaryotic organisms are much larger than those in bacteria, sometimes by several orders of magnitude. At the same rate of replication as in bacteria, it would take days rather than minutes to copy these very long molecules if replication started at a single origin of replication.

And it is not only length that poses a problem. As we noted earlier, these enormously long DNA molecules are packaged in an amazing way into the cell nucleus which is only a few micrometres in diameter. The packaging into the individual units of chromatin, the chromosomes, places further constraints on the speed of replication; overall it proceeds, even under optimal conditions, at 100 or more times more slowly than in bacteria. How then do complex multicellular organisms complete DNA replication in hours rather than weeks? The answer is that instead of starting in one place, DNA replication starts in many places along the chromosome (Fig. 2; [4]). This is seen in single-celled eukaryotes such as yeasts, which have a few hundred origins of replication, right through to those species that have very large genomes which possess hundreds of thousands of origins. Thus, as the genome size has increased, the number of replication origins has also increased. Even so, the more complex multicellular organisms take several hours to complete the replication of their DNA.

These replication origins are places along DNA molecules that are recognised by the biochemical 'machinery' that replicates DNA. The mechanisms involved in this recognition and in the regulation of the overall process lie outside the scope of this chapter. Nevertheless, to those of us who work on this process, these mechanisms represent beautiful examples of structure–function–control relationships in biological systems. For readers who are interested, some of this is dealt with in [4]. The final point to make here is that flexibility is built into the system. There are times when organisms need to replicate their DNA faster than usual, e.g. in early embryonic growth in many animals. At these times, more replication origins are brought into play such that the amount of DNA replicated from each origin is very much reduced. Thus, not only have eukaryotic organisms acquired, during the course of evolution, multiple start points for DNA replication but they have also developed a built-in flexibility such that start points may be used according to the needs of the organism at any particular stage of its life.

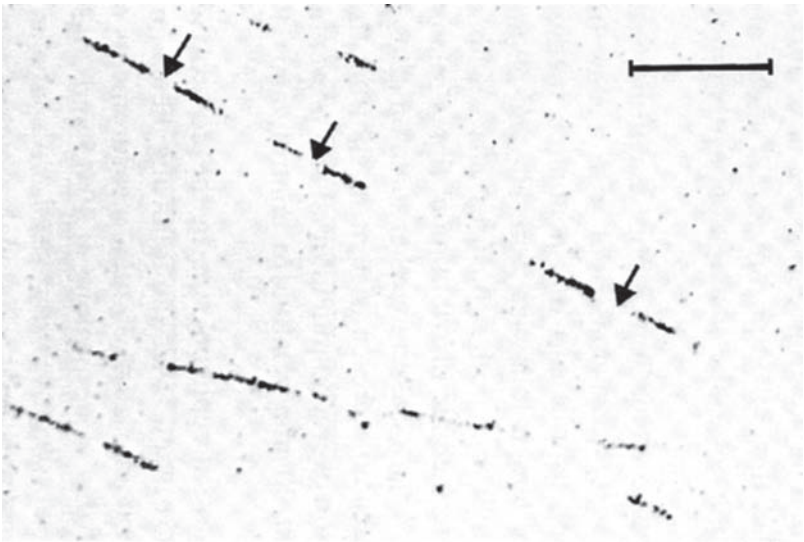


Figure 2: Eukaryotic DNA molecules are replicated from multiple origins of replication, as illustrated here by autoradiographs of replicating pea (*Pisum sativum*) DNA molecules that are incorporating radioactive deoxyribonucleotides at several places along the fibre. The likely position of origins on one of the DNA fibres are indicated by arrows. The scale bar represents 10 μm. Reproduced by permission from Francis, D., Davies, N.D., Bryant, J.A., Hughes, S.G., Sibson, D.R. & Fitchett, P.N., *Experimental Cell Research*, **158**, pp. 500–508, 1985.

4 More on gene structure and function

4.1 Promoters

In Chapter 1 the essential features of gene structure and function were described. The basic features are:

- A promoter 'upstream' of the gene, which is essentially the gene's on–off switch.
- The transcription start and stop sites, which define the stretch of DNA that is copied into messenger RNA (mRNA).
- The coding sequence, which carries the information that instructs the synthesis of a protein during translation.
- At each end of the coding sequence are short stretches of DNA, which although copied in the mRNA molecule are not involved in coding for the protein. These untranslated regions are involved in the mechanics of starting and stopping protein synthesis.

Promoters thus have a key role in regulating the activity of the adjacent gene. How is this done? In general, they do this through the activity of proteins or complexes of proteins called transcription factors. The basic features of the process are the recognition of the promoter sequence by a transcription factor, the binding of the transcription factor to the promoter and making the DNA more accessible to the enzyme (RNA polymerase) that will carry out transcription.

However, a moment's thought will show us that this cannot be the whole story. Even in the simplest bacterium, not all the genes are switched on at any one time. Some genes are obviously needed for the minute-by-minute life of the cell. These are often called 'house-keeping' genes. In genetic terminology, they are constitutively active. However, many genes are switched on or off in response to specific environmental conditions or at particular developmental stages. We can envisage several ways in which this might be done. For example, a negative transcription factor may be bound to a promoter thus keeping a gene switched off except when expression of that gene is required. Alternatively, a positive transcription factor may allow gene expression, either constitutively or under particular circumstances, by binding to the promoter and thus promoting binding and activity of RNA polymerase. Three examples, one from bacteria, one from animals and one from plants, illustrate these principles.

The bacterium *Escherichia coli*, has been for biochemists and geneticists a favourite experimental organism for well over 50 years. In the context of gene expression, it was the organism in which inducible and repressible gene expression was first discovered. That discovery was made by Francois Jacob and Jacques Monod who were later awarded the Nobel prize for their work. The discovery was based on the ability of the bacterial cells to utilise different sugars as a carbon source. In the standard growth medium, the cells are provided with glucose but can actually grow equally well with other sugars. In the classic experiments, cells of *E. coli* were transferred from a growth medium containing glucose to a growth medium containing lactose (the sugar present in milk). Within a few minutes of the transfer, the cells started to make three enzymes that enabled the cells to utilise lactose. The two more important of these are lactose permease (which increases dramatically the rate at which lactose is taken up into the cells) and β-galactosidase (which breaks down the lactose into two smaller sugars, glucose and galactose). The third enzyme, lactose *trans*-acetylase has a minor role in the overall process. What is happening here? First, we note that the genes encoding the three enzymes are activated together. This is an example of *coordinate regulation*. The three genes (named, perhaps confusingly, as the Z, Y and A genes) are actually next to each other in a functional group that Jacob and Monod called an *operon* (this particular operon is known as the *lac* operon). Secondly, the factor that makes this happen is the lactose itself: lactose is therefore an inducer of gene expression. Thirdly, the ability to keep the three genes switched off in the absence of lactose depends on the activity of another gene located next to the *lac* operon. The overall mechanism is illustrated in Fig. 3. The gene whose activity represses the *lac* operon in the absence of lactose produces a repressor protein, i.e. a negative transcription factor that binds to a specific sequence of bases in the promoter. This prevents transcription of the genes. However, when lactose is present, the repressor protein binds to lactose. It has a higher affinity for lactose than for DNA and the binding of lactose changes the shape of the repressor protein so that it cannot in any case bind to DNA. The operon is now available for transcription.

Since this work of Jacob and Monod, further fine-tuning mechanisms operating on this operon have been discovered. However, the key principles remain. Gene expression may be constitutive (as with the gene that encodes the repressor protein) or inducible/repressible (as with the *lac* operon). These principles apply right across the living world, from the simplest to the most complex organism. However, the details vary. Firstly, although coordinate regulation in bacteria is mainly achieved by the organisation of genes as operons under the control of a single promoter, this is not true of eukaryotic organisms, whether they be simple unicellular fungi such as yeast or complex animals or plants. Indeed, genes that are coordinately regulated may be widely spaced from each other, often on different chromosomes. Coordinate regulation is thus achieved because of similar base sequences within their promoters that are recognised by the appropriate positive or negative transcription factors. Secondly, in bacteria many of the inducer substances are small molecules that are either present in the organism's immediate environment, as exemplified by

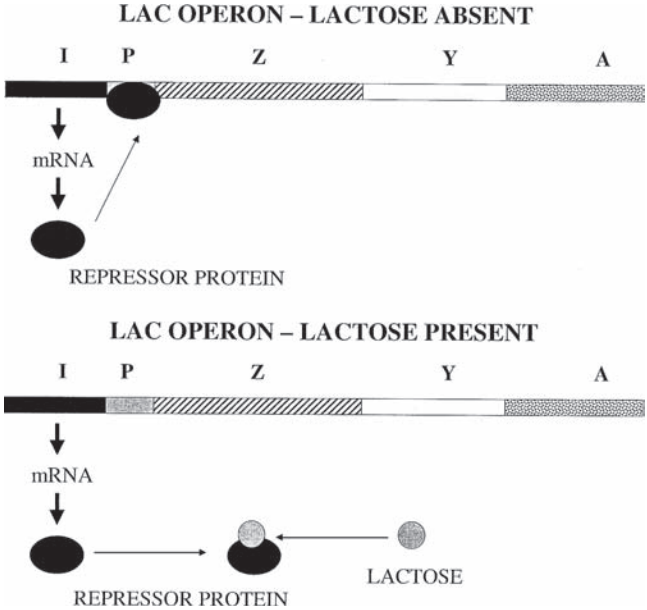


Figure 3: The induction by lactose of gene expression in the *lac* operon of the bacterium *E. coli.* The three genes of the operon, Z, Y and A, are under the control of a single promoter, P. Adjacent to the operon is a single gene, I, whose promoter allows the gene to be switched on all the time. The protein encoded by the I gene binds to a specific sequence in the *lac* operon promoter and prevents the three genes being copied into mRNA. (Hence, none of the three proteins encoded in the operon are produced.) However, the repressor protein has a higher affinity for lactose than for the DNA sequence in the promoter and thus when lactose is present, the repressor protein leaves the *lac* operon promoter and binds instead to lactose. The *lac* operon genes are now available for mRNA synthesis, leading to the production of the proteins that metabolise lactose. Note that the diagram is not to scale (otherwise the lactose molecule would be too small to be visible).

lactose, or generated during cellular metabolism (in which case they may function to induce or repress the expression of enzymes involved in particular metabolic interactions or pathways). In eukaryotes, gene induction or repression by metabolites certainly occurs and indeed is common in unicellular eukaryotes such as yeast. However, in more complex eukaryotes, induction or repression of gene expression is very much more a key part of development than it is in unicellular organisms, which means that developmental cues are very important in gene regulation. In addition, there are also many examples of responses at the gene level to environmental signals, some of which may interact with developmental cues. These principles of developmental and environmental regulation are illustrated by two well-known examples, namely the regulation of gene activity by steroid hormones in mammals and by light in plants.

Different steroid hormones regulate a range of developmental, growth and maintenance processes in mammals. Their action on gene expression is in general to act as an inducer, but they do not interact directly with the relevant genes. When a steroid hormone such as testosterone enters a cell that is competent to respond, it encounters a steroid-binding protein or hormone receptor. The receptor proteins with their bound hormone molecules then form pairs – dimers; the dimerised

····**GGA/TACANNNTGTTCT**···· →
····**CCT/ATGANNNACAAGA**····

Figure 4: The androgen (male hormone) response element, the sequence in the promoters of androgen-responsive genes to which the hormone receptor protein plus hormone binds. The gene is 'downstream' from this sequence, as shown by the arrow.

receptors travel to the cell nucleus and bind to the gene promoters that contain a specific sequence called the androgen response element (Fig. 4). The genes associated with these promoters are thus coordinately induced even though they are spread throughout the genome.

In plants, light is a major regulator of developmental processes, as well as being the energy source for the process of photosynthesis. The effects of light on developmental processes are mediated via a number of light receptor molecules and involve changes in gene expression. The most abundant of these light receptors are the phytochromes which exist in two states. In one of the states, P_r, it is receptive to light in the red region of the spectrum with a peak wavelength at 660 nm. The light energy causes a change in the shape of the non-protein part of the molecule (see Fig. 5a) and in most of the processes involving the phytochrome system it is this form, known as P_{fr}, that is active. The active phytochrome is then able to trigger a range of molecular, cellular or developmental responses, according to the specific situation in which it has been activated. It may also be inactivated by one of three routes (Fig. 5b). Firstly, under the influence of far-red light (peak active wavelength 730 nm), P_{fr} is converted back to its inactive form, P_r. Secondly, this conversion also takes place slowly in the dark. Thirdly, P_{fr} can be slowly broken down, probably by a specific protein-hydrolysing enzyme (protease).

Thus, cells are equipped with sensing systems, many of which are involved in the induction or repression of gene activity. This relies on specific transcription factors and on specific sequence elements within promoters and leads to the controlled expression of genes in response to internal or external conditions or at particular developmental stages. Of necessity, this brief description cannot convey the amazing complexity of gene regulatory mechanisms that exist across the living world. A full description of what we know about these (and there is much that we still do not know) would take a large volume of its own. However, in order to convey some of the wonder of this, let the reader imagine the gene control mechanisms involved in the development of an adult human, consisting of around 70 million million cells (and over 200 different types of cells), from a single cell, the fertilised egg.

4.2 mRNA synthesis in eukaryotic cells

In the previous chapter the basic mechanisms involved in bacterial gene expression were presented. However, in eukaryotes, there are a number of further aspects that we now need to consider. The first of these is the addition of extra residues at each end of the newly synthesised mRNA molecule. At the 5′ end, the end that engages first with the ribosome (Fig. 1 in the previous chapter), a single modified GTP is added in the reverse orientation to normal, making a 5′ to 5′ triphosphate bridge. This structure is known as the mRNA cap and facilitates the binding of the mRNA to the ribosome. At the 3′ end, a series of adenosine monophosphate residues is added. This poly(A) tail is a stability regulator. For example, the mRNA that encodes globin (the protein part of haemoglobin) has a poly(A) tail of about 200 residues and may remain functional for up to 2 weeks in human red blood cells. At the other extreme, the mRNAs encoding the various histones, the major proteins

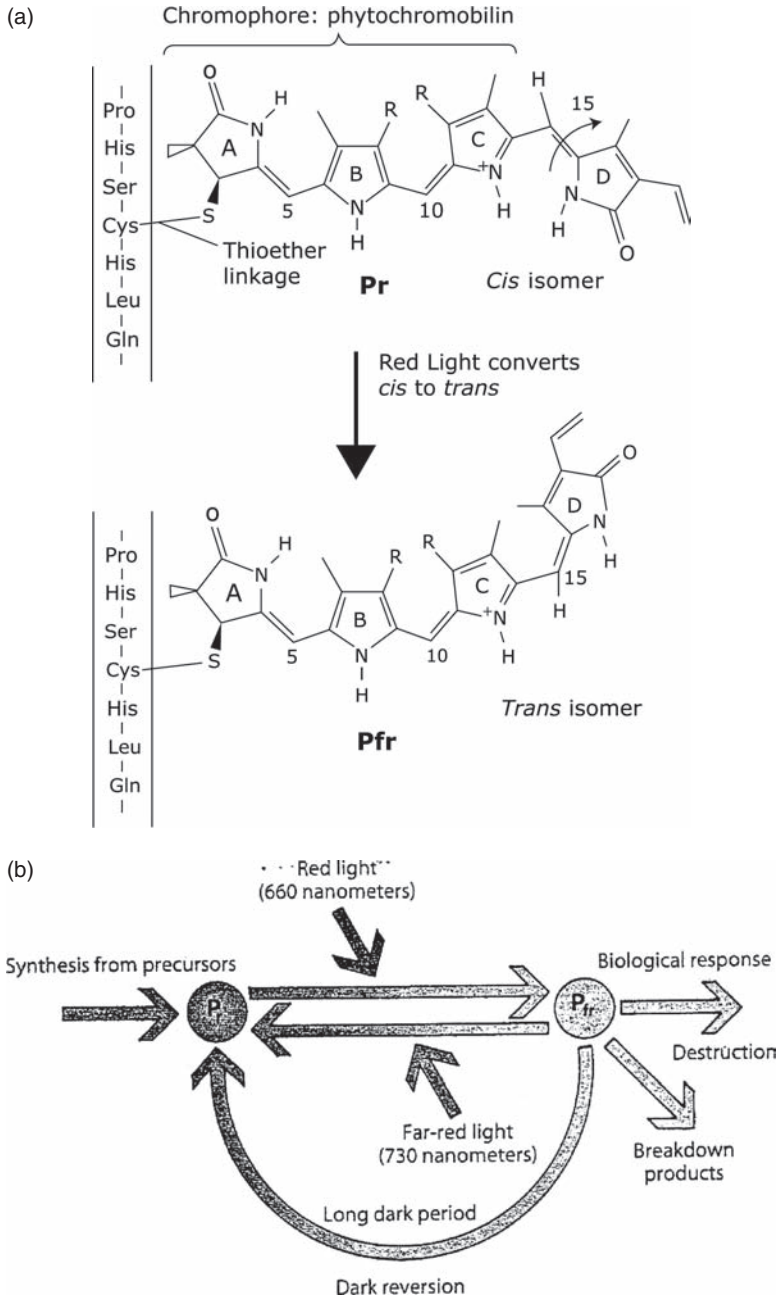


Figure 5: (a) The non-protein part of the plant light receptor phytochrome changes its configuration according to the wavelength of the light it perceives. The diagram shows the 'flip' from one configuration to the other and also indicates the attachment to the protein part of the molecule. (b) Interconversions between the different forms of phytochrome. Reproduced by permission from Raven, P.H., Evert, R.F. & Eichhorn, S.E., *Biology of Plants*, 7th edn, Freeman: New York, 2005.

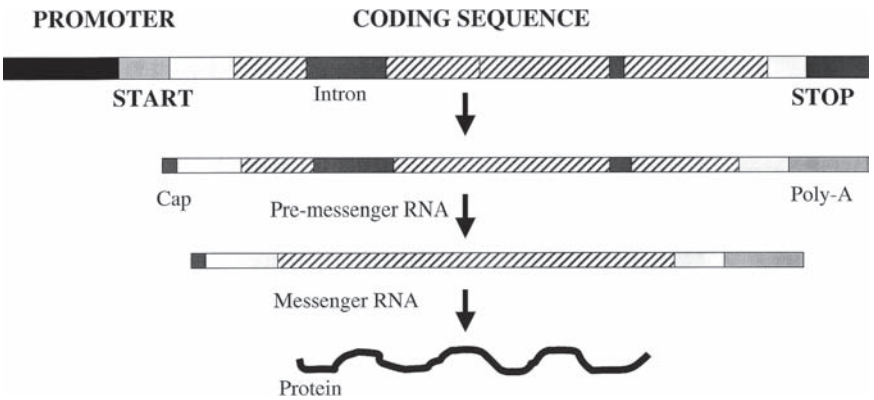


Figure 6: Most eukaryotic genes contain introns, non-coding DNA sequences that interrupt the coding regions. These are copied into pre-mRNA which is immediately modified by the addition of the cap (a modified base) and the poly(A) tail. The introns are then removed by a cut-and-rejoin mechanism in order to obtain finally the uninterrupted mRNA sequence.

associated with DNA in the chromosomes, are only needed for a few hours during the process of DNA replication. These mRNAs have no poly(A) tails.

In a sense, these additions are not remarkable. They are simply reminders that despite the universal genetic code and the general similarities of gene function across the living world, there are variants in the details of the mechanisms themselves and their regulation in different types of organism. However, another of the differences between eukaryotes and prokaryotes, when initially discovered in the 1970s, met with a degree of surprise that bordered on disbelief. The discovery was that many eukaryotic genes (in complex multicellular eukaryotes, most of their genes) contain sequences of DNA which interrupt the coding sequence (Fig. 6) and that these sequences are copied during mRNA synthesis. These interrupting sequences are removed by an editing process after the addition of the cap and the poly(A) tail. Thus, in eukaryotes, we refer to the unedited version as pre-mRNA; the interrupting sequences in the genes are known as *introns* and the coding sections are called *exons*.

4.3 Splicing and shuffling

The removal of the introns from the pre-mRNA must of course be a highly regulated process in order that the correct coding sequence is retained in the mRNA. It will not surprise readers to know that once again, recognition of specific base sequences is involved. The splicing out of the introns is carried out by complexes of proteins and small nuclear RNAs (snRNAs). The snRNAs bind to the splice sites by specific base pairing, enabling the splicing enzyme to cut the RNA at that site. In addition, a looping out of the intron brings the two cut sites at either end of the intron into close proximity so that the cut ends may be joined. The mRNA is thus ‘edited’.

In encountering such mechanisms a biologist will always ask why it has evolved. As with the large excess of DNA in many eukaryotic organisms, it seems so wasteful. However, within the mechanism just described there are possible indicators of evolutionary advantage. The first of these is that in many genes, alternative splicing sites exist. That is, whereas particular sites may be preferred, others may, under specific conditions, be used, thus generating alternative mRNAs

from an individual gene. Whilst this may seem strange, set against the large excess of DNA, it is not so strange if we consider that it may be advantageous to have two proteins whose synthesis is under the control of a single promoter region.

Secondly, by examining the sequences of genes encoding proteins that have some properties in common, it appears that the exon-coding regions may recombine between genes, thus creating new genes encoding new proteins. Readers wishing to know more about this topic will find a clear and concise account in [1].

5 Gene sequence and cellular function

The essential working molecules of cells are the proteins. As described in Chapter 1, proteins are made of building blocks called amino acids, of which there are 20. It is the order of amino acids that is encoded in the DNA and the order of amino acids that governs the overall shape of a protein. This is because the different amino acids, in addition to being joined end-to-end to form the 'backbone' of the protein, can also interact with each other, forming linkages of various types between different parts of that backbone. The protein thus folds up into a specific 3D shape. That specific shape may appear to the non-expert as a ball of tangled string, but it is actually vital for the protein's function. Thus, the ability to mediate specific chemical reactions (i.e. to act as an enzyme) or to perform a structural function or to form a channel for taking metal ions into the cell, or indeed any other function is dependent on the 3D shape of the protein.

As already noted, this 3D shape depends on the order of particular amino acids within the protein chain and their ability to interact with amino acids elsewhere in the chain. In a very real sense then the shape of the protein is built into the genetic code. It is almost the 'holy grail' of protein biochemistry to be able to predict from the DNA coding sequence (from which we know the order of amino acids) what the 3D shape of a protein will be. However, this has eluded all the best efforts of biochemists and bioinformaticists. Nevertheless, we are beginning to understand, thanks to the elegant work of, amongst others, Professor Christopher Dobson, the mechanisms involved in a protein 'getting into shape' [5]. The final structure of the protein is in general the most thermodynamically stable, the configuration of least energy. However, the number of possible interactions between amino acids is very large, especially in large proteins. It is apparent that during the folding process some interactions form the core elements in the final shape and that around these the protein 'tries' different configurations until the final shape is achieved. We should note that the ability to do this involves the reversibility of amino acid interactions and that the less stable interactions are those that are more readily reversed. The final product of this process is a protein that is exquisitely fit for its function. Indeed, the distinguished American biochemist, Dan Koshland, has referred to enzymes as beautiful, finely tuned miniature machines. One amazing example of this is the enzyme ATP synthase which carries out the final step in trapping the energy from respiratory oxidation. This complex enzyme consists of several individual protein molecules, some of which rotate like a tiny turbine, operating at a scale of nanometres. One can only react with wonder, even awe, at such miniature elegance!

6 How many genes are needed?

We have already noted that the amounts of DNA in many, perhaps most, multicellular organisms appear to exceed, often by several orders of magnitude, the amount needed to satisfy the genetic

requirements of an organism. But how many genes does an organism actually need? It is actually quite difficult to obtain an accurate answer to this question but 'ballpark' figures, based on genetic analysis and on biochemical data, are of the order of 25–30,000 for plants and perhaps 80–100,000 for mammals. However, the advent of genome sequencing provides a much more direct approach to the problem: find all the genes and count them. When this is done, most of the data confirm our view that increased biological complexity requires greater numbers of genes. However, the data for mammals are totally surprising. Instead of 80–100,000 expected, humans and other mammals possess only 25–26,000, around the same as a flowering plant. Birds, also complex warm-blooded animals, have 20–23,000 genes. Mammals and birds are clearly much more complex than flowering plants and yet manage with about the same number of genes. However we view this, it is clear from comparison of gene numbers with genetic function that birds and mammals do not seem to have nearly enough genes.

At least a partial answer to this conundrum is that some genes, perhaps a significant proportion of genes, are multifunctional. We have already noted that by splicing a pre-mRNA in different ways it is possible to generate more than one type of mRNA. It is also possible to envisage alternative start sites within genes, and we certainly know of ways in which particular proteins may be processed or modified in different ways after they have been formed. All these mechanisms will increase the number of gene products that are encoded within a set of genes. This provides an efficiency of use of DNA sequences that had previously been encountered only as a few rare examples.

However, these solutions to the gene number paradox in mammals again raise the question of the excess DNA. If mammals can exist with only around 25–35% of the expected gene numbers, then the amount of DNA that is in excess of coding needs is even greater. It is easy to understand the selective pressure to maximise the coding usage of DNA in very small genomes such as those of viruses. It is much harder to understand what the selective pressure has been to maximise the use of individual genes in organisms that appear to have no need to do this.

7 Variations on a theme

In this and the previous chapter I have set out the main ways in which genomes are replicated and genes are regulated. Most genomes consist of DNA and at the heart of the systems that replicate them are the enzymes that copy the template DNA to make the daughter molecules. These are the *DNA-dependent DNA polymerases*. However, in many viruses, the genome is made not of DNA but of RNA. There are two main ways in which these genomes are replicated. First, in the retroviruses, the RNA genome is copied into a DNA version by an *RNA-dependent DNA polymerase*, often referred to colloquially as reverse transcriptase. This DNA version of the genome serves as a template for making new RNA copies using *DNA-dependent RNA polymerases* (similar to the enzymes that make the 'normal' types of RNA in a cell). Some of these retroviruses insert the DNA copy of their genome into the host's genome thus leaving a permanent copy of the virus's genes. An example of this is the human immunodeficiency virus (HIV), the causative agent of acquired immunodeficiency syndrome (AIDS). Secondly, in many RNA viruses, and particularly in those that infect plants, the viral RNA genome is copied by an *RNA-dependent RNA polymerase*. A summary of these various polymerases is shown in Table 2.

Mention of this last group of enzymes, the RNA-dependent RNA polymerases, brings us to consider a very exciting recently discovered group of mechanisms for regulating RNA populations within cells. In the 1970s and early 1980s, investigators studying the infection of plants by RNA viruses such as tobacco mosaic virus were not surprised to find that infected plants contained

Table 2: Enzymes that synthesise nucleic acids in a template-dependent manner.

Enzyme	Template	Product	Notes
DNA-dependent DNA polymerase	DNA	DNA	Used in replicating and repairing DNA
DNA-dependent RNA polymerase	DNA	RNA	Makes RNA copies of genes and other coding sequences
RNA-dependent DNA polymerase	RNA	DNA	Involved in replication of some types of virus
RNA-dependent RNA polymerase	RNA	RNA	Involved in replication of some types of virus; also involved RNA-dependent gene silencing

RNA-dependent RNA polymerases that were involved in replicating the viral genome. However, what was surprising was that uninfected plants were often shown to possess RNA-dependent RNA polymerases of their own, the function of which was completely unknown [6, 7]. We now know, firstly, that double-stranded RNA molecules are targets in plants and animals for a specific degradation pathway. In plants this provides one of the defences against RNA viruses that must go through a double-stranded RNA phase, at least transiently, every time a virus RNA genome is copied. However, more intriguingly, plant and animal cells can target specific mRNAs for this destruction pathway by copying them with RNA-dependent RNA polymerases, thus making them double-stranded. (Thus we at last know why uninfected cells have this type of polymerase.) Further, the ability to target specific messages means that the activity of a particular gene may be quickly silenced, giving rise to the term 'RNA-based gene silencing'. The details of the mechanism for destruction of the double-stranded RNAs lie outside the scope of this chapter. Here we just need to note that the pathway involves the generation of small RNA molecules which are themselves very inhibitory and which, by careful design of the RNA sequence, can be used by scientists to silence individual genes. These small RNA molecules are known as small inhibitory RNAs and the general mechanism as RNA interference (RNAi). Any reader who is interested in finding out more will find clear accounts of this exciting topic in [8] and [9].

8 Concluding remarks

These first two chapters have outlined the key features of information maintenance, transmission, retrieval and use that lie at the heart of life on earth. They have also outlined some of the ways in which these processes are regulated, providing food for thought for anyone who turns to nature for inspiration about regulation of human industry and commerce. However, it is apparent, even from this overview, that there is much that we do not yet know (and some things that we probably never will know). Nevertheless, we are beginning to understand some of mechanisms involved in the way in which these core processes are used in the life of the whole organism. Much of the rest of this book is concerned with this topic, as we examine in more detail, in specific situations, the functioning of genes and proteins before finally moving on to considering some of the higher order functions of particular living organisms.

References

[1] Brown, T.A., *Genomes*, 2nd edn, Bios: Oxford, 2002.
[2] Bryant, J.A. (ed.), *Molecular Aspects of Gene Expression in Plants*, Academic Press: London, 1976.
[3] Gregory, T.R., http://www.genomesize.com/, last consulted on 12/12/2005.
[4] Bryant, J.A., Moore, K. & Aves, S.J., Origins and complexes: the initiation of DNA replication. *Journal of Experimental Botany*, **52**, pp. 193–202, 2001.
[5] Dobson, C.M., Protein folding and misfolding. *Nature*, **426**, pp. 884–890, 2003.
[6] Evans, D.M.A., Bryant, J.A. & Fraser, R.S.S., Characterization of RNA-dependent RNA polymerase activities in healthy and TMV-infected tomato plants. *Annals of Botany*, **54**, pp. 271–281, 1984.
[7] Evans, D.M., Fraser, R.S.S. & Bryant, J.A., RNA-dependent RNA polymerase activities in tomato plants susceptible or resistant to tobacco mosaic virus. *Annals of Botany*, **55**, pp. 587–591, 1985.
[8] Black, D. & Newbury, S., RNA interference – what it is and what it does. *The Biochemist*, **26**, pp. 7–10, 2004.
[9] Zhang, J., Special delivery – small RNAs silencing gene expression. *The Biochemist*, **26**, pp. 20–23, 2004.

This page intentionally left blank

Chapter 3

Green grass, red blood, blueprint: reflections on life, self-replication, and evolution

M. Ciofalo
Dipartimento di Ingegneria Nucleare, Università degli Studi di Palermo, Italy.

Abstract

Following pioneering work by von Neumann in the late 1940s, the goal of achieving self-replication in artefacts has been pursued by a variety of approaches, involving either virtual entities like cellular automata and computer programs or, to a lesser extent, real physical devices. An ample review of the major achievements in these diverse fields is given, and their practical and theoretical relevance is discussed. Possible future developments, notably regarding nanotechnology and space exploration, are also outlined. The most relevant theoretical problems posed by self-replication are discussed in the light of current knowledge regarding life and its origins. Living entities are semiotic systems, in which physical structures have come to perform symbolic functions. The great complexity of biomolecules and of even the most primitive organisms is not a gratuitous complication, but a necessary condition for homeostasis, self-replication and open-ended evolution in a changing environment. Such requisites will have to be matched by artificial devices if their non-trivial self-replication and autonomous development are to be attained.

1 Of crystals and colloids

> Wordsworth's God had his dwelling in the light of setting suns. But the God who dwells there seems to me most probably the God of the atom, the star, and the crystal. Mine, if I have one, reveals Himself in another class of phenomena. He makes the grass green and the blood red.
>
> (J.W. Krutch, 1950, [1])

The lines in the epigraph are excerpted from the famous essay 'The colloid and the crystal', written in 1950 by the American literary naturalist Joseph Wood Krutch. The best known passage of the essay is probably the following: 'A snowflake under a lens and an amoeba under a

microscope ... Crystal and colloid, the chemist would call them, but what an inconceivable contrast those neutral terms imply! Like the star, the snowflake seems to declare the glory of God, while the promise of the amoeba, given only perhaps to itself, seems only contemptible. But its jelly holds, nevertheless, not only its promise but ours also, while the snowflake represents some achievement which we cannot possibly share ... No aggregate of colloids can be as beautiful as the crystal always was, but it can know, as the crystal cannot, what beauty is.'

Krutch's words express in poetic form the naturalist's diffidence against the perfect, but sterile symmetry of inanimate things (the 'crystal') and his unconditional preference for the imperfect, but dynamic living matter (the 'colloid'), of which we ourselves partake. This is the mood in which Goethe wrote 'Grey is all theory. Green grows the golden tree of Life', and the mood underlying *vitalism*, the doctrine that life processes arise from a nonmaterial vital principle and cannot be entirely explained as physical and chemical phenomena. This attitude probably arose as a reaction against the excessive boasts of mechanical and physical sciences in the 18th and early 19th century, and was the more justified in 1950, a few years after 'hard physics' and technology had yielded Hiroshima and Nagasaki, Zyklon B, and the assembly line.

However, as we now see with some clarity, the inability of physical sciences to explain life was not due to an intrinsic irreducibility of life phenomena to within the realm of physical laws, but rather to the vast inadequacy of the available physical and mathematical models on one side, and of our knowledge of the intimate structure and function of living matter on the other side. This was already clear to Ernst Mach who, in '*Knowledge and Error*' [2], wrote 'If one reduces the whole of physics to mechanics, and mechanics itself to the simple theories known today, life will necessarily appear as something hyper-physical.'

By a curious twist of this story, in the very year in which Dr. Krutch wrote of crystals and colloids, Max Perutz [3] was using X-ray diffraction, introduced by Bragg for studying crystals, to unravel the structure of that most life-related of all 'colloidal' substances, the very haemoglobin that makes blood red. Molecular biology was being founded as an autonomous science following these and other important contemporary achievements, including the discovery of the double helix structure of DNA by Watson and Crick in 1953 [4] and the abiogenic synthesis of amino acids by Miller and Urey in the same year [5].

Meanwhile, Ilya Prigogine [6] was founding the non-linear thermodynamics of systems far from equilibrium and dissipative structures; John von Neumann [7] was concentrating his efforts on the problem of artificial self-replication; René Thom was starting the research program that would lead to a mathematical theory of morphogenesis [8] in the spirit of D'Arcy Thompson [9]; and a 'hard-boiled' physicist like Erwin Schrödinger was studying life issues in Dublin [10]. Long before the digital computer established its dominant place in science, and terms now fashionable like complexity or chaos entered common use, the science of *Cybernetics* of Wiener [11] and Ashby [12] and the General Systems Theory of von Bertalanffy [13] were providing powerful paradigms for studying natural and artificial systems under the same light.

These developments were rapidly narrowing the gap between physical and life sciences; so much so that in 1979 Ilya Prigogine might suggest that times were ripe for attempting a new synthesis, capable of bridging this gap altogether and to embrace human sciences as well [14].

Probably no other field of human enquiry is so gap-bridging (to the verge of being regarded by some as an undue intrusion into 'forbidden territory') as the attempt to understand and re-create processes peculiar to life; and, in particular, that most exclusive of all life-related processes, *self-reproduction*. The present contribution is dedicated to this issue.

2 Queen Christina's challenge

> Trying to discover how a biological mechanism works has an advantage over solving problems in non biological areas since one is sure the problem can be solved; we must just be clever enough.
>
> (M. Delbrück, by Laithwaite [15])

According to a popular, and probably apocryphal, anecdote, when René Descartes expressed the view that animals are but clockwork automata to his royal student, Queen Christina of Sweden, she pointed to a clock and challenged him 'Well, Monsieur Descartes, show me how it makes a child'. Needless to say, Christina's challenge could not be seriously taken up for many centuries. In the words of Freitas [16], 'It was not until 1948 that scientists became convinced that machines could be taught to replicate themselves. In that year John von Neumann ... gave a series of historic lectures at the University of Illinois ... He was able to prove, with mathematical rigor, the possibility of building self-reproducing machines.'

Sipper *et al*. [17] made a distinction between the two terms *replication* and *reproduction*:

- replication is an *ontogenetic* developmental process, involving no genetic operators, resulting in an exact duplicate of the parent organism.
- reproduction is a *phylogenetic* (evolutionary) process, involving genetic operators such as crossover and mutation, giving rise to variety and ultimately to evolution.

However, in most works described herein these two terms are considered synonymous and are used interchangeably.

The motivations for studying self-replication are both theoretical and practical and can be summarised as follows.

Theoretical goals:

- understanding bio-molecular mechanisms of reproduction and origins of life;
- understanding complex system dynamics and emergent properties;
- improving *artificial life*, leading to a better understanding of evolution and ecology.

Practical goals:

- achieving self-replicating massive architectures for parallel computing;
- achieving self-repairing and homeostasis in electronic and mechanical machines;
- achieving self-replication in automata (e.g. for nanotechnology or space exploration).

Among the theoretical questions that current research on self-replication strives to solve, perhaps the most deep and fascinating one is whether the appearance of life on our planet has to be regarded as an exceptional event, or rather as the inevitable outcome of the physical conditions prevailing in this corner of the universe [18, 19].

With some inevitable simplification, the two opposite answers to this issue can be represented by Jaques Monod [20], who stresses the role of contingency in the very existence of life, and certainly in our existence ('chance caught on the wing'), and Stuart Kauffman [21], who regards self-replication and life as the inevitable outcome of the complex, autocatalytic set of chemical

reactions that occur whenever organic macromolecules are free to interact ('we the expected ... at home in the Universe').

Roughly speaking, Monod's view – echoed in a somewhat milder form by Francois Jacob [22] – is representative of the more orthodox Darwinian tradition in the natural sciences and evolutionary biology, whereas Kauffman's ideas reflect the school of thought associated with the study of *complexity*. An even more extreme statement of the belief in the inevitability of life is the following, due to Thomas Ray [23]: 'The living condition is a state that complex physical systems naturally flow into under certain conditions. It is a self-organizing, self-perpetuating state of autocatalytically increasing complexity.'

Needless to say, the issue also has practical implications, for example for the policy of scientific research: only a strong belief that life will spontaneously arise whenever suitable conditions are met would justify increased efforts towards its artificial synthesis and the search for extraterrestrial life forms.

3 Different views of life

> There is no spirit-driven life force, no throbbing, heaving, pullulating, protoplasmic, mystic jelly. Life is just bytes and bytes and bytes of digital information.
>
> (R. Dawkins, 1995, [24])

> Life does not depend on the magic of Watson-Crick base pairing or any other specific template-replicating machinery. Life lies ... in the property of catalytic closure among a collection of molecular species.
>
> (S.A. Kauffman, 1995, [21])

In his book *The Problems of Biology* (1986), John Maynard Smith writes 'There are two distinct ways in which living organisms may be viewed. One is of a population of entities which, because they possess a hereditary mechanism, will evolve adaptations for survival. The other is of a complex structure which is maintained by the energy flowing through it.'

The Darwinian, or evolutionary, view of life (represented in the above quotation by Dawkins) places emphasis on hereditary transmission of characters, adaptation and evolution. Extreme neo-Darwinists would define a living organism as *any entity with the properties of multiplication, variation and heredity* (this includes viruses but leaves out mules and bachelors). Research inspired by this evolutionary view has favoured the development of 'virtual worlds' and 'artificial life' systems in which mutations, natural selection and evolution are simulated, while the homeostasis of any individual organism, its interaction with the environment, and even its self-replication with its cumbersome details, may be taken for granted or even be altogether omitted.

For example, Richard Dawkins' *biomorphs* [25] are selected for survival by the user at each generation on the basis of arbitrary (e.g. aesthetic) criteria; the attention is focused on how a few genomic parameters can control the phenotype and on the relation between mutation frequency, selective pressure and overall evolution of 'species'. Similar, more recent, work includes Karl Sims' *Evolved Virtual Creatures* (http://web.genarts.com/karl/) and Thomas Ray's *Aesthetically Evolved Virtual Pets* (http://www.his.atr.jp/%7Eray/pubs/alife7a/index.html). Interactive biomorphs evolution can be played online (http://alife.fusebox.com/morph_lab.html).

The second view places emphasis on organisms as self-maintaining structures open to flows of energy and matter, far from thermodynamic equilibrium, and thus belonging to the class of *dissipative structures* as first defined by Ilya Prigogine and co-workers [14]. However, organisms

differ from other dissipative structures in their ability to exert control on the flux of energy and matter through them [26]. In other words, organisms have a *self* which other systems lack; it makes sense to write that an organism – even a very simple one – maintains it*self*, replicates it*self*, etc., whereas the same terms sound as just metaphors if applied to turbulent eddies or galaxies. Of course, what a *self* is, whether this feeling of being using a metaphor is appropriate, and what to do once the concept of self is applied to human-made machines, are among the most subtle, puzzling, and yet undecided questions in science.

Leaving aside the most subtle issues, let us just remark that this second view of life, called at times the ecological view, has inspired work centred on the principles and mechanisms by which a living organism can exist as such, maintain its integrity, and attain homeostasis, rather than on the way it can reproduce itself, transmit hereditary information, and evolve.

The ecological view has inspired the concept of autopoiesis [27]. An *autopoietic machine* is defined by Varela [28] as 'a homeostatic (or rather a relations-static) system that has its own organization (defining network of relations) as the fundamental invariant'. Studies based on this view have thus focussed on the way complex systems (e.g. collections of macromolecules and of the chemical reactions between them) may attain conditions of permanence of structure. This has led to such concepts as Manfred Eigen's *hypercycles* [29] and Stuart Kauffman's *autocatalytic sets* [21, 30–33].

To the two views of life identified by Maynard Smith, and discussed above, one might add here a third view, i.e. that which privileges the mechanisms and the logical and physical requirements for *self-replication*. In fact, this crucial step for life as we mean it lies midway between the *permanence* (homeostasis) of a living structure and the *transmission* of its informational content (genomic or not) through consecutive generations. This view of life basically lies behind the pioneering studies of von Neumann [7] and the many following attempts to design and understand self-replicating systems, either as *virtual*, mathematical or computer-simulated, entities [34, 36, 37] or as actual *physical* structures [38–40].

4 The beginnings of life on Earth

> Life will always remain something apart, even if we should find out that it is mechanically aroused and propagated down to the minutest detail.
>
> (Rudolf Virchow, 1855)

> The vital forces are molecular forces.
>
> (Thomas Henry Huxley, 1868)

In this section we will give (from a somewhat amateurish standpoint) a brief overview of past and current hypotheses concerning the appearance of life on our planet and of their implications for self-replication. More specialised texts should be seen for a more rigorous account and for a large bibliography [41–43]. Classic, but now outdated, books are those by Orgel [44] and by Miller and Orgel [45]. A highly interesting discussion of the changes in perspective concerning the difference between life and non-life in the last 50 years can be found in [46].

Advanced forms of life existed on earth at least 3.5 billion years ago [47]. Rocks of that age found in Western Australia show microfossils belonging to the Archaea phylum that look much like modern cyanobacteria (highly evolved photosynthetic organisms). There is indirect evidence (e.g. anomalous $^{13}C/^{12}C$ ratio in rocks 3.7∼3.8 billion years old) for an even older origin [42].

On the other hand, it is estimated that the planet was formed, together with the rest of the solar system, some 4.5 billion years ago and remained certainly inhospitable for life for at least half a billion years. The earliest known rocks are about 4 billion years old, and oceans were probably formed at the same time. This leaves a very narrow time window (0.5 billion years at most) for life to have appeared on earth once conditions compatible with it were established.

Much effort has been dedicated, using techniques like the comparison of ribosomal RNA, to the reconstruction of a credible tree of life and to the search for the *last common ancestor* of all currently living species. This was probably an intermediate form between bacteria and archaea, perhaps a thermophilic organism similar to the thermophilic bacteria that live today in hot springs, either terrestrial or oceanic [42].

Apart from unlikely suggestions, e.g. from Francis Crick [48] and Fred Hoyle [49], of an extraterrestrial origin for life, it is now widely accepted that it arose on earth, in a relatively short time, from inorganic matter.

Organic chemistry is, after all, 'just' the chemistry of carbon. The synthesis of urea from ammonium cyanate by Friedrich Wöhler in 1828 is usually regarded as the first demonstration that no special 'vital force' is needed for organic synthesis [50]. There is spectroscopic evidence that organic compounds, probably including amino acids, are present even in cosmic spaces.

In the early 1950s Stanley Miller, working under the direction of Harold Urey (the discoverer of deuterium and a winner of the Nobel Prize for Chemistry), was able to synthesise a variety of amino acids by sending electric sparks through a mixture of hydrogen, methane, ammonia and water vapour, simulating lightning through a hypothetical early earth atmosphere [5]. Since Miller's experiment, research in 'abiotic chemistry' has produced not only all known amino acids but also sugars and purine and pyrimidine bases, which are the constituents of nucleic acids such as RNA and DNA. It is widely accepted that spontaneous processes occurring on the early earth may well have produced – given sufficient time and a suitable variety of environments – the basic building blocks of life.

However, the jump from relatively simple molecules like amino acids and bases to actual living systems is still long. The hypothesis of a casual self-assembly of something as complex as a living system – even a simple bacterium, the simplest living organisms in existence today – is ruled out by the overwhelming size of the contrary odds. Actually, this sort of exercise is one of the favourite weapons of *creationists*, whose pamphlets (now replaced by as many web sites) abound in such calculations, misquoting Hoyle [49] in saying that the chance of a bacterium spontaneously self-assembling from simpler, amino acid – level, building blocks is about the same as 'the probability that a tornado sweeping through a junk yard could assemble a 747 from the contents therein'. Thus, for example, Dr. Art Chadwick of the Earth History Research Center at the Southwestern Adventist University, Keene, TX, tells us that 'Spontaneous origin of life on a prebiological earth is IMPOSSIBLE!', (his emphasis), leaving direct creation as 'the only reasonable alternative' [51].

Of course, the creationists' weapon is blunt, because no serious supporter of the thesis that life arose from the non-living is actually proposing that this happened by chance and in a single step.

However, a real difficulty is met here. Given that living and self-reproducing organisms exist, the Darwinian evolution mechanism of random mutation/crossover under the pressure of natural selection [52] is a powerful accelerator of transformations and will inexorably drive any *phylum* towards a greater differentiation and complexity. The mechanisms for this are too well known to deserve a further discussion here [25, 53]. But we are discussing how the first living organisms came into existence. Therefore, some mechanism of *prebiotic* evolution must be hypothesised in order to circumvent the 'overwhelming odds' objection.

In a nutshell, the idea is that evolution via random mutations and natural selection must have acted for a sufficiently long time on a primitive population of prebiotic beings, endowed with the

only ability to somehow reproduce themselves (even in a partial, approximate and imperfect way), self-replication being the necessary prerequisite for any evolution mechanism. As synthesised by Zindler [54], 'almost certainly, the principle of "survival of the fittest" is older than life itself'.

In their turn, the very first self-replicators (certainly macromolecules, or simple systems of macromolecules) must have emerged by predominantly random processes, which puts an upper bound to the possible complexity of any potential candidate to this role.

The most serious difficulty concerning prebiotic evolution is that in contemporary cells each of the three major types of biological molecules (proteins, RNA, and DNA) requires the other two for either its manufacture or its function [50]. In particular, proteins (enzymes) perform catalytic functions but cannot be manufactured without the information encoded in DNA (although this statement may require a partial revision following the discovery of some limited self-replication capabilities in peptides [33]). In its turn, DNA is merely a blueprint, which by itself cannot self-reproduce nor perform catalytic functions. In fact, DNA replicates almost, but not quite, without the assistance of an enzymatic 'machinery'; the process of filament cleavage, replication, and re-pairing is assisted by three enzymes (DNA endonuclease, DNA polymerase, and DNA ligase) which, in their turn, must be synthesised by a minimum of ribosome–mRNA–tRNA 'machinery'. Therefore, a 'minimum logistic system' is required for DNA replication.

Quoting de Duve [50], 'Considerable debate in origin-of-life studies has revolved around which of the fundamental macromolecules came first – the original chicken-or-egg question'; and McClendon [42] observes 'It must be recognized that, at the early time that metabolism started, enzymes (whether RNA- or protein-based) could not have had the high specificity found today. In fact, metabolism must have started non-enzymatically, using as catalysts amino acids, small peptides, clays, and the same basic cofactors (or their analogues) used today: pyridoxal, nicotinamide, pyrophosphate, thiamin, folic acid, metal ions, etc.'

On this issue, again, creationists don't give development a chance: '... both the DNA and proteins must have been functional from the beginning, otherwise life could not exist' [55].

At this point, it is important to assess more quantitatively what the *minimum* size of a self-replicating system ought to be. This actually implies two different questions, i.e.:

- What is the minimum number of distinct molecular species that can attain self-replication closure?
- What is the minimum size (in terms of mass or number of molecules) required for a self-replicating system to be sufficiently insensible to thermal noise and other disturbances from the environment?

To these a third, more subtle question might be added:

- What is the minimum size required for a population of self-replicating systems to evolve?

Useful hints may come from considering the smallest living systems existing today, as discussed, for example, by Zindler [54]. Apart from viruses, which do not have full self-replicating capabilities and depend on living cells for spreading their genome, the smallest living things today are probably the reproductive bodies of pleuropneumonia-like organisms (PPLO), which are $\sim 0.1\,\mu$m in diameter and contain (not counting water) $\sim$12 million atoms. Their DNA has a molecular weight of 2.88 million daltons and about 500 genes. On the whole, a PPLO elementary body contains $\sim$1200 distinct macromolecular species (DNA, RNA and proteins). Zindler mentions that corresponding theoretically minimum values to keep a living cell 'running' should be about 1.5 million atoms, distributed among a few hundred chemical species, and a DNA of 360,000 daltons with some 200 genes.

The above figures, although far smaller than those holding for more common, 'large' living cells (including bacteria), are still far too large to encourage the hope that such organisms can be close to the first life forms appeared on earth. Organisms of this complexity must be themselves the product of a long and tortuous evolution. First life must have arisen in a different, and much simpler, form.

5 Models of biogenesis: glimpses of the truth or just-so stories?

> The hyena was the most beautiful of jungle animals. He terrorized the tortoise by carrying him up into a tree and leaving him balanced on a high branch. The leopard rescued the tortoise, who rewarded him by painting beautiful spots on his plain yellow coat.
>
> (Forbes Stuart, *The Magic Horns: Folk Tales from Africa*, 1974)

The different 'views of life' discussed in a previous section may affect the preference of different researchers for one or another characteristic of life as the most likely to have appeared first on earth [26]. Thus, a logical priority would become a chronological one – which is not necessarily a good move.

In particular, in the 'replicator first' approach, the likeliest ancestors are self-replicating molecules, appearing before much organisation could develop [25, 56], whereas in the 'metabolism first' approach they are collectively autocatalytic, self-sustaining chains of chemical reactions between organic molecules, none of which by itself is capable of self-replication [21, 57, 58].

Let us start by considering the 'replicator first' approach.

The possibility of different types of molecules (at least two, one informational and one catalytic) having appeared and evolved together appears as highly unlikely to many scientists, who have looked for alternative scenarios.

Among such scenarios is one in which a single biomolecule could perform multiple functions and attain the ability of self-reproducing. The most likely candidate seemed to be RNA. In the late 1970s Sidney Altman at the Yale University and Thomas Cech at the University of Colorado at Boulder discovered RNA molecules that could excise portions of themselves or of other RNA molecules [50]. It appeared (at least theoretically) possible that an RNA molecule could have contained enzymatic sequences catalysing the process of self-reproduction via reciprocal base pairing (a process which cannot occur without the assistance of appropriate enzymes). An enzyme-free self-replication system was demonstrated in 1986 by von Kiedrowski (reported by Cousins *et al*. [59]) and employed the hydrogen bond donor–acceptor pattern between nucleotides.

It was the Harvard chemist Walter Gilbert who coined in the 1980s the term 'RNA world' to designate the hypothetical stage in the appearance of life, in which RNA replicators evolved before other biomolecules such as proteins and DNA came into play [60]. Since then, the proposal has enjoyed a wide popularity [61, 62].

A hypothetical, but not too far-fetched, sequence of steps which, in this 'RNA first' view, may have led to life as we know it today includes

- appearance of first nucleotides from even simpler organic molecules as building blocks, and their assembly into RNA;
- development of RNA replication;
- development of RNA-dependent protein synthesis, and appearance of first membrane-bound compartments (protocells);
- appearance of protein enzymes, and development of the 'modern' genetic code involving the simultaneous cooperation of DNA, RNA and proteins.

A few crucial points, in this as in any other evolutionary model of the appearance of life, are worth pointing out:

- at each step, including the earliest, prebiotic ones, natural selection must have played an essential role, while chance alone would have been sadly inadequate to pick up any successful molecular combination;
- as already noted by von Neumann [7], any replicator, once it appears even in a single copy, will rapidly spread and become a dominant molecular, prebiotic, or protobiotic, species;
- the power of self-replication alone is not unlimited; '. . . the idea of a few RNA molecules coming together by some chance combination of circumstances and henceforth being reproduced and amplified by replication simply is not tenable. There could be no replication without a robust chemical underpinning continuing to provide the necessary materials and energy' [50].

The 'RNA world' model has been a useful paradigm, but it is not without problems and its popularity seems now to be declining. On one hand, attempts at engineering an RNA molecule capable of catalysing its own reproduction have failed so far. On the other hand, RNA itself cannot have arisen by chance, the contrary odds being still overwhelming; its appearance must have been preceded by a phase of 'abiotic' chemistry capable of providing the RNA building blocks steadily and robustly. Also these hypothetical steps (protometabolism) have not been satisfactorily reproduced in the laboratory so far.

In contrast with the 'RNA first' hypothesis, Dyson [58] and Shapiro [63] maintained that only proteins must have existed for a long time, and must have attained some kind of non-genomic metabolism and reproduction, before first RNA, and then DNA, developed. Ghadiri and co-workers showed that a 32-amino acid peptide, folded into an α-helix and having a structure based on a region of the yeast transcription factor GCN4, can autocatalyse its own synthesis by accelerating the amino-bond condensation of 15- and 17-amino acid fragments in solution [64]. These studies also revealed the emergence of symbiosis and error correction in small 'ecosystems' of self-replicating peptides.

In the early 1980s Cairns-Smith made quite a cry (which found its way well into the general press) when he proposed that, long before proper cells could develop, clay (i.e. silicate molecules) may have exhibited rudimentary self-replication abilities and may have provided the 'scaffolding' for the further development of organic (carbon-based) replicators [56, 65, 66]. He argued that clay crystals cannot only replicate, but can even transmit information from one crystal generation to the next. Crystal defects, the analogues of mutations, can be passed on from parent to daughter crystals; thus natural selection can operate on a population of clay crystals, and favour those configurations which happen to bind organic molecules capable of stabilising their micro-environment and of enhancing their chance of survival and reproduction [54].

Also the 'clay first' hypothesis has lost much of its appeal in the last years, although it has never been quite disproved.

Let us now consider the 'metabolism first' approach. A useful track is provided by Casti [18, 19].

The Russian chemist Alexander Ivanovich Oparin was probably the first scientist who strived to devise a realistic mechanism for the rise of early life [67, 68]. He proposed that this mechanism involved the formation of *coacervate* droplets, colloidal particles which form when certain organic molecules associate with one another and precipitate from a solution. Complex coacervates, in particular, are those that form between two different types of macromolecules, for example between a basic, positively charged protein (e.g. gelatine or serum albumin) and an acidic, negatively charged carbohydrate (e.g. gum arabic or sodium alginate). Although coacervates are not technically alive and contain no genetic material, they do have a membrane-like surface layer, can accumulate more organic materials inside themselves (feed), and can even

divide by budding [69]. In such particles, according to Oparin, proteins endowed with catalytic properties (enzymes) managed to develop, perhaps for a long time, before proper genes made their appearance. Oparin reputed coacervates to be the progenitors of living cells and defended this view for over 40 years, from the first Russian publication in 1924 to his latest contributions to international conferences in the 1960s.

Apparently Oparin was the first to use the now popular term 'primordial soup' to denote the compound-rich water environment in which life is often supposed to have emerged. In about the same years, similar ideas were put forth in Europe by J.B.S. Haldane [70].

Sidney Fox of the University of Miami believed that the first life precursors were *proteinoids*, or *thermal proteins* [71–73]. These are short and branching peptide chains (oligopeptides, whereas proteins are much longer linear chains, or polypeptides proper), which are formed when dry amino acid mixtures are heated under appropriate conditions. They possess a number of catalytic properties, including the ability to catalyse the formation of nucleic acids (RNA and DNA) and of other proteinoids as well (autocatalysis). In water, proteinoids spontaneously form microspheres which, like Oparin's coacervates, exhibit a rudimentary 'metabolism', can grow by accretion and can proliferate by fission and budding for several 'generations', thus attaining, albeit imperfectly, *self-reproduction* of a sort. Quoting Enger and Ross [74], 'proteinoid microspheres are "double boundary" structures, which exhibit some membrane-like characteristics. They swell or shrink depending upon the surrounding solution and contain some proteinoids that function as enzymes. Using ATP as a source of energy, microspheres can direct the formation of polypeptides and nucleic acids. They can absorb materials from the surrounding medium and form buds, which results in a second generation of microspheres. Given these characteristics, . . . microspheres can be considered as protocells, the first living cells.'

Fox and a few co-workers [75, 76] have been for decades strenuous and rather vociferous supporters of their 'proteinoid protocell' theory, but, as for Oparin's coacervates, one has the impression that laboratory evidence is far too scanty for such enthusiasm and that the true story must have been more complex and twisted than that.

The Japanese researcher Fujio Egami discovered that, by adding simple compounds such as formaldehyde and hydroxylamine to seawater enriched with trace elements such as iron, zinc and molybdenum, it was possible to obtain the formation of amino acids, lipids and other organic molecules and the formation of lipid-bounded particles, which he dubbed *marigranules* [77]. Like Fox's proteinoid microspheres, marigranules were found to be capable of metabolism of a sort, including growth and undisciplined reproduction.

Wilcox [69] wrote an entertaining account of coacervates and protocells as related to the hypothetical '*bions*' that Wilhelm Reich claimed to have discovered [78]. For a demolition of Reich's '*bions*' see also Martin Gardner [79].

In the 1970s Manfred Eigen, 1967 Nobel laureate in Chemistry, introduced the concept of *hypercycle* [29, 80]. A hypercycle is a system of molecules and chemical reactions which exhibits, as a whole, autocatalytic properties: some of the molecules, say A, catalyse the formation of other compounds B, which in turn, via a more or less long and circuitous route, eventually catalyse the formation of A. Such a system is capable of sustaining itself and may exhibit homeostasis, growth and some rudimentary self-replication ability, i.e. some of the characters of life. Eigen proposed that such hypercycles, of increasing complexity, may have been the precursors of life and may have forerun the appearance of modern proteins and cells. The concept was subsequently borrowed and developed by many other scientists, including John Maynard Smith [81].

The collectively autocatalytic sets of Stuart Kauffman [21, 30, 31] are basically the same thing as Eigen's hypercycles. Both are systems exhibiting molecular replication in which the products as a whole serve as catalysts for their own synthesis [82, 83]. A living cell is in fact such a collectively

autocatalytic system, although in it no individual molecule, including DNA, catalyses its own formation. Most of the cell's catalysts (enzymes) are proteins, so that it is mainly on collective autocatalysis in complex peptide systems that attention has been focussed.

Kauffman [33] describes collective autocatalysis as follows: 'The first step ... is to construct self-reproducing cross-catalytic systems of two peptides, A and B; here A catalyses the formation of B from B's two fragment substrates, and B does the same for A. Such a system would be collectively autocatalytic – no molecule would catalyse its own formation, but the system would collectively catalyse its own formation from 'food' sources – the two A fragments and the two B fragments. If collectively autocatalytic peptide sets with two catalytic components can be constructed, can systems with three, four or hundreds of cross-coupled catalysts be created?'

One of the main arguments used by Kauffman to advocate the origin of life from collectively autocatalytic sets of polypeptides is of a *combinatorial* nature. The number of the possible chemical reactions between polypeptides grows exponentially as the number of chemical species present in a given environment is increased. At some point, it is almost inevitable that a chain of chemical reactions will appear, which is catalytically (enzymatically) promoted by its own products. When this happens, the reaction rates and the amount of products will increase and, provided a sufficient amount of raw matter is available, will attain significant levels at the expense of the other molecular species which do not partake of this 'hypercycle'. Thus, initially amorphous matter will end up with organising itself into complex networks of reactions: *complexity* will arise spontaneously in 'primordial soups' of polypeptides and other organic molecules.

As was mentioned in Section 2, Kauffman's work, which has enjoyed considerable popularity in the media during the last years, is a typical example of the school of thought associated with the study of complexity and having its headquarters at the Santa Fe Institute in New Mexico. His work is closely connected with the concepts of self-organised criticality and of systems at the edge of chaos, developed in recent years by authors like Norman Packard [84], Chris Adami [85] and Per Bak [86].

6 Information aspects of life, self-replication and evolution

> Life seems to be an orderly and lawful behaviour of matter, not based exclusively on its tendency to go from order to disorder, but based partly on existing order that is kept up.
>
> (E. Schrödinger, 1944, [10])

A 70-kg human body contains $\sim 10^{27}$ atoms chosen among 92 elements. At $\ln_2(92) \approx 7$ bits/atom, a complete description of the human body requires $\sim 10^{28}$ bits. But our DNA codes only $\sim 10^{10}$ bits of information. The missing bits of information (to put it mildly) come from the 'substrate', ranging from the laws of physics and chemistry to the actual structure of the Universe and of the Earth, the composition of the maternal ovum, uterus and blood, and of course the environment where the body grows [87].

The fact that metabolism, growth, reproduction and evolution in living beings are all processes that draw from a vast ocean of order and information, provided by the regularity and physical laws of the surrounding world, was already present in the work of D'Arcy Thompson [9], which has been a source of inspiration for the subsequent work on morphogenesis by Conrad Waddington [88] and René Thom [8].

The importance and the nature of information processing in living systems were also quite clear to Erwin Schrödinger [10]. The outstanding pioneering role played by this great physicist in the studies on the origin and ultimate nature of life has been recently acknowledged, for example, by Prigogine [89] and is worth a brief digression.

The lectures published under the title *What is Life?* [10] were given while he was the director of the Institute for Advanced Studies in Dublin. They are not the casual musings of an amateur biologist, but a purpose directed effort by a great scientist towards unravelling the more mysterious aspects of life [19].

According to Schrödinger, the main peculiarity of life with respect to the 'normal' physical systems described by the physical sciences, is that very small entities (molecules of the then unknown genetic material) play an individual role, so that statistical mechanics does not apply (at least as far as the transmission of the genome is concerned). This intuition anticipates by half century a modern concept put forward by Kauffman [21, 32, 33] and other authors: the number of possible proteins is so large that the Universe can have explored so far only a tiny fraction of these possibilities, thus remaining far short of attaining *ergodic* conditions.

Schrödinger believes the gene stuff to be some protein, which is a slight mistake indeed for someone writing in 1944. Again his intuition, many years before Watson and Crick's discovery of the double-helix DNA structure, leads him to imagine it as some 'aperiodic solid', with chemical bonds strong enough to survive thermal noise (we now know that the nucleotides in DNA are actually linked by strong, covalent bonds). Schrödinger uses the term 'miniature code' for the genome.

It is truly remarkable how some very simple physical considerations allow Schrödinger to make a number of estimates concerning the genetic substance, which have been since proved to be basically correct. For example:

- On the basis of studies by Delbrück, who measured the rate of radiation-induced mutations in the living matter, the volume of a single gene was estimated to be of the order of 10^{-27} m^3, corresponding to 10^3 Å^3 and thus to $\sim 10^3$ atoms. This we now know to be the correct order of magnitude for a typical sequence of nucleotides identifiable as a gene.
- The mean survival time to thermal noise at temperature T ($\sim$300 K) is of the order of $\tau = \tau_0 \exp(E/kT)$, in which τ_0 is the period of molecular vibrations (estimated to be of the order of 10^{-13}–10^{-14} s by analogy with known organic compounds), and E is the binding energy (estimated to be 1–2 eV as typical for covalent bonds). Intermediate assumptions yield $\tau \approx$ 15,000 years, which is the right order of magnitude required for the permanence of the genome over long stretches of time (note that this is the lifetime of a gene, which usually far exceeds that of an individual).

The 1944 booklet also deals with the problems of mind and consciousness. Interestingly, despite his own involvement in quantum mechanics (and what an involvement!), Schrödinger explicitly *denies* a direct role of quantum effects in these phenomena: 'To the physicist I wish to emphasise that in my opinion, and contrary to the opinion upheld in some quarters, quantum indeterminacy plays no biologically relevant role . . .'. He *does* mention quantum theory many times, but mainly to emphasise the necessarily discrete nature of configuration changes in the genome molecules, and thus to explain the apparently puzzling ability of such tiny objects to persist unchanged through the years and the generations.

Therefore, the enthusiasm showed by Roger Penrose (a scientist who *does* believe in a strong role of quantum indeterminacy in mind and consciousness) in his foreword to the 1991 edition of Schrödinger's work, seems a bit over the lines: '. . . how often do we still hear that quantum effects have little relevance in the study of biology . . .'.

We will come back to the issue of mind and consciousness (including the rather extreme and pantheistic view supported by Schrödinger) in Section 16, in the context of discussing the role of symbolic representation in life and evolution.

Schrödinger did not yet use the word information for the 'thing' carried by genes, but rather terms like 'number of determinations' and the like (Shannon's work was still to come). However, he already identified this 'something' with *negative entropy* and showed how living beings 'feed' on negative entropy drawn from the environment (in the form of highly organised proteins or, eventually, of sunlight captured by plants).

Of course, information cannot be produced out of thin air, but it can be drawn from the environment and incorporated into the living matter without violating the principles of thermodynamics – provided, of course, that a source of energy is available. Thus a living being can be regarded from the thermodynamics point of view as an open, dissipative system, far from equilibrium and crossed by a continuous flow of energy and matter [14].

An important mechanism which increases the informational content of a living organism (e.g. a cell) is *protein folding*. By folding into its three-dimensional shape, a protein converts a linear string of symbols (a point in a relatively small phase space of one-dimensional strings) into a spatial configuration (a point in a far larger space of three-dimensional configurations). In doing so, the protein effectively *enhances* the original information content of the coding string (drawing from the vast ocean of negative entropy provided by the physico-chemical laws that govern polymer shapes).

As far as the metabolism of individual organisms goes, the energetic cost of this information gathering is relatively low: in fact, information and thermodynamic entropy are related through the very small Boltzmann constant k, so that, in thermodynamic units, information is usually cheap – this is the ultimate reason why we can throw yesterday's newspaper without giving it a second thought, or print a 20-page manuscript just to check for typos.

Also evolution can be regarded as a source of information. The rationale for this is well synthesised by Chris Adami, the creator of the ALife program 'Avida': complexity must increase in evolution in a fixed environment, simply because mutations under natural selection operate like a natural Maxwell Demon, keeping mutations that increase the information about the environment, but discarding those that don't [90].

In the case of evolution, however, the gathering of information regarding the environment is achieved at the cost of 'trying' and sacrificing huge numbers of 'unfit' individuals – a cost which can be considerable indeed also from the point of view of thermodynamics. The point, of course, is that no better solution – no shortcut – is available to Nature!

Ji [91] advocates a view of the evolution of the universe in which a progressive decrease of energy density has been followed, after the emergence of the first self-replicating systems, by a progressive increase of information density (amount of biological information divided by the volume of the biosphere).

It should be kept in mind that this evolutionary increase of information is mainly associated with the increase of biological diversity; it does not necessarily amount to an increase in complexity of the life forms, and does not justify the popular view of evolution as progress [92]. What is about certain is that early evolution was accompanied by a progressive increase of the distance between the genome and the external environment.

The arguments developed in this section suggest that much of the periodically re-surging *querelle* regarding the ability of the genome alone to describe something as complex as an actual living organisms, and the ability of Darwinian natural selection alone to explain the diversity and richness of current life, is misplaced and outdated. Even the most zealous neo-Darwinists accept that natural selection is not the full story, and is *not* sufficient for a full explanation of life on earth. 'Life is embedded in the physical world, and the rules of physics and chemistry must surely impose important constraints on the evolutionary process . . . we cannot understand

life purely through studying the genes, but we must also think about their interactions with the physical medium in which they are embedded' [26].

The great geneticist Conrad Waddington remarked that: 'The hereditary differences which arise in animals are not quite random, like the differences between two heaps of bricks. They result from changes in orderly systems of development, and each new variety has an order of its own, maybe less, but sometimes more, complex than that of the original form from which it was developed' [88]. Along similar lines, Richard Belew and Melanie Mitchell observed that 'common metaphors for the genes (e.g. programs, blueprints, etc.) as the information-carrying component in evolution all rest on an inappropriate "preformationist" view of information, as if information . . . exists before its utilization or expression' [93]. In all cases, 'the genome does not have to encode information about every aspect of the adult organism's design, because some features will just fall into place "for free" as the developmental process unfolds, due to the self-organisational properties of the constituent matter' [26].

Despite the expectations of creationists and other opponents of rational thought, always eager to thrust their crowbar into the crevices that any scientific theory inevitably presents, these limitations have nothing to do with the overall validity of evolutionary theories, and with the possibility of achieving features characteristic of life in artificial systems.

7 Virtual worlds

> You know, the universe is the only thing big enough to run the ultimate game of life. The only problem with the universe as a platform, though, is that it is currently running someone else's program.
>
> (K. Karakotsios, reported in [26])

Much of the recent research on self-replication has been conducted using symbolic systems (virtual worlds). This approach is, of course, a drastic simplification of the real state of affairs, and cannot account for the problems encountered when coping with real-world physical laws. On the other hand, by disposing of these complications, it allows one to concentrate on the logic and symbolic aspects of the problem and may give useful indications for the development of more complex, real world, strategies.

We will give here a brief account of alternative symbolic systems which have been used in self-replication studies. Actual self-replication models developed in such 'virtual worlds' will be presented in the following sections.

7.1 Cellular automata

Wolfram [94] defines cellular automata (CAs) as follows: 'Cellular automata are simple mathematical idealizations of natural systems. They consist of a lattice of discrete identical sites, each site taking on a finite set of, say, integer values. The values of the sites evolve in discrete time steps according to deterministic rules that specify the value of each site in terms of the values of neighbouring sites. Cellular automata may thus be considered as discrete idealizations of the partial differential equations often used to describe natural systems. Their discrete nature also allows an important analogy with digital computers: cellular automata may be viewed as parallel-processing computers of simple construction.'

CAs have been used since the pioneering work of von Neumann as instruments to investigate the conceptual problems of self-replication [95, 96].

A 32-bit Windows-compatible freeware program called Mirek's Cellebration (MCell), which allows playing hundreds of CA rule sets and hundreds of patterns, was presented by Mirek Wojtowicz [97]. It can play rules from 14 of the most popular CA families including Conway's 'Life', 'Generation', and many variants and generalisations by Rudy Rucker and other 'inventors'. A Java applet (MJCell) is also available. The web page cited contains a detailed explanation of rules and patterns.

A large family of CAs are based on a two-dimensional 'chequerboard' (cell space), with each square, or 'cell', having just two states, say ON (alive) and OFF (dead) and the state of each cell at generation $t + 1$ being determined only by

- the state of the cell considered, and
- the number N of ON cells in the 8-cell neighbourhood surrounding it,

at generation t. In this case, the transition rules can be summarised as *sIJK. . ./bPQR. . .*, meaning that an ON cell will remain ON (survive) if $N = I$ or J or $K \ldots$ and an OFF cell will turn ON (be born) if $N = P$ or Q or $R \ldots$

Probably the best known, and certainly one of the most interesting, CAs of this family is Conway's Life [98], defined in the above terms by the transition rule *s*23/*b*3. Whole books have been dedicated to Life [99] and it is too well known for us to dwell on its amazingly rich behaviour here. Among the many CA computer programs implementing Conway's Life, a very fast-running one suitable for Windows and freely downloadable from the author's web site is Life32 by Johan Bontes [100]. Data sets containing hundreds of interesting patterns have been put together by Alan Hensel, Paul Rendell, David Bell and other Life fans and can also be freely downloaded. Some of these patterns implement logical functions such as memory registers, latches, logical gates and even a Turing machine [101], and are connected with the idea of implementing a self-replicating pattern within the rules of Life, as will be discussed later.

7.2 Core wars, viruses, quines: self-replicating computer programs

An approach alternative to CAs is that based on *self-replicating computer programs*. These are best exemplified by Core Wars [102], a game in which two or more processes fight for control of the resources (memory and processor cycles) of an environment (*coreworld*). Apparently it was originally invented as early as in the 1950s by employees of the Bell Laboratories.

The instruction set for Core Wars, Redcode, is a limited, but computationally complete, assembly language. Core Wars programs, or *warriors*, are like a genotype, while the corresponding phenotype is the *behaviour* that the processes realise when they are placed in a Core Wars simulator, or *arena*. A warrior can copy itself and then split to the copy, which is much like cell division, or jump to the copy, which is more like a digital equivalent of movement. A process can cause another process to stop executing ('kill' it), which is somewhat like biological predation. Furthermore, some of the more complex programs have displayed such abilities as setting a trap, self-repair, and mimicry, all of which have biological equivalents.

Core Wars can provide a wide variety of artificial life experiments. Steen Rasmussen and co-workers [103] altered the rules slightly and came up with a self-contained evolutionary environment, sort of a binary version of 'primordial soup' experiments. A potential result of this type of experiments would be to determine whether reproduction or metabolism tends to evolve first within a random coreworld.

Core Wars programs are also the ancestors of modern computer viruses. Rob Rosenberger, a computer consultant who maintains the Computer Virus Myths homepage, thus summarises the origin of computer viruses in the *Scientific American* – Ask the Expert web page:

> Fred Cohen presented the first rigorous mathematical definition for a computer virus in his 1986 Ph.D. thesis. Cohen coined the term 'virus' at this point and is considered the father of what we know today as computer viruses. He sums it up in one sentence as 'a program that can infect other programs by modifying them to include a, possibly evolved, version of itself.'... The media's perception of viruses took a dramatic turn in late-1988, when a college student named Robert T. Morris unleashed the infamous 'Internet Worm' ... Reporters grew infatuated with the idea of a tiny piece of software knocking out big mainframe computers worldwide. The rest, as they say, is history.

A more rudimentary example of digital self-replication is offered by *quines*. A *quine* is a sentence that both mentions and uses a phrase, like, for example:

'*Is not the title of any book*' is not the title of any book

The name is a tribute to the logician Willard van Orman Quine and was introduced by Douglas Hofstadter [34]. A *quine* is also a computer program that generates a copy of its own source code as its output. Writing *quines*, and finding the shortest possible *quine* in a given programming language, is a common hackish amusement. Two examples in Fortran, collected by Gary Thompson [104], are reported below (these work in Microsoft Developer Studio, the most popular current version of Fortran for Windows):

```
      WRITE(6,100)
      STOP
100   FORMAT(6X,12HWRITE(6,100)/6X,4HSTOP/
     .42H 100 FORMAT(6X,12HWRITE(6,100)/6X,4HSTOP/ ,2(/5X,67H.
     .42H 100 FORMAT(6X,12HWRITE(6,100)/6X,4HSTOP/ ,2(/5X,67H.
     .)/T48,2H)/T1,5X2(21H.)/T48,2H)/T1,5X2(21H)/
     .T62,10H)/6X3HEND)T1,5X2(28H.T62,10H)/6X3HEND)T1,5X2(29H)/6X3HEND)
      END
```

(author: Armond O. Friend);

```
      REAL*8F(17)
      DATA F/8H(7X,11HR,8HEAL*8F(1,8H7)/7X,7H,8HDATA F/,,8H5(2H8H,A,
     18H8,1H,)/6,8HX,1H1,6(,8H2H8H,A8,,8H1H,)/6X,,8H1H2,5(2H,8H8H,A8,
     1H,28H,),2H8H,,8HA8,1H//7,8HX,11HWRI,8HTE(6,F)F,8H/7X,3HEN,8HD) /
      WRITE(6,F)F
      END
```

(author: Mike Duffy).

More elegant *quines* can be obtained using other languages, e.g. Lisp. It is possible to write a program which outputs another program which is itself a *quine*, and some people have had the idea of creating a sequence, $p[0] \cdots p[n]$, of programs, such that the output of $p[i]$ is the source for $p[i+1]$ $(0 \leq i < n)$, while the output of $p[n]$ is the source for $p[0]$.

7.3 L-systems

Closely connected with CAs and with self-replicating computer programs are the so-called 'Lindenmayer systems', or L-systems [105]. These were originally conceived by the Dutch

biologist Arstid Lindenmayer as a mathematical theory of plant development [106, 107]. The central concept of L-systems is *rewriting*, a technique for building complex objects by successively replacing parts of an initially simple object according to a set of rewriting rules, or *productions*. Without much loss of generality, one can assume that the rewriting rules operate on *character strings* and generate complex strings starting from very simple ones (*axioms*) after a number of intermediate steps (*derivations*). Of course, the use of the terms axioms and derivations betrays the possible interpretation of L-systems as *formal languages*. *Context-sensitive* productions take into account not only the predecessor character, but also its *environment*, and thus may account for *interactions* between different parts of a string.

L-systems are also closely related with Noam Chomsky's work on formal grammars [108], where the concept of rewriting is used to model the syntactic features of natural languages, and with Douglas Hofstadter's *Typogenetics* [34], which is basically a character-string model of the translation of genes into proteins (see below).

L-systems are usually coupled with a graphical interpretation. The most popular is the so called *turtle interpretation*, in which the string produced by the L-system is interpreted as a sequence of commands to a cursor, or *turtle*, moving in a two-dimensional (computer screen) or even a three-dimensional virtual space.

Thus, L-systems exhibit two noteworthy properties:

- they give rise to growing, one-dimensional strings of characters;
- these strings can be interpreted as two- or three-dimensional structures, or images.

Stauffer and Sipper [35] explored the relationship between L-systems and CAs and showed how L-systems can be used to specify *self-replicating structures*. In particular, they showed that:

- a set of rules can be chosen in an L-system so that an initial string is replicated after a number of derivations;
- the coupled graphical interpretation is that of operations in a two- or three-dimensional cellular space, thus transposing the self-replication process onto a CA context.

The authors applied these concepts to the L-system description of a self-replicating 'loop' of the kind introduced by Langton [36], and pointed out the analogy between the transition from one-dimensional strings to multidimensional CAs and that from a one-dimensional genotype to multidimensional proteins, or phenotypes.

7.4 Typogenetics

Typogenetics is a formal system designed to study origins of life from a 'primordial soup' of DNA molecules, enzymes and other building materials. It was introduced by Douglas Hofstadter in his prized book *Gödel, Escher, Bach: An Eternal Golden Braid* [34]. Typogenetics is a simple model of molecular genetics on the level of DNA replication, ignoring questions about low-level chemistry and the high-level biological phenomena.

The typogenetical alphabet consists of only four characters, A, C, G and T, arranged in one-dimensional strings, or strands. The characters are called bases and their positions in the string, units. The terms *gap* or *blank* are used to indicate no base in a unit position. An example strand is CCAGTTAA, in which base G occupies unit 4.

A and G are called purines and C and T pyrimidines. Purines and pyrimidines are complementary, so that copy operations will copy A into T and G into C (and vice versa).

There are also 15 typogenetic amino acids, examples of which are *cut*, *del*, *cop*. Amino acids are coded for by couples of bases. There are 16 of these (AA, AC, . . . , TT), but the couple AA

does not code for any amino acid and plays the role of a punctuation mark. Sequences of amino acids are called enzymes.

Translation is a non-destructive process by which units called ribosomes crawl along the strand, associate to each pair of bases the corresponding amino acid following the above code, and assemble the resulting amino acids into enzymes. The punctuation mark AA codes for the end of an enzyme. For example, the strand CGCTAATAAGT translates to the enzymes *cop–off* and *rpy–del*. Small differences have been introduced by different authors.

After they are formed, enzymes fold into a secondary structure. This is obtained by writing the enzyme as a linear chain of amino acid blocks and then giving each block a characteristic *kink* ($\leftarrow$, $\uparrow$, $\rightarrow$, $\downarrow$, depending on the specific amino acid).

Enzymes are now let loose to operate on the original base strand or on other strands. The orientation of the link between the last two amino acids determines to which base of the strand it will bind. Once the enzyme is bound to a specific site in a strand, each of its individual amino acids performs some function; for example, *cop* turns on the copy mode and inserts the complementary base above the currently bound unit; *del* removes the base at the currently bound unit, and moves the enzyme one unit to the right; and so on. The final result of the action of the enzymes is that the original strand is transformed into one or more different strands.

Of course, this is a simplified version of the genetic code and of biological processes, in which amino acid sequences (proteins) are assembled in the ribosomes following the instructions coded in the DNA and with the aid of messenger and transfer RNA. All intermediate steps involving RNA are omitted; the number of amino acids is simplified to 15 in lieu of 20, and these are coded for by pairs of bases, rather than by triplets (codons); finally, enzymes (proteins) fold in two dimensions rather than three.

The most interesting applications of typogenetics as related to self-replication involve the application of a set of enzymes to the same strand that originally coded for these enzymes. As suggested by Hofstadter himself, this would mean something along the following lines:

> A single strand is written down. [This is translated] to produce any or all of the enzymes which are coded for in the strand. Then those enzymes are brought into contact with the original strand, and allowed to work on it. This yields a set of 'daughter strands'. The daughter strands themselves [are translated] to yield a second generation of enzymes, which act on the daughter strands; and the cycle goes on and on. This can go on for any number of stages; the hope is that eventually, among the strands which are present at some point, there will be found two copies of the original strand (one of the copies may be, in fact, the original strand). [34]

Actual work along these lines has been presented, for example, by Morris [110], Varetto [111] and Snare [112].

7.5 Virtual chemistry

John Holland [113] developed a theoretical framework, dubbed α-universe, where self-replicating structures capable of heritable variations could arise and interact. A α-universe is a one-dimensional CA space whose cells represent either elements (modelling physical entities, e.g. molecules) or codons encoding elements; starting from an initial configuration, each step in the evolution of the cellular automaton involves the application of a number of operators such as bonding, moving, copying, and decoding.

Such a computer-created virtual world in which particles move around and spontaneously engage in reactions whereby new particles or compound molecules may be formed, leading to further reactions, has been dubbed *virtual chemistry*. The challenge is formulating a virtual

chemistry in which simple *autopoietic agents* can develop. Since the early work of Varela *et al.* [27], these are defined as structures capable of maintaining their own integrity and organisation despite, or because of, turnover in their material components. Such agents would be intrinsically capable of self-repair and even self-replication. 'Successful synthesis of such autopoietic agents would provide considerable insight into basic scientific problems of the origin and organisation of the simplest living organisms; it would also provide a basis for utterly new kinds of technological innovation' (http://www.eeng.dcu.ie/∼alife).

Holland made several estimates of the expected time requiring for the spontaneous emergence of a self-replicating pattern. Unrealistic times (e.g. 10^{43} time steps) were obtained if full self-replication was required, whereas the emergence of partial self-replication was expected within times of the order of 10^8–10^9 steps.

The nature of α-universes likens them to the above-mentioned 'L-systems' and 'typogenetics'.

8 Von Neumann's self-replicating automata

> By axiomatizing automata in this manner one has thrown half the problem out of the window – and it may be the more important half.
>
> (J. von Neumann, 1966, [7])

When John von Neumann started exploring the possibility of a self-reproducing automaton in the 1940s [16, 109, 114, 115], his ambitious requirements for this hypothetical automaton were:

- *universal computation*, i.e. the automaton should have been able to perform arbitrary computations by following proper instructions;
- *universal construction*, i.e. the automaton should have been able to build arbitrary manufacts by following proper instructions;
- *self-reproduction*, i.e. the automaton should have been able to build, in particular, a copy of itself.

The concept of universal computation meant, in practice, that the automaton had to include as a proper subset of itself a *universal Turing machine*, the prototype of the modern computer envisaged in purely theoretical terms by Alan Turing [116]. Universal construction can be regarded as the constructive equivalent of universal computation.

In order to design such a self-replicating system, Von Neumann explored different models [117, 118].

The first self-reproducing machine he imagined (so called 'kinematic' model) would sit in a large stockroom filled with the parts from which it was built (arms, legs, eyes, circuit boards, etc.). The machine would have a memory tape containing the instructions needed to build a copy of itself from these spare parts. Using its robot arm and its ability to move around, the machine would find and assemble parts; the tape program would instruct the device to reach out and pick up a part, look to see if it was the right one, and if not, put it back and grab another. Eventually the correct one would be found, then the next, and the two joined in accordance with the master checklist.

The machine would continue to follow the instructions until it would end up with assembling a physical duplicate of itself. The new robot would be 'uneducated', i.e. it would have no instructions on its tape; the parent would then copy its own memory tape onto the blank tape of its offspring. The last instruction of the parent's tape would be to activate its progeny.

The development of this hardware-based approach met with severe difficulties, basically related with finding, reaching and manipulating parts. Von Neumann assumed these difficulties to be *inessential* to the most crucial issues of self-replication, and thus merely distracting, and looked for alternative approaches.

A 'neuron-type' machine, a 'continuous' machine, and a 'probabilistic' machine [87] were given some consideration. In the end, however, von Neumann turned to a purely 'virtual' model, based on the concept of CAs, in which parts are built 'out of thin air' whenever required, thus avoiding the inessential trouble of getting them out of the stock. Stanislav Ulam seems to have had a major role in directing von Neumann's attention towards this direction.

The most important and universal of the problems tackled by von Neumann was avoiding the *infinite regress* problem which arises when a machine has to build a copy of itself. In order to do so, the machine must possess a description (implicit or explicit) of its own layout (a 'blueprint'). But then the blueprint is part of the machine; therefore, it would seem that it has to include a description of itself (a 'meta-blueprint', i.e. a blueprint of the blueprint), and so on.

The problem can also be stated as follows: *it would seem that machines should necessarily be superior, in size and in organisation, to their output*. But then a machine cannot have a perfect copy of itself as its own output! The objection is strongly supported by many examples of technological manufacturing that surround us: the tiny processor in my computer, with which I am writing these lines, was built in a factory that covers many acres.

Only biological organisms seem to violate this rule (surely Mike Tyson is bigger than his mother). Von Neumann himself wrote 'There is a very obvious trait, of the "vicious circle" type, in nature, the simplest expression of which is the fact that very complicated organisms can reproduce themselves' [119]. The apparent paradox was also clearly described by Rosen [120] and is nicely discussed, for example, in David Boozer's web page [121].

The way out devised by von Neumann was the following:

- a blueprint does exist, but it describes only the remaining part of the machine, and *not* itself;
- in the replication process, the blueprint is not constructed, like all other parts of the system, by following instructions (which would require a meta-blueprint, and so on), but is simply copied (as by a trivial photocopying machine) as the final stage of replication.

In other words, 'the reproductive process uses the assembly instructions in two distinct manners: as interpreted code (during actual assembly), and as uninterpreted data (copying of assembly instructions to offspring). During the following decade, when the basic genetic mechanism began to unfold, it became clear that nature had "adopted" von Neumann's conclusions. The process by which assembly instructions (that is, DNA) are used to create a working machine (that is, proteins), indeed makes dual use of information: as interpreted code and as uninterpreted data' [122].
In greater detail, and following Taylor [26], let:

- A be a *constructor*, which can build a pattern X by following a description $\phi(X)$ of the pattern. This description will be stored in a suitable memory, which will be called 'tape' in the following. In symbols:
$$A + \phi(X) \rightarrow X.$$
- B be a *copier*, which can just copy whatever description $\phi(X)$ is found in the tape; in symbols:
$$B + \phi(X) \rightarrow \phi(X).$$
- C be a *supervisory unit*, which activates first A, then B; in symbols, the result will be:
$$A + B + C + \phi(X) \rightarrow X + \phi(X).$$

Now, suppose that the pattern X described in the instructions is A+B+C. In symbols, substituting A + B + C for X in the last expression above, we are left with:

$$A + B + C + \phi(A + B + C) \rightarrow A + B + C + \phi(A + B + C)$$

i.e. self-replication. Note that:

- it is not necessary for A to be a *universal constructor*; it just suffices that it can build the overall pattern A + B + C;
- it is not necessary for C to be a *universal computer*; it just suffices that it can direct the above sequence.

Actually, von Neumann chose to go beyond the strictest requirements for self-replication, and to make:

- A, a universal constructor, i.e. a device which could build (within the 'physical' limits of the 'universe' in which its patterns lived, namely, those of a cellular automaton), any pattern X for which a suitable description ϕ(X) was provided; and
- C, a universal computer, i.e. a Turing machine (a device which could perform any computation given the appropriate instructions on the tape – we are now surrounded by Turing machines and call them 'personal computers').

The universal constructor, under the guidance of the supervisory unit, builds a new universal constructor and a new supervisory unit. When the construction is complete, some internal switch in the supervisory unit makes it activate the copier, which reproduces the blueprint written on the tape and transfers the copy to the newly built child pattern (Fig. 1).

Mange *et al*. [37] summarise as follows the additional assumptions adopted by von Neumann to develop his cellular automaton model:

- it only deals with the flow of information, while the physical substrate and the power supply are given for granted and not investigated any further;
- the space where the cellular automaton 'lives' is two-dimensional and as large as required, i.e. unbounded;
- this space is homogeneous, i.e. it is composed of identical cells which differ only in their internal state;
- the reproduction is asexual, and the 'child' pattern is to be identical to the 'parent' pattern.

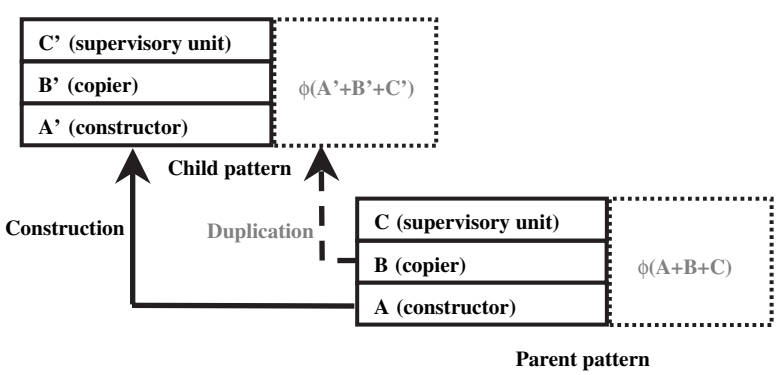


Figure 1: Architecture and mechanism of replication of von Neumann's machine.

On these assumptions, von Neumann devised a finite-state 'machine' in which each element, or cell, had 29 possible internal states and changed state according to the current state of itself and of its four cardinal neighbours (say, north, south, east, and west). Thus, the cell state transition table contains $29^5 \approx 20$ million lines! Most of the 29 cell states served the purpose of specifying the directions of growth of newly formed cells.

A substantial fraction of the cells would serve as the 'tape'; the entire system contains only one copy of its blueprint, or 'genome', stored in its memory. Thus, although it is a large cellular automaton, in a sense it is akin to a unicellular organism (the CA 'cells' are more like the molecules of this organism).

Von Neumann's 'machine' is presented and discussed in great detail, for example, in the 'Early Evolution Course' (http://ool.weizmann.ac.il/courses/ool2000/) at the Weizmann Institute of Science in Israel.

Umberto Pesavento and Renato Nobili implemented von Neumann's designs for a universal computer plus universal constructor, embedded in a CA space [123]. Their implementation relies on an extension of the state-transition rule of von Neumann's original cellular automaton, which was introduced to simplify the design of the constructor. Recently, partial hardware implementations of von Neumann's cellular automaton have been described [124].

Von Neumann's cellular automaton can be thought of as actually implementing the self-replication of a universal Turing machine. Nobody prevents this latter from being programmed to perform additional computational tasks besides acting as a supervisory unit; therefore, the system as a whole can do more than just duplicating itself, and in particular can perform computational tasks dedicating to them exponentially increasing resources. The fact that parent and child structure are identical, and thus necessarily perform identical computational feats, is not a crucial limitation; think, for example, of a Monte Carlo-like computation with random inputs which will be different for parent and child.

Recently, researchers in the field of artificial life [125–126] have pointed out that von Neumann's work should be properly regarded as having addressed (and partly solved) a substantive problem in the *evolutionary growth of complexity*, and that his design for a class of self-reproducing machines can only be properly understood in the context of solving this more general problem. A similar point is raised by Taylor [26], who remarks that von Neumann was as much concerned with the potential evolution of self-reproducing automata as with self-replication itself. Many authors have somehow overlooked this aspect, concentrating only on self-replication. As a consequence, most of their proposals are 'fragile' systems, which can hardly withstand mutations and thus make evolution impossible or, at least, difficult to achieve.

It is also worth noting that purely 'logical', or 'virtual' models, including von Neumann's CA model, lack the competition for 'raw materials' and many other features which would be present in hardware, or 'real world', systems (although concept like the competition for resources can be artificially introduced even in a virtual system, see below). Von Neumann himself was sorely aware of the limitations of purely virtual systems when he wrote:

> By axiomatizing automata in this manner one has thrown half the problem out of the window – and it may be the more important half. One has resigned oneself not to explain how these parts are made out of real things, specifically, how these parts are made up of actual elementary particles, or even of higher chemical molecules. One does not ask the most intriguing, exciting, and important question of why the molecules or aggregates which in nature really occur in these parts are the sort of things they are . . . . [7]

As to the nature of the information encoded on the tape, von Neumann suggested that 'it is better not to use a description of the pieces and how they fit together, but rather a description of

the consecutive steps to be used in building the automaton' [117]. This is called the *developmental* approach: the information is in the form of a *recipe* rather than of a *blueprint* proper.

As will be discussed later, this approach is basically the same adopted by living organisms. Its advantages have been discussed by many biologists, e.g. Maynard Smith [127] and Dawkins [25]. 'From an evolutionary point of view, one of the most important features of the developmental approach is that it allows mutations on the genotype to have a wide range of magnitudes of phenotypic effect. For example, mutations affecting the early developmental process can potentially lead to gross macroscopic changes in phenotype, whereas those affecting later stages can have the effect of "fine-tuning" particular structures' [26].

Von Neumann's approach to self-replication, based on the distinction between blueprint execution and blueprint copying, is not the only possible, and alternative strategies have been reviewed, for example, by Laing [128].

Von Neumann himself discussed the possibility of a machine which built a copy of itself by self-inspection, i.e. by actively inspecting its own parts, without the need for coded design information to be duplicated on a memory 'tape'. CAs which reproduce by self-inspection have been designed by Laing [129] and by Ibáñez *et al*. [130]. However, von Neumann suggested a number of reasons why the genetic architecture would be a more powerful and more general design for this purpose. First, copying a one-dimensional 'tape' is far simpler than copying a two- or three-dimensional structure. Second, copying a *quiescent* representation stored on a tape is far simpler than copying (without disturbing it!) a structure that must necessarily be working during the self-inspection process: '... the reason to operate with "descriptions" ... instead of the "originals" ... is that the former are quasi-quiescent (i.e. unchanging, not in an absolute sense, but for the purposes of the exploration that has to be undertaken), while the latter are live and reactive. In the situation in which we are finding ourselves here, the importance of descriptions is that they replace the varying and reactive originals by quiescent and (temporarily) unchanging semantic equivalents and thus permit copying' [7].

Another possible approach would be to start from *two* identical machines in two different states, one active and the other passive. The active machine 'reads' the passive one and builds two copies of it. Finally, the active machine activates one of the two copies. Thus we are left with a second couple of machines, made up of an active one and a passive one; the initial couple, as a whole, has replicated itself.

A third example involves two (generally different) machines A and B. A 'reads' B and builds a duplicate B′; in its turn, B 'reads' A and builds a duplicate A′. We end with a new couple $A' + B'$, identical to $A + B$. Again, the initial couple as a whole has performed self-replication.

Freitas and Gilbreath [87] review yet more alternative approaches. They also remark that machines do not necessarily have to self-replicate perfectly; they might produce 'offspring' which is slightly different from the 'parent', provided it is still capable of (imperfect) self-replication. Also, a machine might augment its 'genome' during its lifetime with valuable information drawn from its environment, and then transmit to the offspring this improved genetic code. Lamarckism (i.e. the inheritance of acquired traits), proved to be false in biology, might well come back with a vengeance in artificial self-replication!

9 In von Neumann's tracks

> You know, I occupy the John von Neumann chair at Princeton ... I think it was just a coincidence; I wouldn't have been interested in building bombs.
>
> (John Horton Conway, interview to *The Sciences*, 1994)

9.1 Progressive simplification of self-replicating CAs following von Neumann

Mostly because of the universal construction and universal computation requirements, von Neumann's 'machine' was far too complex, and neither von Neumann himself, nor anybody else, ever came up with an actual implementation. Estimates [37] put to about 200,000 the minimum number of cells that would be required for a working example (which, in any case, ruled out even the possibility of 'running' the replication process in a computer of the kind available to von Neumann!).

Following von Neumann's pioneering work, there was a curious 'downward race' towards simpler and simpler CAs which could still exhibit self-replication of the 'non-trivial' type (trivial versus non-trivial replication will be discussed in detail in Section 15).

Codd [131] devised a 5-neighbour cellular automaton which had only 8 states available to each cell in place of von Neumann's 29, but retained 'universal construction' capabilities. He estimated that a working self-replicating implementation would require some 10^8 cells. Later, a simplified version of Codd's cellular automaton, requiring only ~95,000 cells, was discussed by Devore and Hightower [132]. Also Smith [133] and Banks [134] discussed how cell space and its rules could be simplified, while still maintaining the ability to 're-capitulate' von Neumann's cellular automaton, and Vitànyi [135] proposed an 8-state, 5-neighbour sexually reproducing cellular automaton requiring some tens of thousands of cells for a working implementation.

Langton [36] took a bolder step. He dropped von Neumann's requirement that the self-reproducing automaton should be capable of universal computation and universal construction, and thus obtained a considerable simplification. He devised an 8-state, 5-neighbour cellular automaton, whose transition table consisted of 'just' $8^5 = 262,144$ lines; actually, only 219 internal transitions were used. Unlike von Neumann's construct, which remained in the stage of a conceptual study and was never actually laid down in its details, Langton's pattern was a working – though, of course, virtual, or *in silico* – self-reproducing structure. It exhibited a loop-shaped configuration with a constructing arm and a 'genome' (replicating code), which produced a child loop identical to itself after 151 clock periods, and consisted of just 94 cells, 24 of which were occupied by the genome. A simple sketch of the self-replication process is shown in Fig. 2.

Animations of self-replicating Langton loops and their variants can be seen, for example, on Hiroshi Sayama's web site (http://necsi.org/postdocs/sayama/sdsr/index.html). Of course,

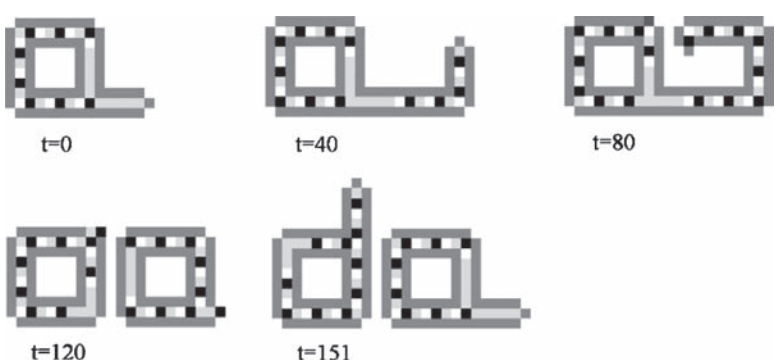


Figure 2: Self replication of Langton's 'loop'.

simplification could only be obtained at the expenses of generality; the loop does nothing apart from reproducing itself.

Following Langton, much work was done based on variants of his basic 'loop' architecture, and usually in the context of patterns that just reproduce themselves. Byl [136] presented a 12-cell, 6-state, 5-neighbour self-replicating cellular automaton which required 57 replication rules and produced a first replica after 25 steps. A number of even simpler structures were presented by Reggia and co-workers [137]; these include 5- and 6-cell, 6- and 8-state, 5-neighbour CAs which, unlike Langton's loop, do not require a 'sheath' of buffer cells lining the 'genome'. Even these simple structures, though, require a relevant number of replication rules (e.g. 58 for structure UL06W8V).

A further, considerable simplification was obtained by Sipper [138, 139] using a modified cellular space model, in which a given cell can change a neighbouring cell state and can copy its rule into a neighbouring cell. For example, a replicator requiring only 5 cells, 2 states, 9 neighbours and 10 transition rules was reported. For one of these diminutive replicators (a 6-cell pattern in a 5-neighbour, 8-state cellular automaton), Sipper and Reggia [96] showed how the self-replication cycle could be implemented on a chessboard using standard chess pieces and hand-played in just nine steps following simple transition rules!

9.2 A self-replicating pattern in Conway's 'Life'

In the above designs, simplicity was sought as a compromise between few states, few neighbours (thus, small transition table), and few cells. If the number of cells in the pattern is not regarded as a problem, then even simpler rule sets can support non-trivial self-replication.

William Poundstone, in his exceptionally beautiful book *The Recursive Universe* [99], dedicates the whole of chapter 12 ('Self-reproducing life patterns') to a detailed discussion of how self-reproduction, universal construction, and universal computation, could possibly be implemented in the most popular of all CAs, i.e. Conway's Life [98], which is based on a very simple 2-state, 9-neighbour transition table with just $2^9 = 512$ lines. Conway himself demonstrated in principle this possibility. For those who know the game of Life, the demonstration involves the general design of logical gates realised mainly by streams of 'gliders' which implement a flow of information akin to the flow of currents in electric circuits; the details are extremely ingenious and cumbersome.

Conway's proof is not a constructive one, and nobody has ever designed an actually self-replicating Life pattern in detail. Poundstone estimates that the smallest pattern exhibiting self-replication would probably require billions of active cells and would be embedded in an area of some 10^{13} cells of the Life plane, i.e. some 3×10^6 cells across. If a cell was the size of a computer monitor pixel, i.e. $\sim$0.3 mm, then the self-replicating pattern would occupy an area of $\sim$1 square km – the size of a town! In all probability the number and the complexity of the steps required would be even more astronomically large.

Poundstone addresses also the question of how such enormous self-replicating patterns might spontaneously emerge in a random Life field. His answer is that this should indeed be possible, but that the minimum size of a Life board on which such spontaneous formation might occur would be fantastically large, perhaps of the order of $10^{1,000,000,000,000}$ cells – much (much!) larger than the Universe.

Of course, the reason for the extreme complexity of a self-replicating pattern in Life is that Life was not designed to this purpose; the mere possibility of its existence is a bit of a miracle, and supports the view that any sufficiently rich substrate of rules can support self-replication.

9.3 Self-replicating CAs also capable of construction and computation

As discussed above, the race towards simpler and simpler self-replicating CA patterns has led to skinny structures which, however, can do nothing else than self-replicating, thus contravening the spirit of the pioneering studies by von Neumann (who wished to endow its 'creatures' with universal construction and computation capabilities). In recent years, CA studies have been presented which try to overcome this difficulty and to retrieve as much as possible of von Neumann's program, albeit in the context of still simple and feasible CAs.

Tempesti [140] presented a 6-state, 9-neighbour, 52-cell cellular automaton which, besides self-replicating, possessed additional construction and computational capabilities. The structure was similar in design to Langton's loop, one main difference being that the parent structure now remained active and capable of program execution. Along the same lines, Perrier *et al*. [141] presented a 63-state, 5-neighbour, 127-cell self-replicating structure exhibiting universal computation. The structure included a Turing machine model (W-machine), programmed using a small instruction set, and a string of data. After a daughter structure is produced, it can execute a W-machine program on the replicated data. Therefore, it belongs to the class of CAs which cannot only reproduce themselves, but also accomplish some useful task.

The problem with the data is that they are necessarily the same for all copies, so that these all perform the same computational task. This difficulty, of course, is rather crucial for 'useful' self-replication and is mentioned by Perrier *et al*. as one of the still not satisfactorily solved issues in 'von Neumann – like' CA machines. Possible ways out are:

- using random data, e.g. for a Monte Carlo-like calculation;
- read different data for each copy from an 'environment'.

Another alternative, explored by Chou and Reggia [142], is to allow for each replicant to receive a different partial solution to the given problem, which is modified during replication. Under artificial selection, replicants with promising solutions proliferate while those with failed solutions die. The authors presented an application to the so-called satisfiability (SAT) problem in predicate logic, a classic NP-complete problem.

Arbib [143] argued that the design complexity of von Neumann's and Codd's self-replicating automata could be greatly reduced if the fundamental components were more complex. Arbib's automaton (Constructing Turing machine, or CT-machine), is embedded in a two-dimensional cellular space. However, the cells of this space are themselves finite-state automata capable of executing short (22-instruction) programs. Composite structures are obtained by 'welding' individual cells. As in von Neumann's cellular automaton, self-replication was not actually implemented, but rather demonstrated by mathematical proof.

9.4 Emergence of self-replicating structures and evolution in a CA space

All the CA models of self-replication discussed so far must be initialised with an original (parent) copy of the replicating structure to be 'bred', and adopt a rigid set of rules tailored to this specific structure.

In the field of two-dimensional CAs, models in which self-replicating structures (replicants) spontaneously emerge from an initial random state were created by Chou and Reggia [144]. These CAs employ a general rule set that can support the replication of structures of different size and their growth from smaller to larger sizes, as well as the interaction of replicants with each other and with other (non-self-replicating) structures within the virtual cellular-automata space.

The basic architecture of the replicants is based on Langton's 'loops' [36, 137], together with the concept of *functional division* of data fields [145]. The bit depth of the CA cell (8 bit in this work) is functionally divided into different fields, each encoding different functions (in this case four fields, i.e. component, growth, bound and special). Systematic simulations show that the emergence and growth of replicants occurs often and, within certain limits, is basically independent of cellular space size, initial density and initial random pattern.

Lohn [146] developed procedures, based on genetic algorithms, to search the rule-table space of CAs for structures capable of self-replications. Besides CAs, Lohn considered also 'effector automata', in which rules specify actions, such as movement and automaton division, rather than cell state transitions. He discussed also alternative definitions of the *fitness function* characterising the degree of self-replication ability attained by a given structure. Similar work was presented by Pargellis [147, 148] in the context of self-replicating sequences of instructions.

The work by Hiroki Sayama on the increase of complexity in populations of self-replicating 'worms' [149] also broadly falls within the same field of evolution in CAs. Sayama's 'worms' are deterministic, 5-neighbour, shape-encoding CAs based on Langton's 'loops' but engineered in such a way that random collisions between 'worms' may give rise to variations ('mutations') and to the exchange of 'genetic information'.

The implications of the above studies for research on the origin of life from prebiotic precursors are obvious. In fact, they are borderline applications between 'simple' self-replicating CAs and the studies on 'artificial life' which will be described below.

At the same time, especially if one takes into account recent work on the possibility of programming replicating loops to solve computational problems besides replicating [140, 141], abstract models of replication may lead to practical applications, which will probably lie, at least in an initial stage, in the field of computation.

10 Artificial life

> I don't know a lot about this artificial life stuff – but I'm suspicious of anything Newsweek gets goofy about.
>
> (Aaron Watters, *Python quotations*, 29 September 1994)

According to one of his pioneers, Chris Langton [36], Artificial Life (ALife, AL) is 'a field of study devoted to abstract the fundamental dynamical principles underlying biological phenomena, and recreating these dynamics in other physical media such as computers, thus making them accessible to new kinds of experimental manipulation and testing'. It could be defined as a way of investigating life by synthesis rather than by analysis. In ALife, emphasis is usually on the relations between individual organisms and species, while relatively less attention is dedicated to the mechanism of reproduction proper (self-replication).

The 'organisms' of an ALife system are machine-language programs, competing for computational resources (RAM, CPU time) and capable of self-replication – of course, within the very special milieu offered by the computer. In most realizations, the 'genotype' is represented by the instructions that made up a program, and the 'phenotype' by the *actions* that a program performs as its instructions are executed [150, 151].

In the most interesting applications, AL organisms can undergo *mutations* and are thus subject to evolution by 'artificial natural selection'. They may be compared with the simple RNA replicators supposed by many to have been the first self-replicating structures to have appeared on earth

(see Section 5 on the origin of life). A difference exist in the way the ALife dynamics are initiated; one may:

- 'inoculate' the environment with hand-written programs (ancestors) already capable of self-replication, and let them evolve; or
- start from simpler, non-self-replicating, programs and look for the spontaneous emergence of self-replication capabilities.

The optimum computational environment for ALife experiments is quite different from ordinary computers, operating systems and programming languages. In particular, such environment should be particularly fault-tolerant, or otherwise most mutations (e.g. single-bit changes in machine-language programs) would be fatal and nothing interesting would ensue. Such 'special', ALife-oriented, computational environments could certainly be built, in which the ALife 'organisms' would still exist as virtual creatures made of software; but, more commonly, these environments are simply simulated within the hardware and operating systems of ordinary computers, so that the ALife beings live a doubly virtual existence, twice removed from 'blood and flesh' life.

Despite this rather ethereal nature, digital ALife organisms exhibit many of the properties of living entities and, besides their obvious theoretical interest, can be used to test self-replication, evolution, and ecology. For example, ALife has been used to compare rival biological theories, such as Darwin's gradualism versus Gould's theory of *punctuated equilibria* [52, 152]. And, as Langton puts it, 'Artificial Life can contribute to theoretical biology by locating life-as-we-know-it within the larger picture of life-as-it-could-be'.

Of course, the 'artificial physics' of the computer model must resemble, at least in some crucial respects, the physics of real world if ALife is to be treated as a *science* rather than as a sophisticated computer *game* [26].

It may be added that there are currently two 'tastes', or positions, with respect to ALife:

- the 'weak' position is to regard it as a *model* of life;
- the 'strong' position is to regard it as an *instance* of life. This rather extreme view is expressed, for example, by Thomas Ray, who holds that '. . . digital organisms are not "models" of life, they are rather forms of a different life . . . digital organisms live, to all practical purposes, in another universe where the physical laws (including thermodynamics) do not hold, whilst the only "natural laws" are the logic and rules of CPU and operating system' [23].

ALife studies have largely made use of non-standard programming techniques such as *genetic programming* (http://www.genetic-programming.com/coursemainpage.html). The method is based on John Holland's research on genetic algorithms during the 1970s and 1980s [113] and was introduced in ALife studies by John Koza at Stanford University [153, 154]. 'While a programmer develops a single program, attempting to perfect it as much as possible, genetic programming involves a population of programs. The initial population, referred to as the first generation, consists of randomly created programs. The following generations are formed by evolution so that in time the population comes to consist of better (fitter) programs (here, fitness is simply defined by the user in accordance with the particular problem at hand). Each generation is created by applying genetic operators to the previous generation programs. These operators are known in biology as crossover and mutation' [122].

A characteristic aspect of ALife is *emergence*, by which phenomena at a certain level arise from interactions at lower levels. In the words of Sipper [122], 'In physical systems, temperature and pressure are examples of emergent phenomena (an individual molecule possesses neither!). Examples of emergence are von Neumann's model, where the basic units are grid cells and

the observed phenomena involve composite objects consisting of several cells [7], and Craig Reynolds' work on flocking behaviour in birds [155]'.

Another important keyword in ALife is *subsumption*, i.e. the hierarchical ordering of structures, each dealing with a specific level of complexity. Again in the words of Sipper [122], '... the underlying principles of ALife stand also at the core of Rodney Brooks' work on robots capable of functioning in a 'noisy' environment [156]. Brooks' robots possess 'brains' comprised of a hierarchy of layers, each one performing a more complex function than the one underneath ... This architecture, dubbed the subsumption architecture, roughly resembles that of our own brains where primitive layers handle basic functions such as respiration and high-level layers handle more and more complex functions up to abstract thinking'.

The subsumption architecture is also reminiscent of Marvin Minsky's work [157], which describes the operation of the brain in terms of an *ensemble of agencies*, each responsible of a simple functionality.

The work by Thomas Ray [158] addresses the question of whether *open-ended evolution* can occur, and proceed without any human guidance, in an ALife virtual world which he dubbed Tierra. As in 'Core Wars', the creatures in Tierra are self-replicating programs which compete for the natural resources of their computerised environment, namely, CPU time and memory.

Ray did not wish to investigate on how self-replication is attained, but rather on what happens *after* its appearance on the scene. He inoculated his system with a single self-replicating organism, (Ancestor), which is the only engineered creature in Tierra, and then set the system loose. Results were highly provocative; for example, parasites evolved, i.e. small creatures using the replication machinery of larger organisms such as the Ancestor to self-replicate.

In recent work [159], Tierra was developed to account for multicellular organisms, simulated by *parallel processing*, i.e. by multiple machine-language programs sharing the same 'genome' and executing their task (basically self replication) in parallel. Note that the various 'cells' share the same code but process different data. This approach was claimed to mimic the development of multicellular organisms on earth following the so-called 'Cambrian explosion' some 600 million years ago. The evolution of the multicellular organisms by random mutations and 'natural' selection led to an increased level of parallelism and to a more efficient co-ordination of the parallel programs.

Further developments of Tierra have included the possibility for its organisms to perform additional tasks besides mere self-replication [160], along the lines followed for CAs by Tempesti [140], Perrier *et al*. [141] or Mange *et al*. [37] and discussed in a previous section of this paper.

Other 'virtual worlds' in which artificial life and evolution occur, similar to Tierra, include AVida by Chris Adami and Titus Brown [161] and Jacob Skipper's 'Computer Zoo' [162].

A recent example is Tim Taylor's 'Cosmos' [26]. The whole of Cosmos is modelled after living cell metabolism, with promoters–repressors, cell fission etc. A noteworthy characteristic is that in reproduction the 'genome' is first copied within the 'mother' cell, and then the whole cell is split (as in real life).

Among the research centres that have been most active in the last years in the field of ALife, one should mention the Digital Life Laboratory at the California Institute of Technology (http://dllab.caltech.edu/research); the Department of Computer Science at the Ben Gurion University in Israel (http://www.cs.bgu.ac.il) and the Artificial Life Laboratory at the City University in Dublin (DCU), the city where Schrödinger developed his pioneering views on the genetic code (http://www.eeng.dcu.ie/∼alife). ALife results obtained at DCU have been presented, among others, by McMullin [125, 126, 163, 164] and by McMullin and Varela [165].

Following a first workshop in Los Alamos promoted by Chris Langton in 1987, regular biennial international conferences have been devoted to the subject. There have been also periodic

European Conferences on Artificial Life and a journal of the same title, currently edited by Mark Bedau and Charles Taylor, has been published by MIT Press since 1995.

A list of on-line publications related to the field of Artificial Life can be found in the web pages of the School of Cognitive and Computing Sciences (COGS) of the University of Sussex (http://www.cogs.susx.ac.uk/users/ezequiel/alife-page/alife.html). Huge selections of links to Internet sites related to Artificial Life are available in the web pages of Ariel Dolan (www.aridolan.com), Tim Tyler (http://www.alife.co.uk/) and Eric Max Francis (http://www.alcyone.com/max/links/alife.html).

Among ALife available programs, a shareware version of Rudy Rucker's 'Boppers' is available online [166]. It is based on the rules described in the book *Artificial Life Lab* by the same author [167]. The program shows creatures (boppers) grouped into up to three colonies. The boppers have 'genes' which are bit strings specifying a number of parameters, and their fitness level is determined in a co-evolutionary manner, as in a predator–prey system. Cloning, mutation, and crossover operators are implemented. Different styles of boppers are possible. The two main types are 'turmites' and 'boids'; the turmites are two-dimensional Turing machines, while the boids obey a flocking algorithm of the type devised by Craig Reynolds [155].

To give an idea of the current status of research in ALife, we report a list of 14 open problems in 'Artificial Life' as recently been proposed by Bedau *et al*. [168] (in the spirit of Hilbert's famous open mathematical problems proposed at the dawn of the 20th century). The problems come under three major headings.

A (problems 1–5): How does life arise from the non-living?

1. Generate a molecular proto-organism *in vitro*.
2. Achieve the transition to life in an artificial chemistry *in silico*.
3. Determine whether fundamentally novel living organisations can exist.
4. Simulate a unicellular organism over its entire life cycle.
5. Explain how rules and symbols are generated from physical dynamics in living systems.

B (problems 6–10): What are the potentials and limits of living systems?

6. Determine what is inevitable in the open-ended evolution of life.
7. Determine minimal conditions for evolutionary transitions from specific to generic response systems.
8. Create a formal framework for synthesising dynamical hierarchies at all scales.
9. Determine the predictability of evolutionary consequences of manipulating organisms and ecosystems.
10. Develop a theory of information processing, information flow, and information generation for evolving systems.

C (problems 11–14): How is life related to mind, machines, and culture?

11. Demonstrate the emergence of intelligence and mind in an artificial living system.
12. Evaluate the influence of machines on the next major evolutionary transition of life.
13. Provide a quantitative model of the interplay between cultural and biological evolution.
14. Establish ethical principles for artificial life.

Chou and Reggia [144] wrote: 'Attempts to create artificial self-replicating structures or 'machines' have been motivated by the desire to understand the fundamental information processing principles involved in self-replication'. This quotation makes it clear that, in studies on

self-replication, the attention has been focussed so far mainly on the (abstract) *information-processing* aspects of the phenomenon. Little work has been done on the *concrete physical mechanisms* (actual manipulation and transformation of raw materials) upon which self-replicating systems work or might work.

11 Real-world self-replication

If you can't beat your computer at chess, try kickboxing.

(Anonymous)

When we move from the virtual world of purely symbolic systems to the real world of physical objects and physical laws, a host of new difficulties is encountered. Most of the work done in this field has regarded so far only simple systems, exhibiting only rudimentary self-replication capabilities and requiring the availability of parts only slightly less complex than the whole.

However, if the difficulties are great, the rewards would be great too; the self-replication of physical machines would open great possibilities concerning, for example, the economic production of manufacts and the feasibility itself of molecular-scale nanotechnology.

11.1 Molecular self-assembly

An essential prerequisite for 'real world' self-replication is probably self-assembly, i.e. the ability of individual constituent parts to spontaneously assemble together with the proper spatial orientation into complex, composite structures [169]. Of course, self-assembly has much less significance in virtual (e.g. CA-based) self-replicating systems.

Self-assembly occurs both in nature and in artificial (built or, at least, designed) systems. In nature, it is commonly observed in cell metabolism. An impressive example (Fig. 3) is the self assembly of *clathrin* molecules to form first *triskelia*, i.e. aggregates of three molecules in the form of three-armed 'swastikas', and then *coated vesicles*, polyhedral 'cages' shaped as regular icosahedra, which shuttle through the cell a range of important bio-molecules [170]. Another well-known example is the process by which membrane pores are formed by the random encounter of component proteins, freely floating in the 'two-dimensional fluid' represented by the lipid bilayer [171]. Strictly related to self-assembly is also the folding of linear protein strands into their stable tertiary structure.

Also, on increasingly larger levels, one may cite as examples of biological self-assembly:

- the self-assembly of a *viral particle* from its molecular 'bricks';
- the process by which individual proteins, continuously produced by randomly scattered ribosomes, find their way to the 'correct' site in the architecture of a living cell;
- the unfolding of a full living organism from a fertilised egg during embryogenesis.

All these processes occur without a 'constructing arm', of the kind envisaged by von Neumann, placing individual components into their correct positions, and thus contrast with von Neumann's view of a 'universal constructor'.

In the realm of the artificial, Julius Rebek and co-workers at the Skaggs Institute for Chemical Biology (La Jolla) have produced a number of synthetic macromolecules which exhibit self-assembly and self-replication from smaller units [172]. Details and animations can be found in their web site (http://www.scripps.edu/skaggs/rebek).

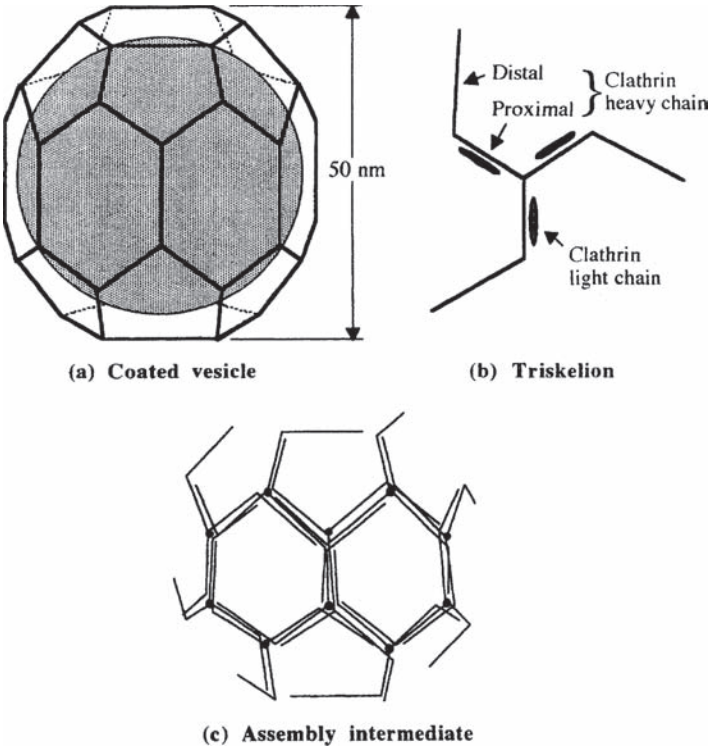


Figure 3: Structure and assembly of a coated vesicle. (a) A typical coated vesicle contains a membrane vesicle about 40 nm in diameter, surrounded by a fibrous network of 12 pentagons and 8 hexagons. The fibrous coat is constructed of 36 clathrin *triskelia*. (b) Detail of a clathrin *triskelion*. Each of three clathrin heavy chains is bent into a proximal arm and a distal arm. A clathrin light chain is attached to each heavy chain, most likely near the centre. (c) An intermediate stage in the assembly of a coated vesicle, containing 10 of the final 36 *triskelia*, illustrates the packing of the clathrin *triskelia.* Each of the 54 edges of a coated vesicle is constructed of two proximal and two distal arms interwined. The 36 *triskelia* contain $36 \times 3 = 108$ proximal and 108 distal arms, and the coated vesicle has precisely 54 edges [171, 207].

Self-assembly has been seriously considered as a possible intermediate stage towards self-replication in nanotechnology. For example, Solem [173] discusses how the self-assembly of simple polyhedral 'bricks' into larger aggregates of a required shape can be obtained by controlling the distribution of electric charge on the faces of the elementary polyhedra (however, much of the original nanotechnology work, e.g. by Drexler [40, 174], still relies on a von Neumann-like 'constructor arm' to place individual molecules and atoms into their correct places).

In many instances of self-assembly, it is crucial that the components can move freely in space in order to find their way to the appropriate equilibrium positions. This is most easily achieved if the components can float in a fluid medium. In turn, this raises delicate issues of hydrodynamic interactions at very small scales, where Stokes flow, or even molecular dynamics below the threshold of continuum, applies [171]. Interestingly, also a variant of the first self-replication

models envisaged by von Neumann, i.e. the so called kinematic model, required components to 'float' freely in a pond or a similar fluid environment.

11.2 Simple mechanical and electro-mechanical devices which exhibit self-replication

In the late 1950s, the issue of self-replication in real physical systems attracted the interest of the British physician and geneticist Lionel Penrose, better known for his pioneering work on mental retardation and Down's syndrome; he was the first to demonstrate the significance of the mother's age. (Speaking of mental retardation, Lionel Penrose was father to Roger, the well known physicist and Nobel laureate; Jonathan, lecturer in psychology and perhaps the strongest chess player in Britain in the 1950s and 1960s, with an ELO score above 2500 at the age of 17; and Oliver, who went on to become professor of mathematics first at the Open University, then at Heriot-Watt University in Edinburgh.)

Besides theoretical work [175–177], Lionel Penrose, aided by his son Roger, built simple mechanical units, or bricks, which exhibited some rudimentary self-replication (Fig. 4). An ensemble of such units was placed in a box, which was then shaken. 'In fanciful terms, we visualized the process of mechanical self-replication proceeding somewhat as follows: Suppose we have a sack or some other container full of units jostling one another as the sack is shaken and distorted in all manner of ways. In spite of this, the units remain detached from one another. Then we put into the sack a prearranged connected structure made from units exactly similar to those already within the sack . . . Now we agitate the sack again in the same random and vigorous manner, with the seed structure jostling about among the neutral units. This time we find that replicas of the seed structure have been assembled from the formerly neutral or "lifeless" material' [38].

Incidentally, the shape of the blocks will surely ring a bell in the head of readers familiar with Roger Penrose's work on the aperiodic tiling of the plane!

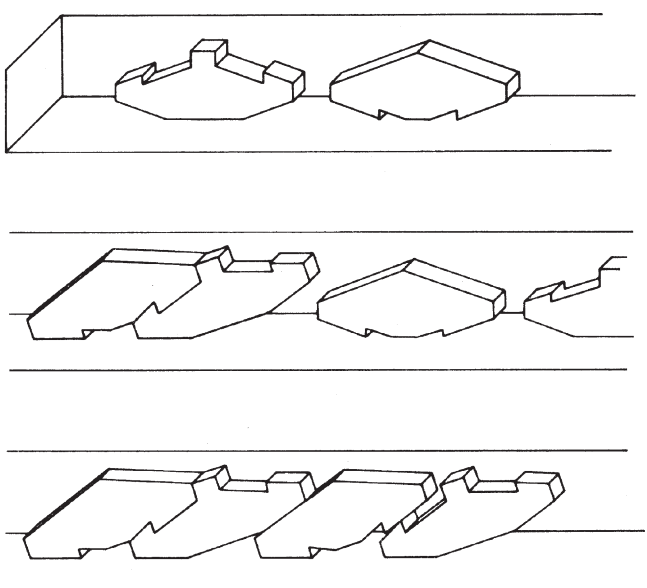


Figure 4: Lionel Penrose's [38] 'self-replicating' block assemblies (from [99]; reproduced with permission).

About in the same years, Homer Jacobson, also known for his studies on the information content of biological and engineered systems, built a self-replicating contraption using toy train parts running around a track [178]. A simple electro-mechanical replicator was presented by Morowitz [179].

More recently, designs for simple electro-mechanical self-replicating systems have been presented by Lohn *et al*. [180] in the context of developing strategies for the self-replication of molecular assemblers in nanotechnology. Their paper includes also a schematic descriptions of the above systems by Penrose and Morowitz.

A number of somewhat more complex self-replicating robots have recently been built by Chirikjian and co-workers at the John Hopkins University [181]. Assembled out of Lego Mindstorm part kits, they operate in an environment which is composed of half-assembled robot parts and includes also some auxiliary fixtures. The self-replication process basically consists of the final assembling steps necessary to make a copy of the whole robot (Fig. 5).

Closely connected with self-replication are those research project which aim to attain self-assembly, self-repair and self-reconfiguration in robots and machines. A rather extreme example is the infinitely reconfigurable, nanotechnology-based 'utility fog' envisaged by Storrs Hall [182], while a more typical example is the work carried out by Satoshi Murata, Eiichi Yoshida and co-workers (http://www.mel.go.jp/soshiki/buturi/system/menu-e.htm) at the Agency of Industrial Science and Technology in Japan.

Another related field is that dubbed 'evolutionary robotics' [183]. Here 'an evolutionary process operates on a population of candidate robots, each composed of some repertoire of simple building blocks. The process iteratively selects fitter machines, creates offspring by adding, modifying and removing building blocks using a set of operators, and replaces them into the population' [184]. Gradually, fitter and fitter robots are thus obtained (fitness being somewhat arbitrarily defined from the outside, in what can be regarded as a hardware equivalent of the genetic algorithms mentioned in a previous section.

It should be observed, however, that neither self-repair nor evolutionary robotics necessarily imply self-replication proper.

Skidmore *et al*. [185] summarise the current state-of-the-art in real-world self-replication as follows: 'The self-replicating entities of the biological arena are often used as an existence proof for the possibility and inevitability of self-replicating machinery. Simpler, man-made self-replicating machinery, however, has not seen much progress since the demonstrations of the first rudimentary self-reproducing machine demonstrations . . . The first artificial, self-replicating mechanical structures were built by Penrose, and the first electro-mechanical systems were created by Jacobson decades ago . . . These authors simplified self-replication by showing that it is not inherently a complex problem and that a rich environment of complex parts and simple reproduction steps reduces the self-replicating problem to a tractable one'.

11.3 Ontogenetic hardware: midway between virtual-world and real-world self-replication

Sipper *et al*. [17, 186] define as *ontogenetic hardware* an integrated circuit environment in which functions characteristic of cellular organisation, including growth, self-repair, and self-replication, can be implemented.

In principle, any function which can be implemented in a virtual world, i.e. in the form of software structures such as CAs and computer programs, can also be implemented in hardware form on a suitable substrate. Typically, the device used to this purpose is the field-programmable gate array (FPGA) [187]. An FPGA is a two-dimensional array of processors (cells) in which

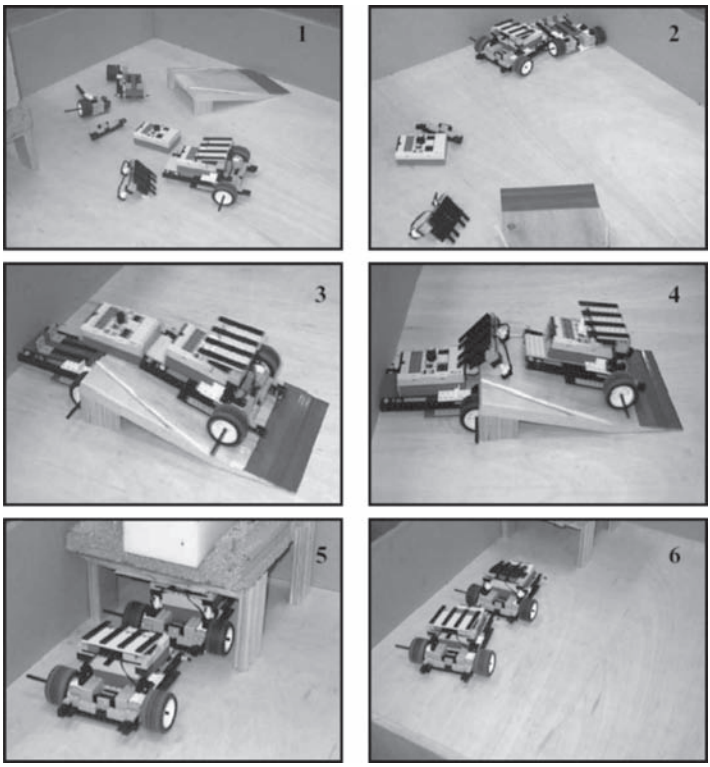


Figure 5: Stages of the self-replication process for robots assembled with Lego Mindstorm kits (from [181]; reproduced with permission). Two fixtures are used: a ramp and a tunnel-like cave with a wedge on the ceiling, used to physically force the connector in place. The process begins with the original robot dragging the right part (which consists of half of the chassis, the right wheel and the right motor) to a wall. Then the left part (which consists of half of the chassis, the left wheel and the left motor) is pushed to connect with this right part. The left and right parts of the replica are securely merged by adding the bumper which has interlocks to both subsystems. The combined subsystems are moved and oriented to a desired position and orientation next to the ramp. The original robot then pushes the controller up to the ramp, and gently drops and inserts the controller on the top of the previous combination of subsystems. The connector is fixed in its place in the same fashion. The last step is to physically force the connector to be in contact with the controller by pushing the replica in the tunnel-like area with a wedge on the ceiling. After pushing the replica several times, the electronic connectors on the replica finally make contact. The replica is able to operate in the same way as the original does.

the function of each cell, the interconnection between cells, and the inputs and outputs, can be programmed by loading a string of bits. While programming a standard processor amounts to loading data into a memory area without changing its hardware, programming an FPGA actually amounts to physically changing the configuration of the underlying circuit at gate level, as in ROMs (firmware) and in neural networks. Any action of a cell amounts to writing a bit-string program into a neighbouring cell which is initially in a neutral, ('blank', or non-programmed) state,

thus changing its configuration. For example, a cell can copy itself into neighbouring locations in the cell array (self-replicate).

Clearly, this kind of ontogenetic hardware lies midway between purely virtual 'artificial worlds' and a physical, 'real-world', status. As in the case of robots building copies of themselves from half-assembled, complex parts, discussed above, a highly organised environment is still a prerequisite for metabolism and self-replication. In the present case, this environment or substrate is the array of 'blank' non-programmed cells: of course, FPGA processors cannot actually build other physical processors.

Mange and Stauffer [188] presented a cellular automaton structure which was complex enough for universal computation, but still simple enough for a physical implementation through commercially available digital circuits. They dubbed the unitary cell of this structure a *biodule* (for 'biological module'), and the associated overall architecture *embryonics* (for 'embryo electro nics'). Using this approach, Mange *et al*. [37] realised a universal Turing machine which, while replicating, could perform a simple computational task (parenthesis checking) and some limited self-repair; and Mange *et al*. [189] presented a self-repairing structure known as the *biowatch*.

In von Neumann's cellular automaton, the whole parent structure contains only one copy of the blueprint (genome) and thus is akin to a unicellular being (of course, cell is used here with a meaning different from that in cellular automaton). By contrast, biodule architectures are truly 'multicellular' automata, with the genome repeated in each of the cells. Multicellular structures based on 'biodules' promise to have important practical applications especially in the field of self-repairing, fail-proof devices and mechanisms.

12 Self-replicating probes for space exploration

> Uniquely in history, we have the circumstances in which we can create Athens without the slaves.
>
> (British Agriculture Minister Peter Walker, reported in [16])

Within the realm of 'real world' self-replication, what still today remains the most advanced and coherent effort towards a self-replicating machine is certainly the 'lunar factory' project described by Freitas and Gilbreath [87].

By the late 1970s NASA had become aware that a gap existed because their space exploration projects and the most advanced technological and scientific achievements of the time. Around 1977, a decision was made to get updated with robotics and artificial intelligence studies, which was initially pursued with the Woods Hole workshop (1979) and the 'Telefactors Working Group on future applications of very advanced automation to space missions'.

In June 1980, a symposium was held at Pajaro Dunes on these topics. One of the projects which were identified as worth developing was the design of a *lunar factory* which, starting from an earth-manufactured 'seed', could lead within short times to a large-scale industrial output by means of self-replication. A summer study was dedicated to this subject in the period June–August 1980, with an overall effort of some 10^4 man-hours and a cost of some 12 million dollars.

The main result of the study was a realistic proposal for a self-replicating automated factory system, capable of exponentially increasing productive capacity. As a first scenario, it was assumed that such a factory would be located on the surface of the moon. The 'seed', manufactured on earth, would not exceed $\sim$100 ton of weight (corresponding to a realistic payload). Starting from this seed, growth would occur in a centrifugal fashion with a doubling time of $\sim$1 year. During the growth period, the factory would mainly produce further parts of itself, to switch to

the production of useful output once some prescribed size were achieved; as an alternative, both growth and the production of useful output might be pursued from the beginning. Useful output might include products (useful manufacts, hazardous chemicals or rare minerals) and/or services (sky monitoring for SETI, asteroid surveillance or military purposes).

Assuming that the output consisted of refined minerals, the study estimated that the production rate might rise to 10^6 ton/year within 10 years. By then, the initial 'seed' would undergo a thousand-fold increase, attain a size of $\sim 10^5$ ton and require a power supply of $\sim$2 GW (mainly provided by solar cells), with a data throughput of $\sim$16,000 Gbits/s and a mass storage capacity of $\sim 3 \times 10^5$ Gbits. A production rate of $\sim 4 \times 10^9$ ton/year (equivalent to the current annual industrial output of all human civilisation) would be attained within 18 years.

The proposal, Freitas comments, was 'quietly declined with barely a ripple in the press'.

As in self-replication in general, and in the case of molecular assemblers which will be discussed further on, also for the 'lunar factory' a *closure* problem arose [190]. Complete closure would be achieved if the system were totally self-sufficient, i.e. required only lunar soil minerals and available solar energy as the 'raw material' to be processed. Such a complete autonomy, however, would be an unnecessarily hard goal. It was judged that a degree of closure of 90–96% would be quite satisfactory; the factory would then need only a moderate supply of components (e.g. sophisticated electronics) from the earth. These would be close analogues of vitamins (chemicals that a living organism cannot synthesise and that must be provided from the outside).

In the *very* long run, a self-replicating factory could become part of interstellar missions which, in principle, could allow the exploration of the nearest 10^6 stars within $\sim 10^4$ years and of the entire galaxy within a timeframe of the order of 10^6 years.

A detailed discussion of the possibilities opened by self-replicating interstellar probes is given by John Barrow and Frank Tipler in their celebrated book *The Anthropic Cosmological Principle* [191]. In section 9.2, 'General theory of space exploration and colonisation', they discuss the possibility of a self-reproducing universal constructor with human-level intelligence, built according to the general principles laid down by John von Neumann [7], and including an interstellar rocket. They call such a device a 'von Neumann probe'. It would be instructed to reach a 'nearby' stellar system, search out suitable construction material, build one or more copies of itself and of the original probe rocket engine, and launch it, or them, towards further stellar systems. As a nice extra touch, a von Neumann probe could be instructed to synthesise a number of human beings from the appropriate genetic information.

Incidentally, as is discussed in their subsequent section 9.3 (Upper bounds on the number of intelligent species in the Galaxy), the same line of reasoning leads to the conclusion that, most probably, we are the only intelligent species in our Galaxy (inter-galactic colonisation was intentionally left out of the discussion since it raises difficult issues such as the possibility of *very* long-life manufacts). The rationale for the above, somewhat gloomy, conclusion is that, if other intelligent species existed, they would have developed, sooner or later, the technology for von Neumann probes and would be here already – which, apart from UFO aficionados, nobody believes. The concept was so synthesised by Enrico Fermi: 'if they existed, they would be here'.

The above, eminently anthropic, concepts have some obvious relevance to the Kauffman-Monod *querelle* of 'we at home in the Universe', or 'we the expected' against 'chance caught on the wing'. Uniqueness of intelligence, of course, does not imply uniqueness of life – it may well be that life pops out easily enough whenever the appropriate conditions are met, as held by Kauffman, whereas intelligence remains a unique, exceptional event in the history of Universe ('chance caught on the wing'). Yet, it is difficult to avoid the sensation that a true proof of the uniqueness of intelligence would somehow shed strong doubts on the idea of the ubiquity of life

('if we as *intelligent* beings are here by sheer chance, why can't we be here by sheer chance also as just *living* beings?').

13 Self-replication and nanotechnology

> ... within hours of his arrival at the remote testing center, Jack discovers his wife's firm has created self-replicating nanotechnology – a literal swarm of microscopic machines.
>
> (Amazon's editorial review of Michael Crichton's SF novel *Prey* [192])

The invited lecture given by Richard Feynman in 1959 at the annual meeting of the American Physical Society (West Coast Section), under the title 'There's Plenty of Room at the Bottom' [193] is widely regarded as the founding stone of nanotechnology.

The word itself has been in use since the early 1970s to denote the manufacturing of structures in which the size, shape and relative position of parts and components are specified to nanometre scale, or imaging and measurement techniques having nanometre resolution [194].

Many recently developed techniques do allow the measurement or the manipulation of matter at nanometre scales. They include imaging devices which can also be used as atom-scale 'tools', like the atomic force microscope, the scanning tunnelling microscope and the total internal reflection microscope. Considerable progress has also been achieved in the field of electronics and semiconductor technology thanks to new nanoscale techniques like ultra-thin epitaxial film growth or electron-beam, ion-beam and X-ray beam lithography. Another engineering area which falls within the above 'loose' definition of nanotechnology is *micromachining*, i.e. the fabrication of mechanical devices, involving moving parts, having typical sizes of a few μm to a few mm; although metals and any other conventional engineering materials can be involved, the most recent and striking achievements have been based on the machining of silicon by essentially the same methods (like epitaxy and lithography) used for electronic microcircuits [195].

Most of these applications are characterised by a 'top-down' approach to miniaturisation: smaller and smaller sizes and tolerances are achieved by pushing to their extreme limits some relatively conventional macro-scale techniques, originally developed without reference to nanotechnology. A rather different and more radical meaning, based on an opposite 'bottom-up' approach, has been given to the word 'nanotechnology' by Karl Eric Drexler and by a small but influential group of supporters including Marvin Minsky at MIT, one of the fathers of Artificial Intelligence, and Bill Joy, founder of Sun Microsystems. In their sense nanotechnology (also referred to as molecular nanotechnology, or molecular manufacturing) is the building of structures to complex, atomic specifications by means of mechanosynthesis, i.e. by chemical synthesis controlled by mechanical means to enable direct positional selection of reaction sites.

The book *Engines of Creation* [40] made these ideas popular and elicited a review by Dewdney in his *Scientific American* column 'Computer recreations' [196]. In synthesis, these ideas are:

- engineered nanoscale systems, built with atom-by-atom control, will become feasible soon;
- within realistic time scales, it should be possible to manufacture nanoscale assemblers, able to produce other nanodevices, or even replicators, able to produce copies of themselves (given the appropriate molecular bricks);
- such devices, used of course in large numbers and exploiting the inherent advantages of ultra-miniaturisation (very short response times, very large surface to volume ratio and very easy coolability) will be capable of performances unattainable by any other approach (computing at 10^{10} Mips/watt, electromechanical power conversion at 10^9 MW/m^3 and mechanosynthesis

at 10^6 operations per device, leading to the assembly of 1 kg of macroscopic products in less than 10^{-9} s);
- atom-by-atom control on macroscopic scales would allow the manufacturing of engineering materials and components having exceptional properties (e.g. monocrystals exempt from defects having tensile strengths of 5×10^{10} GPa, and composite materials with prescribed atom-level structure);
- nanoscale assemblers and manipulators will be able to perform molecular-level maintenance and repair work on the very biological machinery of living cells and tissues.

Of course, in a preliminary stage nanoscale assemblers will have to be built by using macroscopic tools capable of nanometre precision; but these, as discussed above, are already on their way.

Drexler's ideas were more rigorously put down in a second book, *Nanosystems* [174], where potential applications are barely mentioned while the stress is on the physical constraints imposed by the laws of nature on molecular nanotechnology (e.g. thermal noise problems); on the detailed design of some basic components (like bearings and manipulator arms); and on manufacturing strategies for nanodevices. Specific topics include gears and bearings, molecular electronic devices, mechanical nanocomputing, and strategies for molecular systems engineering.

Figure 6 shows a relatively 'large' nanodevice, i.e. a manipulator arm to be used during assembly operations. It would be made of more than 4×10^6 atoms and would allow an excursion of some 50 nm. Figure 7 compares the size of some typical nanomachines (including a nanocomputer having the same power as a current mainframe) with that of some biological particles. A relatively complex nanomachine like the manipulator arm in Fig. 6 would still be much smaller than a virus and comparable in size with a DNA helix or a microtubule.

Under many respects, Drexler's nanotechnology is more akin to chemistry – especially to the new macromolecular chemistry – than to conventional technologies based on a 'cut and mill' approach. Current technologies do not allow yet the degree of control over matter required for

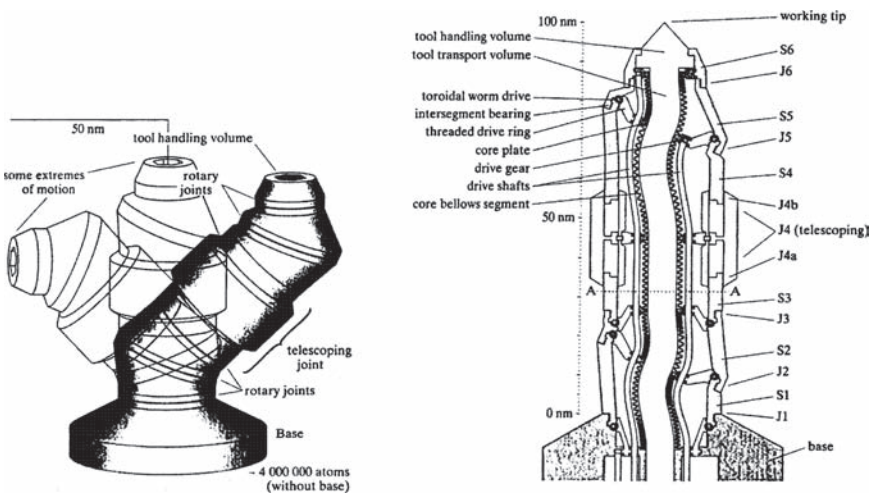


Figure 6: Design of a nanoscale manipulator arm (universal constructor). Left: external shape and range of motion; right: cross section and schematic of internal parts (from figures 13.10 and 13.11 in [174]; reprinted with permission from John Wiley & Sons, Inc.).

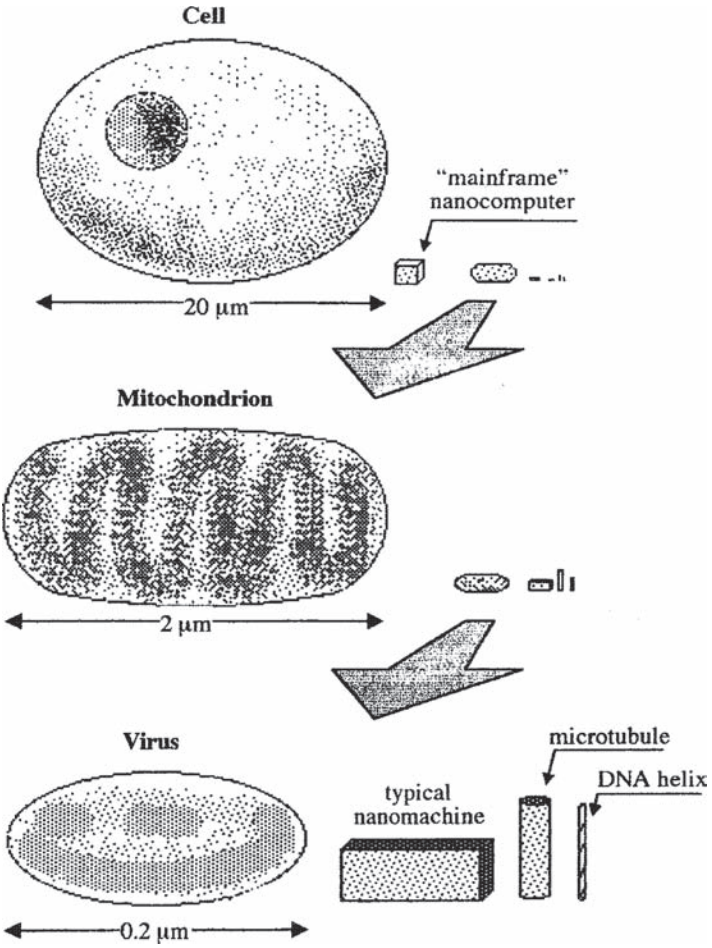


Figure 7: Comparison of biological and artificial 'machinery' [221].

Drexler's 'engines of creation' to come true; however, several recent technological and scientific achievements can be regarded as promising steps towards such goal.

In this more advanced, 'bottom-up' view, the existence itself of 'molecular machinery' in living organisms which is capable of self-assembly and self-replication plays the important role of an ultimate proof that such devices are, indeed, possible. More specifically, design solutions for man-made devices can be suggested by existing biological structures performing similar functions (which, incidentally, is what the pharmaceutical industry does every day with drugs).

It should be observed that, while the overall complexity and organisation of living matter will probably remain for a long time unsurpassed by artificial creations, yet in individual, specific tasks the biological solutions (shaped by evolution from the available 'raw materials') are not necessarily optimal in an engineering sense. For example, Drexler [174] points out that diamond-like structures, based on covalent bonds of carbon and strengthened where required by suitable atoms not available in living tissues, may well be simpler, tougher and more durable than corresponding organic structures, mainly based on proteins in aqueous solutions. He also remarks that designing

de novo stiff and reliable protein-like structures capable of performing specific tasks may actually be simpler than predicting the shape and tertiary structure of a given natural protein.

Many of the potential applications of molecular nanotechnology were identified, since Drexler's first book (1986), with biomedical tasks made possible by programmed nanodevices 'let loose' in the blood stream and capable of killing undesired cells or repairing faulty genes. The title of Dewdney's (1988) paper nicely summarises these views; an attached illustration depicts a nano-submarine, smaller than red blood cells, attacking a fat deposit in a small vessel.

Since the hypothetical nanodevices envisaged in molecular manufacturing are *small*, most applications would require a huge number of these devices simultaneously working on a project (although some applications, notably of biomedical nature, may require a relatively small number of individual nanomachines). Self-replication would be the most economical – and, in many cases, the only viable – approach to the production of such huge numbers of devices, as it would allow, at least in principle, an exponential growth of the available devices starting from a single 'seed' or a small number of seeds. As commented by Skidmore *et al.* [185], 'self-replication and nanotechnology are commonly discussed together as it has been suggested that no large-scale nanotechnology industry can develop without self-replication'.

In the words of John Storrs Hall, one of the research fellows at the Institute for Molecular Manufacturing, 'to achieve the levels of productivity that are one of the desirable prospects of a molecular manufacturing technology, the techniques of self-reproduction in mechanical systems are a major requirement. Although the basic ideas and high-level descriptions were introduced by von Neumann in the 1940s, no detailed design in a mechanically realistic regime (the "kinematic model") has been done' [197]. The author then proceeds to discuss some alternative architectures for replicating systems, and concludes in favour of a multilevel, hierarchical scheme in which devices at a given level of complexity self-replicate for a few generations and then switch to the status of assemblers for the devices of the next level (and so on), until the final desired nanomachine (not necessarily endowed with self-replication capabilities) is obtained.

Of course, besides architectural aspects, also the details of molecular manufacturing have still to be worked out. For example, a novel approach invented by Bruce Smith [198] uses DNA–protein conjugates produced by biotechnology as molecular building blocks for the assemblers. This technique utilises the well-understood complementarity of DNA sequences to assemble proteins into specific configurations.

Molecular nanotechnology in general, and the problem of self-replication in particular, have been the subject of a rather hot debate in a recent issue of *Scientific American* (September 2001). The most radical criticism was raised by Nobel laureate Richard Smalley [199] who concluded 'Self-replicating, mechanical nanobots are simply not possible in our world.' Two principal objections were advanced, neither of which specifically related to self-replication but rather to the possibility itself of precise molecular-level manipulation:

- the 'fat fingers' problem ('. . . there just isn't enough room in the nanometer-size reaction region to accommodate all the fingers of all the manipulators necessary to have complete control of the chemistry');
- the 'sticky fingers' problem ('. . . the atoms of the manipulator hands will adhere to the atom that is being moved. So it will often be impossible to release this minuscule building block in precisely the right spot . . .').

'Both these problems are fundamental', Smalley wrote, 'and neither can be avoided'. As could be expected, Drexler and co-workers presented a rebuttal of these objections. The most convincing counter-objection, and the strongest argument in favour of molecular nanotechnology, is indeed that which appeals to the actual existence of nanoscale manufacturing in living organisms: 'For a

biological example, consider the ribosome. This ubiquitous biological molecular assembler suffers from neither the "fat finger" nor the "sticky finger" problem. If, as Smalley argues, both problems are "fundamental", then why would they prevent the development of mechanical assemblers and not biological assemblers? If the class of molecular structures known as proteins can be synthesized using positional techniques, then why would we expect there to be no other classes of molecular structures that can be synthesized using positional techniques?' [200].

At any rate, some are bracing themselves for a future of spreading molecular nanotechnology. Eric Drexler founded the 'Foresight Institute' (www.foresight.org) and the 'Institute for Molecular Machinery' (www.imm.org) with the declared purposes 'to guide emerging technologies to improve the human condition', 'to carry out research to develop molecular manufacturing (molecular nanotechnology, or MNT)', and to promote 'guidelines for research and development practices that will minimise risk from accidental misuse or from abuse of molecular nanotechnology'. Ralph Merkle, together with James Von Ehr II, former owner of Altsys, founded Zyvex, 'the first molecular nanotechnology company' (www.zyvex.com).

Merkle [201–203] has observed that in Drexler's assemblers, as well as in the von Neumann architecture and in living systems, constructor and computer are both embodied in each machine. However, this is not a logical necessity in a general manufacturing system. One might, for example, separate the 'constructor' from the 'computer', and allow many individual constructors to receive instructions broadcast from a single central computer. This eliminates the need for each constructor to store internally the plans for what it is going to construct (including itself), thus eliminating most of the mechanisms involved in decoding and interpreting those plans.

Such a 'broadcast architecture' is sketched in Fig. 8. It would reduce the size and complexity of the self-replicating component, and would also allow self-replicating components to be rapidly redirected to build something novel. Also, if the central computer is macroscopic and under direct human control, the 'broadcast architecture' is inherently safe since individual constructors lack sufficient capability to function autonomously. The author comments 'This general approach is similar to that taken in the Connection Machine, in which a single complex central processor broadcasts instructions to a large number of very simple processors. Storing the program, decoding instructions, and other common activities are the responsibility of the single central processor;

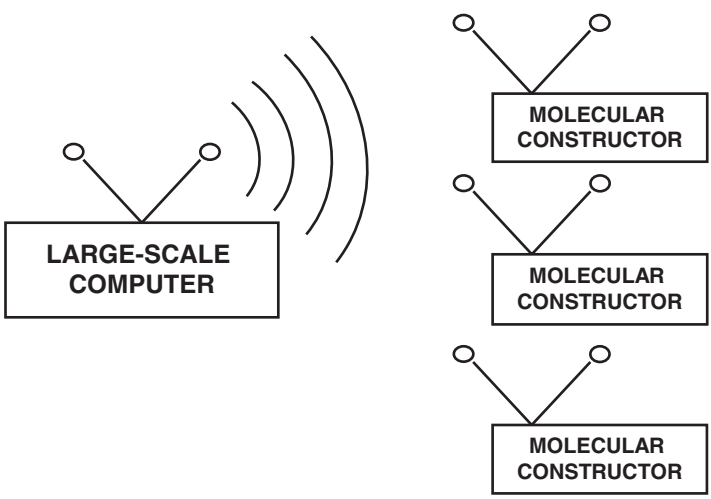


Figure 8: 'Broadcast architecture' for nanoscale assemblers.

while the large number of small processors need only interpret a small set of very simple instructions ... It is interesting to view the cell as using the broadcast architecture with the nucleus as the "central computer" broadcasting instructions in the form of mRNA to perhaps millions of ribosomes' [202].

As regards the form of the transmission, among the many feasible approaches, a simple mechanical proposal, that interfaces easily to the mechanical computational system and also works well in a liquid environment, would be to transmit information *acoustically*. 'Each assembler would be equipped with a pressure sensor, and would note the rise and fall of the local pressure. Transmission rates in the megahertz range should be relatively easy to achieve and would be sufficient for the purposes considered here' [201].

In a similar spirit, Skidmore *et al*. [185] observe that 'completely self-replicating systems may not prove necessary for nanotechnology, as simpler systems containing some replicating aspects will undoubtedly be easier to implement'. The authors describe the conceptual design of an array of 'assembly stations' having individual rotational degrees of freedom but shared translational degrees of freedom and shared control. The stations rely on pre-manufactured parts, and their operation is limited to assembling further copies of themselves from these parts. A single, 'manually' assembled station, is necessary to start the whole process, which then proceeds with an exponential growth of the production rate and of the overall product. This somewhat partial self-replication process is dubbed 'exponential assembly'. It is yet another example of how self-replication is the more easy and feasible as the complexity of the environment (what Freitas and Gilbreath call the 'substrate', and von Neumann himself the 'milieu') is allowed to increase. It also shows that 'partial' self-replication starting from half-assembled and relatively complex parts, if somewhat unsatisfactory as a model of *life*, can nevertheless have useful technological applications.

14 A comparison of natural and artificial self-replication

> ... any automaton sufficiently complex to reproduce itself would necessarily possess certain component systems which are strictly analogous to those found in a cell.
>
> (M. Denton, 1986, [204])

Having given an overview of natural (biological) and artificial self-replication, it is now possible to step back and have a more detached look at the main points of contact and of divergence. Some of the points at issue, mainly concerning the details of the self-replication process, will be touched here. More general and philosophical aspects will be discussed in the following and conclusive sections.

14.1 Homeostasis, metabolism, and replication

Laing [128, 129, 205] and many other authors have compared von Neumann's ideas, as expressed in his CA self-replication model, and biological processes as occurring in living cells. Such comparisons often parallel the construction/copying (i.e. blueprint execution/blueprint duplication) steps in von Neumann's cellular automaton to the DNA transcription/DNA translation steps in cells.

Actually, things may be more complex and are worth analysing in greater detail.

In his work on self-replicating automata, von Neumann envisages a rather clear-cut replication process, in which there is a sharp distinction between 'blueprint' and 'body' and which occurs in two stages:

- a 'universal constructor', directed by the blueprint, reproduces the original system (automaton) with the exception of the blueprint itself (thus avoiding the logical loops implied by the problem of self-description);
- in the end, the blueprint is simply copied and the copy is added to the newly created 'daughter' automaton.

As regards living cells, first of all it should be observed that here the *reproductive* function is strictly intertwined with the metabolic function. This reflects the inextricable link between genes and proteins. Things may have been different in the past, when reproduction may have been relatively separated from metabolism [19].

Freeman Dyson [58] asks: 'Is life one thing or two things? Are metabolism and replication connected? Can metabolism exist without replication, or replicative life exist without metabolism?'. He argues '. . . either life began only once, with the functions of replication and metabolism already present in a rudimentary form and linked together from the beginning, or life began twice, with two separate kinds of creatures, one capable of metabolism without exact replication, the other kind capable of replication without metabolism.'

Actually, the *main* function of DNA in a living cell is to direct the continuous, uninterrupted synthesis of proteins which is essential to the cell's survival, while its function in replication, though of course crucial, occurs by necessity only exceptionally (once in a lifetime!). This fact has been somehow overlooked in self-replication studies, and marks a difference with most self-replicating systems in which the 'blueprint' shows up only in the climatic stage of the drama – the very act of self-replication.

The most orthodox Darwinists might point out that the entire existence of a living cell, and in particular its growth, are necessary and preliminary steps to that single crucial event that is reproduction: in a way, a DNA strand would perform its ultimate function when it unwinds, duplicates itself, and settles in the two daughter cells, while the whole previous humble work, the survival and growth of its 'host' cell, would just be a long, preparatory stage meant to insure the conditions for the safe survival of the duplicated DNA. Even so, however, the fact remains that survival and growth of cells and organisms are essential features of life and as such cannot be overlooked in any study of self-replication. After all, even Roger Rabbit might say to von Neumann 'I self-reproduce only occasionally, sir – most of the time I just live'.

Several authors – especially those working with CAs, from Langton [36] to Lohn and Reggia [206] – liken the blueprint copying–blueprint execution stages of von Neumann-like self-replicating systems to the transcription–translation functions of living cells. However, in biological cells transcription is the step leading from DNA to RNA, and translation is the step from RNA to proteins. This dichotomy does not match that between blueprint copying and blueprint execution. Actually, the closest biological analogue of the blueprint execution step in von Neumann's self-replication is the overall process (transcription + translation) leading from DNA to proteins, while the closest analogue of the final, blueprint-copying step is DNA duplication.

Even in the replication stage proper, the sequence of events in biological cells is quite different than in von Neumann's cellular automaton. *Mitosis* begins with DNA duplication and continues with the spatial separation of the two twin DNA sets towards opposite halves of the cells – before actual cleavage and division occurs [207]. For non-specialists, the clearest description of the mechanism of cell replication and of the role of the genetic code I have ever read is

that given by another non-specialist, Douglas Hofstadter [34]. Thus the analogy should go as follows:

- the blueprint is DNA;
- the universal constructor is the ribosome (or rather the extended system formed by the ribosomes and by any further reaction catalysed by enzymes produced by the ribosomes);
- there is no real supervisory unit, and the synchronisation of the whole process is guaranteed by the laws of physics and (bio)chemistry;
- duplication of DNA comes first, and construction of daughter cells is a continuous process involving the growth of the mother cell, the cleavage and division into daughter cells, and their own growth following division.

14.2 Constructor arms vs. self-assembly

'Universal construction' is achieved in von Neumann's 'machine' by means of a constructor arm. A hardware equivalent of such a device is the 'manipulator arm' of Drexler molecular assemblers (nanomachines) which will be described in a subsequent section, and approximations are the manipulator arms commonly found in current industrial robots. In CA implementations, the idea of a 'constructor arm' was taken up by Langton for his self-replicating 'loops' (see Section 15), and was henceforth inherited by many subsequent self-replicating CA.

All these instances, however, are far removed from the way 'construction' is accomplished in actual living systems. At the smallest scales, the closest approximation is probably the assembling of proteins in ribosomes; these, however, do not actually 'reach' for parts but rely amply on random encounters occurring in a fluid medium, encounters which are made possible with significant rates by the small size of the 'building blocks' (e.g. amino acids). From this point of view, construction in cells resembles more closely the method which would be followed in von Neumann's kinematic model of self-replication. Then, if we turn our attention to structures larger than proteins, we find that in living systems they are never assembled by direct manipulation, but 'come together' spontaneously mainly by self-assembling (see Section 11), a method which, too, relies on small size and rapid diffusion to achieve significant rates, and which is almost completely absent in proposed software or hardware implementations of self-assembling.

14.3 Genomic vs. non-genomic evolution

Kauffman [21, 32, 33] envisages the replication by division of a (membrane-bound) 'autocatalytic set' as occurring without a 'blueprint':

> Numerical studies show that such autocatalytic systems can evolve without a genome. They do so by incorporating fluctuations in the spontaneous reaction graph 'penumbra' surrounding the collectively autocatalytic set. Such fluctuations can bring forth molecular species that are catalyzed from the autocatalytic set and can abet the set itself. When this happens, old molecular species may be 'ejected' from the set (from Lecture 2 in [32]).

Thus, some evolution is regarded as possibly occurring without a genome. However, the rather obscure formulation seems to point to evolution by mutation/selection occurring at the level of macromolecules. It is not clear how mutations thus occurring can be transmitted to the 'offspring' and can be exposed to the effect of natural selection at the level of the individual (the prebiotic system of reacting macromolecules, probably surrounded by a boundary, or 'proto-membrane', and somehow distinct from the surrounding environment).

Von Neumann's arguments in favour of a blueprint-based replication are rejected on the basis that they only hold if a 'universal constructor' is required, while living systems are not universal, but rather highly specialised, constructors: '. . . von Neumann assumed a universal constructor. This assumption then required separate Instructions. But cells and other collectively autocatalytic systems are not universal constructors. Cells are special purpose constructors, ones that happen to create themselves via closure in the appropriate phase space . . .' [32].

Kauffman describes large 'autocatalytic sets' of molecular products and reactions.

Now, the working machinery of actual living cells can indeed be regarded as such an autocatalytic set. The distinction between transcription (enzyme-assisted DNA→RNA step), translation (ribosome-actuated RNA→protein step) and duplication (DNA doubling by complementary base pairing) belongs more to our conceptual framework than to reality. What we have in reality is a continuous flow of complex chemical reactions, which obviously achieve *closure* in that a complete cell is able (a) to achieve homeostasis against a changing environment, and (b) to replicate itself.

The main peculiarity of actual living cells is that, unlike Kauffman's sets, they store all the relevant information in a single copy (the 'blueprint', i.e. DNA). This feature seems to be crucial for the possibility of Darwinian evolution *via* random mutations and natural selection, i.e. it seems to be a crucial prerequisite for evolvability.

Kauffman himself writes of the 'evolution of evolvability'. The DNA-based information storage system adopted by current cells may well be the final outcome of such evolution. The basic assumption is that early protocells evolved in the absence of a nucleic acid-based genome, and coded information storage emerged only later. In this initial phase, there was non-genomic evolution. The rationale for this hypothesis is that nucleic-acid based information coding (as used in today's cells) is too complex to have emerged all of a sudden by random mechanisms. Therefore, some pre-genomic evolution must have preceded (and must have led to) its appearance.

The proposed mechanism of non-genomic, or pre-genomic, evolution is the random emergence of peptide-bond forming proto-enzymes (ligases). Probably these were initially very low-efficiency catalysts. However, some of the peptides generated with the aid of these catalysts could have been better catalysts of peptide bond formation than the proto-enzymes that formed them. Thus, protocells would 'search' the space of all peptides at faster and faster rates.

At the same time, some of the peptides generated in this search would presumably function as proteases, cutting peptide bonds. Since functional peptides tend to have some degree of ordered structure, and since proteases cleave unstructured peptides more rapidly than structured ones, then functional peptides would preferentially survive.

Finally, newly produced peptides would occasionally turn out to be capable of performing some new function which could be integrated into the protocell's metabolism. This process might eventually have generated nucleic acid and their integration with peptides to form a genomic system.

The only prerequisites for the above scenario to be credible are:

- that protocells existed, capable of growing either by acquiring amphiphilic material from the environment, or by producing it internally, and capable of dividing giving birth to 'offspring' protocells, even in the form of imperfect replicas of the 'parent' one;
- that polymers existed capable of performing constructive and destructive processes, and that the cleavage (hydrolysis, destruction) of non-functional polymers was, at least slightly, preferred.

Ghadiri and co-workers [64, 82, 83, 208] produced a self-replicating peptide system with an inherent error-correction mechanism and demonstrated the evolution of populations of peptides. Yao *et al*. [209] constructed a peptide system capable of auto- and cross-catalysis and of generating new (not initially present) self-replicating peptides.

New and Pohorille [210] describe a computational model simulating the essential biochemical features of a non-genomically evolving system. Different peptides are initially present, characterised by their length and their catalytic efficiency as ligases and proteases. Suitable probability distributions describe such efficiencies and account for differences in chemical composition which are not described in detail. When two peptides are joined together, or one peptide is cleaved (hydrolysed) into two smaller fragments, the resulting peptides 'inherit' (in a probabilistic sense) the catalytic efficiencies of their 'parents'. Encounters and chemical reactions are then simulated by a Monte Carlo technique, and the changes in the peptide population are recorded and analysed.

Such simulations show that, provided the rates with which ligases and proteases are formed are in a certain ratio, genomic evolution occurs: the variety of polymers within a protocell, their average length, and their catalytic efficiency all increase over consecutive generations.

The authors observe that, in this prebiotic and non-genomic evolution, the sloppy replication of protein sequences is an asset rather than a drawback, since it allows for a more rapid exploration of the space of possible proteins and for a more rapid discovery of new functions. However, a 'truly advanced' (*sic*) protocell would have to find a more accurate method of transmitting information to the offspring (presumably, the genomic method in use today!).

15 Trivial vs. non-trivial self-replication

> One of the difficulties in defining what one means by self-reproduction is that certain organizations, such as growing crystals, are self-reproductive by any naive definition of self-reproduction, yet nobody is willing to award them the distinction of being self-reproductive. A way around this difficulty is to say that self-reproduction includes the ability to undergo inheritable mutations as well as the ability to make another organism like the original.
>
> (J. von Neumann, 1949, [117])

A rather controversial issue in self-replication is the establishment of a clear boundary between 'trivial' and 'non-trivial' self-replication. The issue was addressed by von Neumann himself [117] and has been discussed, among others, by Langton [36], Poundstone [99], and Perrier *et al*. [141].

If self-replication is seen as *the organisation of 'raw matter' into a 'pattern', induced by the pattern itself*, then some rudimentary form of self-replication is displayed, for example, by:

- crystals (raw matter = atoms, ions or molecules, pattern = crystal);
- zipper tooth pairs (raw matter = open teeth, pattern = closed zipper);
- domino tiles (raw matter = upright tiles, pattern = fallen tiles).

The above examples are perhaps a bit lacking because it is difficult to identify the *individual* that is being replicated, since it does not possess clear boundaries. More convincing are perhaps the following two examples, respectively pertaining to 'virtual' and 'real world' self-replication.

The first example, taken from the field of CAs, is the (self?) replication of any given pattern in the 1357/1357 variant of Life. In this cellular automaton, a cell will be ON at generation $k + 1$ if, and only if, it has 1, 3, 5 or 7 (i.e. an odd number of) ON neighbours at generation 'k'. Surprisingly, the result of such a simple rule is the spreading and multiplication of any arbitrary initial pattern, as exemplified in Fig. 9.

The second example, belonging to the field of 'real world' self-replication, is the block contraption devised by Lionel Penrose [38], presented in a previous section and sketched in Fig. 4. In this case, the 'raw matter' is made of loose blocks of two types, whereas the self-replicating

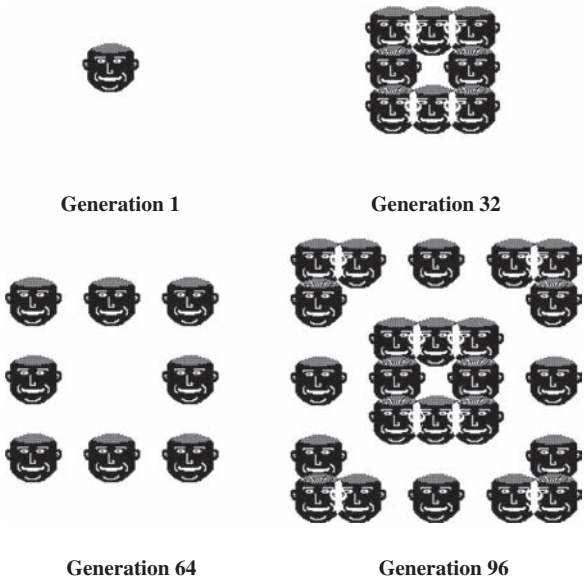


Figure 9: Replication of an arbitrary initial pattern in the 1357/1357 variant of Life.

'pattern' is the couple of two blocks locked together, which 'catalyses' the fusion of further blocks into locked couples.

Of course all the above examples – including the last two – leave a deep feeling of dissatisfaction as instances of self-replication. In fact, they lack at least two of the features that we expect to find in 'real' self-replicating systems such as living cells:

- the self-replication process is entirely coded within the 'physics' of the systems (either 'real world' physics as in crystals, domino tiles, zippers or Penrose's replicators; or CA rule sets, as in the Life variant discussed above), rather than being actively directed by the structure that is being replicated;
- the replicated 'pattern' is only slightly more complex than the 'raw matter' from which it emerges.

The former aspect was emphasised, for example, by Perrier *et al*. [141]. Sensible as it may look at first sight, it actually raises very subtle issues pertaining to the distinction between text and context, or system and environment, or even 'self' and 'non-self', and to the existence of different levels of description for the same phenomenon.

Take, for example, a cellular automaton like Langton's 'loop'. Any behaviour of such a structure will eventually result from the application of simple transition rules governing the change in state of an individual cell as a function of the state of its neighbours. It is true that we can describe its self-replicating behaviour in terms of higher-order structures, as in the following excerpt: 'The structure, embedded within an 8-state CA space, consists of a looped pathway, containing instructions, with a construction arm projecting out from it . . . Upon encountering the arm junction, the instruction is replicated, with one copy propagating back around the loop again and the other copy propagating down the construction arm, where it is translated as an instruction when it reaches the end of the arm' [114]. However, it would be possible, although certainly pointless, to recast the above sentences in terms of state transitions of individual cells. In this description, there

would be no 'pathway', no 'construction arm', no 'instructions' and no 'propagation' – indeed no 'self', just a two-dimensional cell space in which some low-level activity goes on.

This objection against a simplistic differentiation of 'trivial' versus 'non-trivial' replicators on the basis of their degree of 'self-direction' does not apply only to CAs and other virtual entities. Also the functioning of a living cell, once observed at a molecular level, shows nothing but the unfolding of chemical reactions in accord with strict laws of physics, and only by switching to a higher level of description we can identify such abstract entities as an information-carrying 'genome' and processes as 'transcription' and 'translation'. Thus, unless we already have an idea of the correct level of description at which to look, we cannot distinguish between a 'self-directed' and a 'forced' replication process.

We will come back to these subtle issues in the next section, in the context of discussing the question of the 'epistemic cut' between symbols and physics, as raised, for example, by Pattee.

The latter issue, i.e. the lack of a clear difference in complexity between 'raw matter' and replicated 'pattern' in 'trivial' self-replication, is probably more amenable to a quantitative treatment. In fact, methods to quantify the complexity and the information content of both natural and artificial structures have been devised since the early days of information science [211–214].

If we identify the complexity of a structure with its information content (say, the minimum number of bits required for a complete description), then it seems unquestionable that a closed zipper, a fallen domino tile, or a couple of interlocked Penrose blocks, have little (if any) more complexity than an open zipper, an upright tile, or two disjointed blocks, respectively. By contrast, a living cell is astronomically more complex than the 'soup' of molecules from which it can assemble a replica of itself.

The issue of self-replication as an addition of complexity, or information, to pre-existing components can also be viewed as a problem of individual–environment relationship. In *The Theory of Self-reproducing Automata* von Neumann [7] wrote:

> . . . living organisms are very complicated aggregations of elementary parts, and by any reasonable theory of probability or thermodynamics highly improbable. That they should occur in the world at all is a miracle of the first magnitude; the only thing which removes, or mitigates, this miracle is that they reproduce themselves. Therefore, if by any peculiar accident there should ever be one of them, from there on the rules of probability do not apply, and there will be many of them, at least if the milieu is reasonable.

Now, the apparently harmless last clause may be taken as the starting point of far-reaching considerations. One moment's reflection shows that, if the milieu – the environment – is very reasonable, than it is easy to build a self-replicating machine. Just think, following Freitas and Gilbreath [87], of a 'robot' which includes, as its essential component, a fuse. If such a robot is released in an 'environment' consisting, as available 'raw matter', of loose fuses and loose fuseless robots, all it has to do in order to 'self-replicate' is inserting a fuse into its appropriate slot in a fuseless robot. Surely, designing a robot with this sole ability should not be too difficult . . . Of course, few of us would acknowledge such a behaviour the status of self-replication proper.

The same authors observe that the degree of 'non-triviality' of self-replication depends on how complexity and information are distributed between the self-replicating entity (the 'machine') and its environment (the 'substrate'), ranging between two extreme cases:

- at one extreme, all complexity and information reside in the substrate and the action of the machine in making a new machine is trivial, as in the above example of robots and fuses;

- at the opposite extreme, all complexity and information reside in the machine and the substrate is utterly undifferentiated (a 'plasma').

Even von Neumann's kinetic automaton 'feeds' on component parts that just happen to have been manufactured to exact specifications and organised in bins. Strictly speaking, no advanced life form on earth self-replicates, because we depend on the rest of ecosystem to survive. Partial exceptions are those organisms – bacteria and plants mostly – that do not depend on other organisms for their subcomponents (proteins, carbohydrates, etc), but instead can self-replicate from raw energy and simple chemicals. However, any living cell operates in a milieu which is far from being an amorphous 'plasma', and would actually die in a totally undifferentiated environment. In the field of artificial devices, the closest approximation to complete self-replication is probably the 'lunar factory' described above.

In the context of CAs, Lohn and Reggia [206] propose a rather axiomatic definition of non-trivial self-replication, based on the following criteria:

1. S is a structure comprised of more than one non-quiescent cell, and changes its shape during the self-replication process;
2. replicants of S, possibly translated and/or rotated, are created in neighbour-adjacent cells by the structure;
3. for any positive integer N and for infinite cellular spaces, there must exist a time t_N such that S can produce N or more replicants within t_N;
4. the first replicant becomes detached from the parent structure at some time t_D.

The authors add that 'the issue of triviality was circumvented in early models by requiring universal computation and universal construction ... starting with Langton (1984), more recent models, inspired by biological cells, have abandoned this requirement by insisting that an identifiable instruction sequence be treated in a dual fashion: interpreted as instructions (translation), and copied as raw data (transcription)'.

In conclusion, no clear-cut distinction can be made between 'trivial' and 'non-trivial' self-replication. It is the existence of a symbolic level (self-description), in addition to some minimum complexity, that better characterises 'non-trivial' systems [96].

16 Epistemic cut and semantic closure

> Traditional philosophy sees [the relation between material and symbolic aspects] as the question of reference, or how symbols come to stand for material structures ... I have always found the complementary question, of how material structures ever came to be symbolic, much more fundamental.
>
> (H.H. Pattee, 1995, [215])

The act of measurement in physics requires that a *cut* be made between the system S on which the measure is to be taken, and the measuring apparatus M. This cut expresses an *epistemic* necessity and not an *ontological* condition; it does not reside in the things themselves, but nevertheless it is a necessary condition for knowledge to be acquired, i.e. for the measurement to produce a result.

Von Neumann [216] discussed the necessity of the epistemic cut in measurement '... we must always divide the world into two parts, the one being the observed system, the other the observer. The boundary between the two is arbitrary to a very large extent ... but this does not change the

fact that the boundary must be put somewhere . . .'. We would be tempted to include M in the physical system to be measured, and to describe the act of measurement as a purely physical process, with no epistemic cut, involving the overall system S + M. But, in order to do so, we must measure some property of (S + M), which implies a second measurement device M′ distinct from (S + M) . . . and so on. Note that the problem is specially known in quantum mechanics, where it takes the form of the wave function collapse and enters the very formalism of the laws, but is still present also in purely classic measurements.

A perfectly analogue problem exists in the case of control. The possibility itself of describing a physical phenomenon in terms of control and information processing postulates an epistemic cut between the controlled system S and the controlling entity C. A similar infinite-regress problem arises if we try to include C into the system and describe the control process as the physical behaviour (with no cut) of S + C. This, in fact, still requires some control to be exerted on (S + C) by means of some apparatus C′ . . . and so on.

Actually, measurement and control are dual processes: measurement transforms physical states into information (symbols), while control transforms information (symbols) into physical states. In the former, in order to measure something without having to measure everything, we must reduce the degrees of freedom of the measuring device to the few that are semantically relevant for the measurement (the physical state of a thermometer collapses to a single quantity – a temperature – if we are using it to measure temperatures; the thermometer acquires again the status of a complex physical system only if we want to study thermometers). Similarly, in control we have to reduce the degrees of freedom of the controlling apparatus to a few relevant variables which admit a symbolic expression (a toggle in a machine can only be 'ON' or 'OFF', which is a highly simplified but semantically sufficient description of its actual state).

The epistemic cut is of the same nature as the distinction between hardware and software and, to some extent, as the distinction between mind and body. Its necessity makes itself apparent whenever we have to design a complex artefact, to analyse its functioning, or to explain it to an audience. Doing so without ever mentioning subject/object, controller/controlled, meter/measure distinctions, though perhaps possible in principle, would be absurdly difficult and would not qualify as proper understanding. The epistemic cut is thus similar to what Daniel Dennett [217] calls the 'intentional stance', i.e. the attribution of purposeful intentions to the subjects of an action as a powerful method for compressing our descriptions. 'It was raining, therefore I opened my umbrella': any attempt to describe this simple state of affairs without postulating that some purposeful action is taking place would be doomed from the start.

Probably no one has analysed the nature, meaning and implications of the epistemic cut with greater insight and lucidity than Howard Pattee [46, 215, 218, 219].

Pattee [46] defines a symbol as 'a relatively simple material structure that, while correctly describable by all "normal" physical laws, has a semantic function (significance) that is not describable by these laws'. The apparent paradox of a physical structure which nevertheless cannot be completely described by physical laws is solved by Pattee invoking the difference between laws and boundary conditions, or constraints.

Laws and constraints are of the same nature. However, laws express highly correlated events whose description can be highly compressed, whereas constraints and boundary conditions express loosely correlated, apparently gratuitous facts whose description cannot be significantly compressed. Murray Gell-Mann [220] calls this incompressible information 'frozen accidents', of which DNA is an illustrious example. A symbol is basically such a constraint (more subtle distinctions could be made between constraints and boundary or initial conditions, but we will omit these fine-grained differences in the present discussion).

In terms of control theory, there are state variables, obeying inflexible physical laws, and control variables, expressing constraints; the latter account for all the flexibility observed in the system.

Are symbols independent of physical laws? A closely related question (discussed, for example, by Hofstadter [34] as applied to the genetic code) is: are symbols completely arbitrary? To both questions the answer is not clear-cut – it is rather a matter of extent. A digital computer is the most familiar example of almost totally symbolic control. Maintaining data and program in memory *does* require some physical action (e.g. a power supply and a thermal sink), but these are non-specific actions which constrain very little the arbitrariness of the code. To the user, a computer appears as a 'physics-free' device, a syntactic machine based only on arbitrary symbolic rules.

The epistemic cut is not necessarily associated with a physical boundary. There are complex systems in which the distinction between hardware and software rests on a purely logical ground, and does not imply the possibility of singling out corresponding physical sub-systems.

At the Science Museum in London, visitors can behold a Fourier analyser designed in the 19th century by Lord Kelvin. It is an impressive contraption, the size of a large lathe, mostly made of shining brass. The user grabs a special stylus connected with the machine's innards, moves it along an arbitrary curve drawn on a sheet of paper which rests in a precise position on a tray, and, hey presto, six dials indicate the amplitudes of the first six harmonics in the Fourier expansion of the function represented by the curve.

In this, as in all analogical computing devices, there is no obvious distinction between software and hardware. The software – the purposeful design – is, so to speak, distributed throughout the bulk of the machine, reflected in each of the thousand minute gears of which it is made up.

The spread of digital computers and integrated processors has accustomed everybody to a different architecture, and alternative (Kelvin-like) designs now look strange and outdated. All logical and control functions are now usually demanded to segregated and specialised sub-systems, in which software has the better hand on hardware and whose size is but a tiny fraction of the whole. We build machines big and dumb and then we control them by small, smart computers.

In such digitally controlled devices, the epistemic cut naturally coincides with the physical boundaries separating the brutish bulk from the cute brain (strictly speaking, identifying a logical cut with a physical one is a categorical error, but it captures well enough our common perception of the issue).

It should be kept in mind that, even in completely analogical devices, the epistemic cut is still there – only it is not associated with precise physical boundaries. Also, a purpose-directed artefact is not found in the street, it has to be designed; and in the design stage the symbol-reference duality must necessarily jump out with clarity, or design itself would be impossible.

Let us now focus our attention on systems that exhibit self-replication, starting with those that are certainly lavishly available around us, i.e. living organisms.

In principle, a living organism, e.g. a *cell*, might well be a Kelvin-like analogical machine. *Per se*, the processes of self-maintenance and self-replication that we regard as the characteristic hallmarks of life could be accomplished without an explicit distinction between hardware and software. For example, although the issue was not fully explored by the author, the collectively autocatalytic sets envisaged by Kauffman [21] as the paradigms (and likely forerunners) of life would be such Kelvin machines.

However, real life has chosen a different route. While some aspects of life, e.g. sensorimotor activity and maybe even some cognition phenomena, can be described by a temporally coded dynamics with no static symbolic structures, it is a fact that all the functions of a cell which pertain to self-maintenance and self-reproduction are demanded to clearly identifiable, morphologically and chemically distinct, structures, of which the most prominent is the DNA-based genome. It is

a fact, in other words, that Nature has chosen the route of developing a clear distinction between genotype and phenotype.

The genotype–phenotype distinction is indeed an epistemic cut of the kind discussed above. Actually, the main distinguishing trait of living organisms is that they realise *within themselves* this epistemic cut between controller and controlled, subject and object, knower and known. 'Metaphorically, life is matter with meaning. Less metaphorically, organisms are material structures with memory by virtue of which they construct, control and adapt to their environment' [218]. Also, since the memory structures that constitute the blueprint of the organism are themselves built by the organism's molecular machinery, living beings attain what can be defined a semantic closure of the subject–object loop [46].

Symbols can only exist in systems with a function, including living organisms and artefacts. Thus, speaking of the epistemic cut before life appeared would be meaningless. Pattee's sentence in the epigraph of this section point to this fundamental fact: that in living beings some specific material structures (the genome) have been invested of a symbolic role. Evolution has created a language, whose grammar – notably, the genetic code – is basically the same throughout life on this planet, with only minor, 'local' differences.

Thus, a living organism, and in particular a cell, can be regarded as the realisation of what Pattee calls a semiotic system. This is the totality composed by:

- a discrete set of quiescent symbol vehicles (DNA);
- a set of interpreting structures, playing the role of constraints (RNA, ribosomes and enzymes implementing the genetic code);
- a system in which an interpreted symbol has some function (the cell as a whole).

It is important to keep in mind that, in order to perform its symbolic function, the genotype must consist of quiescent molecular structures and must involve energy-degenerate, rate-independent configurations.

The importance of quiescent structures, which are not altered by the functioning of the system and thus can act as a readable memory, was already pointed out in Section 8 in discussing possible alternatives to the basic von Neumann architecture for a self-replicating automaton.

Energy-degenerate means that the alternative configurations encoding the genetic information must be substantially equivalent as regards energy level, chemical properties, etc., so that the choice between alternative configurations is not derivable (almost) from the physical laws. If the presence of one or another of two alternative bases at a given location of a DNA strand implied a significant difference in the overall stability, reactivity etc. of the strand, then we could not speak of symbolic control.

This latter requirement is closely connected with the question asked by von Neumann [7] (see Section 8) regarding why 'the molecules or aggregates which in nature really occur in these parts are the sort of things they are . . .'. It has been argued that only sufficiently heavy macromolecules can provide a large number of alternative configurations with roughly equivalent energy levels and overall chemical properties.

The issue of protein size is also discussed by Ji [91]. Ji invokes the cybernetic principle known as the 'law of requisite variety' [12], which states that the number V_E of different environmental conditions (variety of the environment), the number V_M of the possible internal states of a machine (variety of the machine) and the number V_O of the possible outcomes of interactions between machine and environment (variety of the outcomes) must satisfy the inequality

$$V_O V_M \geq V_E.$$

Thus, the only way of maintaining the number V_O of outcomes low – i.e. of achieving substantial homeostasis – under a broad range of environmental conditions (V_E) is to increase the number of internal states (V_M). This, in turn, requires large macromolecules.

Finally, from an evolutionary point of view, large proteins are necessary so that the many internal degrees of freedom provide a smooth dependence of the folded structure upon amino acid sequence – otherwise most mutations would throw the phenotype farther off in the fitness landscape, preventing smooth evolution (this is what happens most of the times in virtual worlds like 'Core Wars' and, partly, Tierra, Avida and similar environments [21, 32, 33]). One might say that only large macromolecules can exhibit fluidity in the sense of Hofstadter [34].

Thus, the molecules of life could *not* be much smaller than what they are ('miniaturize this', De Niro might say – and even Feynman would have to give up).

A cell can be defined in two distinct ways:

- first, it is the well defined portion of matter delimited by a membrane boundary, so that molecular exchanges within the cell dominate by far on exchanges across the membrane, and that these latter occur under strictly controlled conditions;
- second, it is the portion of organic matter associated with a single, individual copy of the genome. This 'one membrane, one genome' rule is almost universal, the only exceptions occurring when (a) a cell divides (two identical copies of the genome can then temporarily coexist within the same membrane boundary), and (b) a gene-deficient cell (such as a mammalian red blood cell) is created. This latter case can be disregarded by observing that such particles are *not* cells proper since they lack the ability to self-reproduce.

A remarkable fact concerning life is that the two above definitions actually coincide. This fact is not obvious, and its significance has often been overlooked. A cell (in the sense of a membrane-bounded particle) might well have more than one active copy of the genetic material. In principle, nothing prevents the cell's molecular machinery (including messenger- and transfer-RNA, ribosomes, and enzymes) from simultaneously operating on multiple copies of the cell's DNA: certainly the constructive functions of this machinery would not be impaired.

The necessary uniqueness of the genetic information associated with a cell can only be understood (a) by considering the mechanism of evolution, and (b) by considering the cell as a control system, either in its normal state of metabolic, self-preserving activity or in its occasional state of self-reproduction.

As regards (a), evolution requires:

- the possibility of variations (mutations);
- the inheritability of these variations.

It is difficult to figure out how these two requirements might be simultaneously satisfied unless the cell's blueprint exists in a single copy.

As regards (b), it is difficult to imagine how the genetic material could perform its regulatory functions without conflicts and instabilities if multiple copies of the genome were simultaneously active throughout the cell. The reproduction process, in particular, is one in which a rigid spatial organisation of the cell's parts is strictly required. Not only the genetic material but also other structures (e.g. the centriole) play a precise, individual role during this process, and their hypothetical multiplication would have disastrous consequences.

What lessons can we learn from nature regarding the possibility of self-replication in artefacts, and the best means to achieve it?

Certainly it is not casual nor irrelevant that, after billions of years of evolution, trials and errors, nature has stuck to a single, well-defined strategy for the preservation and the spreading of life. Summarising the considerations developed above, this strategy can be characterised by:

- the existence of a genotype–phenotype distinction ('internal' epistemic cut), in virtue of which certain physical structures acquire a symbolic value, and semantic closure is attained;
- the adoption of 'constructive' symbolic coding, in which symbols (memory units) represent not the description of corresponding physical structures, but rather the *constructive steps* necessary to assemble them (a recipe rather than a blueprint);
- the presence of membrane-bounded compartments, or *cells*, as irreducible living units, and the storage of a *single copy* of the genetic 'recipe' in each compartment.

It is true that this strategy was already present in the earliest known fossil micro organisms (see Section 4), and thus must have been selected within the 500 million years at most elapsed since the Earth began a habitable place. It is also true that, once the first successful life architecture filled most of the available ecological niches, rival forms didn't have much of a chance. Therefore, we cannot be absolutely certain as to whether the above requirements are strictly indispensable for life and open-ended evolution, or alternative strategies are in principle possible. Studies of artificial life and progress in the development of self-replicating artefacts will certainly help clarify these, and many other, issues.

In the meantime, what we know for sure is that this strategy has worked; life as we know it has enjoyed since its first appearance a remarkable success and has withstood the harshest of predicaments.

Surely, it is not casual either that very similar architectural choices, at least as far as the hardware–software distinction is concerned, are found in the most complex human artefacts. Perhaps it is not strictly a matter of independent convergence of design: both the basic concept of the universal Turing machine and the von Neumann (software–hardware) architecture of current computers may have been directly inspired by what was being discovered in the same years about life and its organisation (the reverse is probably also true: the search for a deep, symbolic-level, layer of structures controlling maintenance and reproduction in living organisms may have been partly inspired by contemporary advances in computation theory, and the history of the interaction between these two fields of knowledge during the first half of the last century has not been completely written yet).

It is difficult, however, to imagine what alternative route we might have followed: without the explicit distinction between hardware and software and the physical separation between the 'bulky', 'dumb' parts of a machine and its 'smart' control devices, the design of artefacts as complex as even the most primitive of current industrial robots would have been nightmarishly more difficult. Purely analogical devices, it seems, can reach only a certain level of complexity (probably in the range of Lord Kelvin's Fourier analyser), beyond which some level of symbolic control and some internal epistemic cut become necessary.

At the same time, it is not sufficiently appreciated that a purely formal model of self-replication is only part of the problem (as von Neumann's remark in the epigraph of Section 8 reminds us). The difficult part stays in the details, and in the material, physical details as to that. The appreciation of this point in robotics, for example, has increased as more and more real-world problems were tackled. Hundred years ago, if asked what would be more difficult for a machine, (a) crossing the road or (b) playing chess at the grand master level, few would have answered 'a'. Yet, we now have many machines that easily defeat the (human) world chess champion, but few (if any) that can cross the road (granted, this task has become more difficult in the last hundred years).

17 Epilogue

> History arises when the space of possibilities is too large by far for the actual to exhaust the possible.
>
> (Stuart Kauffman, 1995, [21])

In [171] the authors observed 'the differences between biological systems and engineering, man-made devices, which appear quite large if we look at them "from a distance", tend to shrink as we proceed towards the nano-scales. At large scales, biological tissues are "wet", "soft" and "fuzzy", whereas mechanical or electronic equipment is "dry", "hard" and "clear cut"; however, even a living cell, once adequately resolved . . . decomposes into beautifully complex, but rather conventional, pieces of machinery (the opposite is also, at least in part, true: a Cray-2 computer, on the whole, looks quite wet and even bubbling!). It is the level of organisation and complexity, and not an intrinsic, irreducible extraneousness, that differentiates biological and man-made devices'.

This remark is certainly true, but, with hindsight, it seems to me that it is missing a point. What makes living matter 'wet, soft and fuzzy' – in a word, the fluidity of living matter – is the fluidity of the swarm. Living matter is fluid for the same reason why fluids are fluid: because it consists of myriads of particles. In a word, because it is very, very much complex (true, numerosity does not amount to complexity – but it surely helps!).

To come back to Dr. Krutch with whom we started: when we consider the geometric, crystal-like symmetry of an artificial device – say, a gearbox – and seek its equivalent in biological systems, we commit an error of perspective if we look with our naked eye, or even with Dr. Krutch's optical microscope: this will only show us colloidal jelly. It is farther, farther down that we must look, there in the sub-micron scale, where only the electron microscope will take us: then the equivalents of the gearbox will make themselves visible in the form of the bacterial flagellar motor, the pores of the nuclear membrane, or the ribosomal machinery steadily crunching RNA tapes. And by converse, if we zoom back from these scales up to the cell level, we will have made such a journey as would take us from the gearbox to the factory and the city.

Trim gearbox with trim ribosome, fluid cell with fluid city: this is the correct comparison. Any serious attempt to mirror essential features of life in artificial devices (real or virtual) is doomed unless it is ready to cope with this degree of complexity – a complexity that, as has been argued at length in previous sections, is necessary, and not gratuitous.

The feature called self-replication is, of course, no exception. Thus, the first lesson to be learned (partly from the scientific knowledge acquired in the last half century, and partly from the much more ancient knowledge gathered by life during billions of years) is that only by stacking several levels of increasing complexity between the whole and the ultimate simple parts (be these molecules, gears or logical gates) we will be able to build artificial systems capable of homeostasis, self-maintenance and self-replication in a complex environment. This holds for real and virtual systems alike: the several billion molecules of a typical living cell match the billions of active squares probably required to assemble a self-replicating pattern in Conway's Life. Neither molecules nor Life squares were made for self-replication, and this can only be achieved as an emergent property requiring levels on levels of organisation and complexity.

The second lesson is that only systems provided with a complete semiotic structure (a set of symbol vehicles, a set of interpreting devices, and a set of codes and rules to give symbols function and meaning) are likely to achieve non-trivial self-replication. This semiotic structure will not necessarily take the form of the software/hardware distinction, and will not necessarily imply the existence of separate control subsystems, but surely will recognisably be there, at least in the stage of *designing* such a system.

The third lesson is that real-world self-replication is much more difficult to achieve, but also much more rewarding, than self-replication in *virtual* contexts. Research on virtual life, of course, will continue to give important indications, but only if it will manage to incorporate more and more of the physical complexity of the real world.

The last lesson is that self-replication cannot be separated from homeostasis and self-maintenance, on one side; and open-ended evolution, on the other side. In living organisms, the structures that provide for reproduction are the same that provide for metabolic activity and self-maintenance (autopoiesis) and that also allow evolutionary changes by exposing individuals to the pressure of natural selection. Probably machines, too, will attain satisfactory self-maintenance and self-replication when they will also attain the ability of evolving.

In works like this, touching sensitive topics such as life, reproduction and nature-mimicking technology, it has become customary in recent years to include a section on philosophical and ethical issues. As the reader will have noticed, in this work philosophical aspects are only marginally touched (in Section 16), and ethical aspects are not discussed at all.

This by no means implies that I am underestimating the relevance of these problems. Simply, I am convinced that philosophical and ethical issues should only be discussed on the basis of an adequate knowledge and understanding of the more technical and scientific aspects. In Italy, it has become customary to have the most delicate matters of bioethics, politics and technology discussed in TV talk shows by actors, starlets, singers and priests (none of whom usually exhibits the least understanding of the real points at issue), occasionally mixed with one or two experts. This, I am afraid, has not helped much.

I can only hope that this somewhat lengthy survey has helped to improve somebody's understanding of some fascinating, controversial and, surely, delicate problems of science and technology. Preparing it has certainly improved my understanding – for which I am grateful to Prof. Michael W. Collins, who persuaded me to put it down in writing. I am also indebted with my colleagues and students at the Dipartimento di Ingegneria Nucleare in Palermo, who have often exchanged views on this or that aspect and have patiently waited for me to run a *quine* or play with Life while more mundane business (like meetings and exams) was pressing.

References

[1] Krutch, J.W., The colloid and the crystal. *The Best of Two Worlds*, William Sloane Associates: New York, 1950.

[2] Mach, E., *Knowledge and Error: Sketches on the Psychology of Enquiry*, D. Reidel Publishing Company: Dordrecht, NL, 1976 (1st German as *Erkentniss und Irrtum*, 1905).

[3] Perutz, M.F., New X-ray evidence on the configuration of polypeptide chains. *Nature*, **167**, pp. 1053–1058, 1951.

[4] Watson, J.D. & Crick, F.H.C., A structure for deoxyribose nucleic acid. *Nature*, **171**, pp. 737–738, 1953.

[5] Miller, S. & Orgel, L., *The Origins of Life on the Earth*, Prentice-Hall: Englewood Cliffs, NJ, 1974.

[6] Prigogine, I., Schrödinger and the riddle of life. *Molecular Theories of Cell Life and Death*, Chapter 2, ed. S. Ji, Rutgers University Press: New Brunswick, New Jersey, pp. 238–242, 1991.

[7] von Neumann, J., *The Theory of Self-reproducing Automata*, ed. A.W. Burks, University of Illinois Press: Urbana, Chicago, IL, 1966.

[8] Thom, R., *Stabilité Structurelle et Morphogenèse – Essai d'une Théorie Général des Modèles*, InterEditions SA: Paris, 1972.
[9] Thompson, D'Arcy W., *On Growth and Form*, Cambridge University Press: Cambridge, UK, 1942 (First published in 1917).
[10] Schrödinger, E., *What is Life? The Physical Aspect of the Living Cell*, Cambridge University Press: Cambridge, 1944 (republished by The Folio Society, London, 2000).
[11] Wiener, N., *Cybernetics; or, Control and Communication in the Animal and the Machine*, John Wiley & Sons: New York, 1948.
[12] Ashby, W. Ross, *An Introduction to Cybernetics*, Chapman & Hall: London, 1956.
[13] von Bertalanffy, L., *General System Theory: A New Approach to Unity of Science*, John Hopkins Press: Baltimore, 1951.
[14] Prigogine, I., Thermodynamics of irreversible processes. *Applied Mechanics Reviews*, **5**, pp. 193–195, 1952.
[15] Laithwaite, E., *An Inventor in the Garden of Eden*, Cambridge University Press: Cambridge, UK, 1994.
[16] Freitas, R.A. Jr. & Gilbreath, W.P., (eds) *Advanced Automation for Space Missions: Proceedings of the 1980 NASA/ASEE Summer Study*, NASA Scientific and Technical Information Branch, Conf. Publication 2255, Washington, 1982 (http://www.islandone.org/MMSG/aasm/). See in particular Chapter 5, *Replicating Systems Concepts: Self-replicating Lunar Factory and Demonstration*.
[17] Sipper, M., Studying artificial life using a simple, general cellular model. *Artificial Life*, **2(1)**, pp. 1–35, 1995.
[18] Casti, J.L., *Paradigms Lost: Tackling the Unanswered Mysteries of Modern Science*, HarperCollins: New York, 1990.
[19] Casti, J.L., *Paradigms Regained: A Further Exploration of the Mysteries of Modern Science*, HarperCollins: New York, 2000.
[20] Monod, J., *Chance and Necessity*, Fontana: London, 1972.
[21] Kauffman, S.A., *At Home in the Universe*, Oxford University Press: Oxford, 1995.
[22] Jacob, F., *The Possible and the Actual*, Pantheon Books: New York, 1982.
[23] Ray, T.S. & Hart, J., Evolution of differentiated multi-threaded digital organisms. *Artificial Life VI*, eds. C. Adami, R. Belew, H. Kitano & C. Taylor, pp. 295–304, MIT Press: Cambridge, MA, 1998.
[24] Dawkins, R., *River Out of Eden*, HarperCollins/Basic Books: New York, 1995.
[25] Dawkins, R., *The Blind Watchmaker*, Norton: New York, 1987.
[26] Taylor, T.J., *From Artificial Evolution to Artificial Life*, Ph.D. thesis, University of Edinburgh: UK, 1999 (http://www.dai.ed.ac.uk/homes/timt/papers/thesis/html; http://www.timandpete.com/tim/daiweb/research.html).
[27] Varela, F.J., Maturana, H.R. & Uribe, R., Autopoiesis: the organization of living systems, its characterization and a model. *BioSystems*, **5**, pp. 187–196, 1974.
[28] Varela, F.J., *Principles of Biological Autonomy*, North Holland, Amsterdam, 1979.
[29] Eigen, M. & Schuster, P., *The Hypercycle*, Springer-Verlag: Berlin, 1979.
[30] Kauffman, S.A., Autocatalytic sets of proteins. *Journal of Theoretical Biology*, **119**, pp. 1–24, 1986.
[31] Kauffman, S.A., *The Origin of Order*, Oxford University Press: Oxford and New York, 1993.
[32] Kauffman, S.A., *Investigations: the Nature of Autonomous Agents and the Worlds they Mutually Create*, Santa Fe Institute Preprint: 1996 (http://www.santafe.edu/sfi/People/kauffman/Investigations.html).

[33] Kauffman, S.A., Even peptides do it. *Nature*, **382**, pp. 496–497, 1996 (see also reprint in http://www.santafe.edu/sfi/People/kauffman/sak-peptides.html).
[34] Hofstadter, D.R., *Gödel, Escher, Bach: an Eternal Golden Braid*, Basic Books: New York, 1979.
[35] Stauffer, A., Mange, D., Goeke, M., Madon, D., Tempesti, G., Durand, S., Marchal, P. & Nussbaum, P., FPPA: a field-programmable processor array with biological-like properties. *Proc. Fourth Reconfigurable Architectures Workshop – RAW 97*, eds. R.W. Hartenstein & V.K. Prasanna, IT Press Verlag: Bruchsal, pp. 45–48, 1997.
[36] Langton, C.G., Self-reproduction in cellular automata. *Physica D*, **10**, pp. 135–144, 1984.
[37] Mange, D., Madon, D., Stauffer, A. & Tempesti, G., Von Neumann revisited: a Turing machine with self-repair and self-reproduction properties. *Robotics and autonomous systems*, **22**, pp. 35–58, 1997.
[38] Penrose, L.S., On living matter and self-replication. *The Scientist Speculates: An Anthology of Partly-baked Ideas*, eds. I.J. Good, A.J. Mayne & J. Maynard Smith, Heinemann: London, 1962.
[39] Freitas, R.A. Jr., Building Athens without the slaves. *Technology Illustrated*, **3**, pp. 16–20, 1983 (http://www.rfreitas.com/Astro/BuildingAthens1983.htm).
[40] Drexler, K.E., Machines of inner space, *1990 Yearbook of Science and the Future*, Encyclopedia Britannica, Inc., 1989 (reprinted as Appendix A in *Nanotechnology – Research and Perspectives*, eds. B.C. Crandall and J. Lewis, MIT Press: Cambridge, MA, 1992).
[41] Dennett, D.C., *The Intentional Stance*, MIT Press: Cambridge, MA, 1987.
[42] McClendon, J., The origin of life. *Earth Science Reviews*, **47**, pp. 71–93, 1999.
[43] Barbieri, M., *The organic codes – An Introduction to Semantic Biology*, Cambridge University Press: Cambridge, UK, 2002 (http://www.biologiateorica.it/organiccodes/).
[44] Orgel, L.E., *The Origins of Life: Molecules and Natural Selection*, John Wiley & Sons: New York, 1973.
[45] Miller, S.L., A production of amino acids under possible primitive earth conditions. *Science*, **117**, pp. 528–529, 1953.
[46] Pattee, H.H., The physics of symbols: bridging the epistemic cut. *Biosystems*, **60**, pp. 5–12, 2001 (special issue *Physics and Evolution of Symbols and Codes*).
[47] Schopf, J.W., The oldest fossils and what they mean. *Major Events in the History of Life*, ed. J.W. Schopf, pp. 29–63, Jones and Bartlett: Boston, 1992.
[48] Crick, F., *Life Itself*, Simon & Schuster: New York, 1981.
[49] Hoyle, F., *The Intelligent Universe*, Michael Joseph: London, 1983.
[50] Deamer, D. & Fleischaker, G., (eds), *Origins of Life: The Central Concepts*, Jones & Bartlett: Boston, 1994.
[51] Chadwick, A.V., *Abiogenic Origin of Life: A Theory in Crisis*, Earth History Research Center web pages, 2001 (http://origins.swau.edu/papers/life/chadwick/default.html).
[52] Darwin, C., *The Origin of Species*, 1859 (reprint of the first 1859 edition by Penguin Books: London, 1968).
[53] de Duve, C., The beginnings of life on Earth. *American Scientist*, September–October 1995 (http://www.americanscientist.org/template/AssetDetail/assetid/21438).
[54] Zindler, F., *How did Life Begin?* Round Earth Society web pages, 2001 (http://www.str.com.br/English/Scientia/life.htm).
[55] Sarfati, J., Self-replicating enzymes? *Creation Ex Nihilo Technical Journal*, **11(1)**, pp. 4–6, 1997.
[56] Cairns-Smith, A.G., *Genetic Takeover and the Mineral Origin of Life*, Cambridge University Press: Cambridge, MA, 1982.

[57] Maturana, H.R. & Varela, F.J., *Autopoiesis and Cognition: the Realization of the Living*, D. Reidel Publishing Company: Dordrecht, Holland, 1980.
[58] Dyson, F., *Origins of life*, Cambridge University Press: Cambridge, 1985.
[59] Cousins, G.R.L., Poulsen, S.-A. & Sanders, J.K.M., Molecular evolution: dynamic combinatorial libraries, autocatalytic networks and the quest for molecular function. *Curr. Opin. Chem. Biol.*, **4(3)**, pp. 270–279, 2000.
[60] Gilbert, W., The RNA world. *Nature*, **319**, p. 618, 1986.
[61] Schwartz, A.W., The RNA world and its origins. *Planet. Space Sci.*, **43(1/2)**, pp. 161–165, 1995.
[62] Bartel, D.P. & Unrau, P., Constructing an RNA world. *Trends in Genetics*, **14(12)** (Millennium issue), pp. M9–M13, 1999.
[63] Shapiro, R., *Origins–A Skeptic's Guide to the Creation of Life on Earth*, Bantam Doubleday Dell Pub., 1987.
[64] Lee, D.H., Granja, J.R., Martinez, J.A., Severin, K. & Ghadiri, M.R., A self-replicating peptide. *Nature*, **382**, pp. 525–528, 1996.
[65] Cairns-Smith, A.G., *Seven Clues to the Origin of Life: A Scientific Detective Story*, Cambridge University Press: Cambridge, MA, 1993.
[66] Cairns-Smith, G., The origin of life: clays. *Frontiers of Life*, **1**, pp. 169–192, Academic Press, 2001 (4-Volume Encyclopaedia, eds. D. Baltimore, R. Dulbecco, F. Jacob & R. Levi-Montalcini).
[67] Oparin, A.I., *Proiskhozhdenie Zhizni* [*The Origin of Life*], Izd. Moskovskii Rabochii: Moscow, 1924.
[68] Oparin, A.I., *Vozniknovenie Zhizni na Zemle* [*The Origin of Life on Earth*], 1st edn, Izv. Akad. Nauk SSSR, 1936. Translations published by Macmillan: NY, 1938; Dover Publications: New York, 1953; Academic Press: New York, 1957.
[69] Wilcox, R.M., *A Skeptical Scrutiny of the Works and Theories of Wilhelm Reich as Related to Bions*, http://home.netcom.com/∼rogermw/Reich/bions.html, 11 November 2002.
[70] Haldane, J.B.S., The Origin of Life. *Rationalist Annual*, **3**, pp. 148–153, 1929. Reprinted in *Science and Human Life*, Harper & Brothers: New York and London, 1933.
[71] Fox, S.W. & Dose, K., *Molecular Evolution and the Origin of Life*, Freeman: San Francisco, 1972.
[72] Fox, S., *The Emergence of Life: Darwinian Evolution from the Inside*, Basic Books: New York, 1988.
[73] Fox, S., My scientific discussion of evolution for the Pope and his scientists, Harbinger symposium *Religion and science: the best of enemies – the worst of friends*, 1997 (http://www.theharbinger.org/articles/rel_sci/fox.html).
[74] Enger, E.D. & Ross, F.C., *Concepts in Biology*, 8th edn., WCB Publishers: Dubuque, IA, 1997. Also McGraw-Hill (10th edn), 2002.
[75] Rohlfing, D.L., The development of the proteinoid model for the origin of life. *Molecular Evolution and Protobiology*, eds. Koichiro Matsuno, Klaus Dose, Kaoru Harada & Duane L. Rohlfing, pp. 29–43, Plenum Press: New York, 1984.
[76] Pappelis, A., *History of Biology*, Online lecture notes for course BIOL-315-2, Southern Illinois University at Carbondale, USA, 2002 (http://mccoy.lib.siu.edu/projects/bio315/index.htm).
[77] Egami, F., Chemical evolution in the primordial ocean and the role of transition element ions. *Izvestiya Nauk SSSR, Seriya Biologicheskaya*, **4**, pp. 519–526, 1980.
[78] Reich, W., *Selected Writings: An Introduction to Orgonomy*, Farrar, Strauss and Giroux, New York, 1973.

[79] Gardner, M., *Fads & Fallacies in the Name of Science*, Dover Publications Inc.: New York, 1957.
[80] Eigen, M., *Steps Towards Life*, Oxford University Press: Oxford, New York, 1972.
[81] Maynard Smith, J., Hypercycles and the origin of life. *Nature*, **20**, pp. 445–446, 1979.
[82] Lee, D.H., Severin, K. & Ghadiri, M.R., Autocatalytic networks: the transition from molecular self-replication to molecular ecosystems. *Curr. Opin. Chem. Biol.*, **1**, pp. 491–496, 1997.
[83] Lee, D.H., Severin, K., Yokobayashi, Y. & Ghadiri, M.R., Emergence of symbiosis in peptide self-replication through a hypercyclic network, *Nature*, **390**, pp. 591–594, 1997.
[84] Packard, N.H., Adaptation toward the edge of chaos. *Dynamic Patterns in Complex Systems*, eds. J.A.S. Kelso, A.J. Mandell & M.F. Shlesinger, World Scientific: Singapore: pp. 293–301, 1988.
[85] Adami, C., Belew, R., Kitano, H. & Taylor, C., (eds), *Artificial Life VI (Proceedings of the Sixth International Conference on Artificial Life)*, MIT Press: Cambridge, MA, 1998.
[86] Bak, P., *How Nature Works: The Science of Self-organized Criticality*, Oxford University Press: Oxford, New York, 1997.
[87] Freitas, R.A. Jr., A self-reproducing interstellar probe, *Journal of the British Interplanetary Society*, Vol. 33, pp. 251–264, July 1980.
[88] Waddington, C.H., *The Strategy of the Genes: A Discussion of Some Aspects of Theoretical Biology*, George Allen and Unwin: London, 1957.
[89] Prigogine, I. & Stengers, I., *La Nouvelle Alliance. Métamorphoses de la Science*, Gallimard: Paris, 1979.
[90] Adami, C. & Brown, C.T., Evolutionary learning in the 2D artificial life system "Avida". *Artificial Life IV*, eds. R. Brooks & P. Maes, pp. 377–381, MIT Press: Cambridge, MA, 1994.
[91] Ji, S., Biocybernetics: a machine theory of biology. *Molecular Theories of Cell Life and Death*, Chapter 1, ed. S. Ji, Rutgers University Press: New Brunswick-NJ, pp. 1–237, 1991.
[92] Maynard Smith, J., Evolutionary progress and levels of selection. *Evolutionary Progress*, ed. M.H. Nitecki, University of Chicago Press: Chicago, pp. 219–230, 1988.
[93] Belew, R.K. & Mitchell, M., (eds), *Adaptive Individuals in Evolving Populations: Models and Algorithms*, Addison-Wesley: Reading, MA, 1996.
[94] Wolfram, S., Cellular automata. *Los Alamos Science*, **9**, pp. 2–21, Fall 1983 (http://www.stephenwolfram.com/publications/articles/ca/)
[95] Reggia, J.A., Chou, H.-H. & Lohn, J.D., Cellular automata models of self-replicating systems. *Advances in Computers*, ed. M. Zelkowitz, **47**, pp. 141–183, Academic Press: New York, 1998.
[96] Sipper, M., An introduction to artificial life. *Explorations in Artificial Life* (special issue of *AI Expert*), pp. 4–8, Miller-Freeman: San Francisco, 1995.
[97] Wojtowicz, M., *What is Mirek's Cellebration (MCell)?* http://www.mirekw.com/ca/whatis_mcell.html, 2002.
[98] Berlekamp, E., Conway, J. & Guy, R., *Winning Ways*, Academic Press: New York, 1982.
[99] Poundstone, W., *The Recursive Universe*, Oxford University Press: Oxford, New York, 1987.
[100] Bontes, J., *Life32 by Johan Bontes*, http://psoup.math.wisc.edu/Life32.html, 2002.
[101] Rendell, P., *A Turing Machine in Conway's Game Life*, http://www.rendell.uk.co/gol/tm.htm, 2002.

[102] Dewdney, A.K., Computer recreations: in the game called Core War hostile programs engage in a battle of bits. *Scientific American*, **250(5)**, pp. 15–19, May 1984.
[103] Rasmussen, S., Knudsen, C., Feldberg, R. & Hindsholm, M., The coreworld: emergence and evolution of cooperative structures in a computational chemistry. *Physica D*, **42**, pp. 111–134, 1990.
[104] Thompson, G., *The Quine Page*, http://www.nyx.net/~gthompso/quine.htm, 1999.
[105] Green, D.G., *L-Systems*, Charles Sturt University preprint: 1993 (http://life.csu.edu.au/complex/tutorials/tutorial2.html).
[106] Lindenmayer, A., Mathematical models for cellular interaction in development (parts I and II). *Journal Theoretical Biol.*, **18**, pp. 280–315, 1968.
[107] Prusinkiewicz, P. & Lindenmayer, A., *The Algorithmic Beauty of Plants*, Springer: New York, 1990.
[108] Chomsky, N., Three models for the description of language. *IRE Trans. Inform. Theory*, **2(3)**, pp. 113–124, 1956.
[109] Sipper, M., Mange, D. & Stauffer, A., Ontogenetic hardware, *Biosystems*, **44(3)**, pp. 193–207, 1997.
[110] Morris, H.C., Typogenetics: a logic for artificial life. *Artificial Life*, ed. C.G. Langton, Addison-Wesley: Redwood City, CA, pp. 369–395, 1989.
[111] Varetto, L., Typogenetics: An artificial genetic system. *Journal of Theoretical Biology*, **160**, pp. 185–205, 1993.
[112] Snare, A., *Typogenetics*, BSc Thesis, School of Computer Science & Software Engineering, Monash University: Australia, 1999 (http://www.csse.monash.edu.au/hons/projects/1999/Andrew.Snare/thesis.pdf).
[113] Holland, J.H., Studies of the Spontaneous Emergence of Self-Replicating Systems Using Cellular Automata and Formal Grammars. *Automata, Languages, Development*, eds. A. Lindenmayer & G. Rozenberg, North-Holland: New York, pp. 385–404, 1976.
[114] Sipper, M., Tempesti, G., Mange, D. & Sanchez, E., (eds), von Neumann's legacy on self-replication. Special issue of *Artificial Life*, **4(3)**, Summer 1998.
[115] Sipper, M., Fifty years of research on self-replication: an overview. *Artificial Life*, **4(3)**, pp. 237–257, 1998.
[116] Turing, A., On computable numbers with an application to the *Entscheidungsproblem*. *Proc. London Math. Soc.*, **42**, pp. 230–265, 1936.
[117] von Neumann, J., *Re-evaluation of the Problems of Complicated Automata – Problems of Hierarchy and Evolution*, Fifth Illinois Lecture, December 1949 – published in *Papers of John von Neumann on Computing and Computer Theory*, pp. 477–490, eds. W. Aspray & A. Burks, MIT Press, 1987.
[118] Burks, A., von Neumann's self-reproducing automata. *Essays on Cellular Automata*, ed. A. Burks, University of Illinois Press: Urbana, IL, pp. 3–64, 1970.
[119] von Neumann, J., The general and logical theory of automata. *Cerebral Mechanisms in Behavior*, ed. L.A. Jeffress, John Wiley & Sons: New York, pp. 1–31, 1951 (Proc. of the *Hixon Symposium*, Pasadena, CA, September 1948).
[120] Rosen, R., On a logical paradox implicit in the notion of a self-reproducing automaton. *Bull. Math. Biophys.*, **21**, pp. 387–394, 1959.
[121] Boozer, D., *Self Replication*, http://www.its.caltech.edu/~boozer/symbols/self_replication.html, 2001.
[122] Sipper, M., *The Artificial Self-replication Page*. http://www.cs.bgu.ac.il/~sipper/selfrep/, 2002.

[123] Pesavento, U., An implementation of von Neumann's self-reproducing machine. *Artificial Life*, **2(4)**, pp. 337–354, 1995.
[124] Beuchat, J. & Haenni, J., Von Neumann's 29-state cellular automaton: a hardware implementation. *IEEE Transactions on Education*, **43(3)**, pp. 300–308, 2000.
[125] McMullin, B., Artificial Darwinism: the very idea!. *Autopoiesis & Perception*, Report BM CM 9401, eds. B. McMullin & N. Murphy, Dublin City University: Dublin, pp. 71–94, 1994 (proceedings of a workshop held in Dublin City University, August 25–26, 1992).
[126] McMullin, B., *Computational Autopoiesis: The Original Algorithm*, working paper 97-01-001, Santa Fe Institute: Santa Fe, NM 87501, USA, 1997.
[127] Maynard Smith, J., *The Problems of Biology*, Oxford University Press: Oxford, 1986.
[128] Laing, R.A., Some alternative reproductive strategies in artificial molecular machines. *Journal of Theoretical Biology*, **54**, pp. 63–84, 1975.
[129] Laing, R.A., Automaton models of reproduction by self-inspection. *Journal Theoretical Biology*, **66**, pp. 437–456, 1977.
[130] Ibáñez, J., Anabitarte, D., Azpeitia, I., Barrera, O., Barrutieta, A., Blanco, H. & Echarte, F., Self-inspection based reproduction in cellular automata. *Advances in Artificial Life: Third European Conference on Artificial Life*, eds. F. Morán, A. Moreno, J.J. Merelo, & P. Chacón, Lecture Notes in Artificial Intelligence, Springer, Berlin, pp. 564–576, 1995.
[131] Codd, E.F., *Cellular Automata*, Academic Press: New York, 1968.
[132] Devore, J. & Hightower, R., The Devore variation of the Codd self-replicating computer, draft presented at the *Third Workshop on Artificial Life*, Santa Fe, NM, November 1992.
[133] Smith, A.R. III, Cellular automata and formal languages. *Proc. 11th Annual Symposium on Switching and Automata Theory*, Santa Monica, 28–30 October 1970 – IEEE, New York, pp. 216–224, 1970.
[134] Banks, E.R., Universality in cellular automata. *Proc. 11th Annual Symposium on Switching and Automata Theory*, Santa Monica, 28–30 October 1970 – IEEE: New York, pp. 194–215, 1970.
[135] Vitànyi, P.M.B., Sexually reproducing cellular automata. *Mathematical Biosciences*, **18**, pp. 23–54, 1973.
[136] Byl, J., Self-reproduction in small cellular automata. *Physica D*, **34**, pp. 295–299, 1989.
[137] Reggia, J.A., Armentrout, S.L., Chou, H.-H. & Peng, Y., Simple systems that exhibit self-directed replication. *Science*, **259**, pp. 1282–1287, 1993.
[138] Sipper, M. & Reggia, J.A., Go forth and replicate. *Scientific American*, **285(2)**, pp. 26–35, 2001.
[139] Sipper, M., Sanchez, E., Mange, D., Tomassini, M., Perez-Uribe, A. & Stauffer, A., A phylogenetic, ontogenetic, and epigenetic view of bio-inspired hardware systems. *IEEE Trans. Evol. Comput.*, **1(1)**, pp. 83–97, 1997 (http://www.cs.bgu.ac.il/%7Esipper/poe.html).
[140] Tempesti, G., A new self-reproducing cellular automaton capable of construction and computation. *Lecture Notes in Artificial Intelligence*, **929**, pp. 555–563, Springer-Verlag: Berlin, 1995.
[141] Perrier, J.-Y., Sipper, M. & Zahnd, J., Toward a viable, self-reproducing universal computer. *Physica D*, **97**, pp. 335–352, 1996.
[142] Chou, H. & Reggia, J., Problem solving during artificial selection of self-replicating loops. *Physica D*, **115**, pp. 293–312, 1998.
[143] Arbib, M.A., Simple self-reproducing universal automata. *Information and Control*, **9**, pp. 177–189, 1966.
[144] Chou, H. & Reggia, J., Emergence of self-replicating structures in a cellular automata space. *Physica D*, **110**, pp. 252–276, 1997.

[145] Toffoli, T. & Margolus, N. *Cellular Automata Machines*, MIT Press: Cambridge, MA, 1987.
[146] Lohn, J.D. & Reggia, J.A., Exploring the design space of artificial self-replicating structures. *Evolution of Engineering and Information Systems and Their Applications*, ed. L.C. Jain, CRC Press: Boca Raton, FL, pp. 67–103, 2000.
[147] Pargellis, A.N., The spontaneous generation of digital "life". *Physica D*, **91**, pp. 86–96, 1996.
[148] Pargellis, A.N., The evolution of self-replicating computer organisms. *Physica D*, **98**, pp. 111–127, 1996.
[149] Sayama, H., Self-replicating worms that increase structural complexity through gene transmission. *Artificial Life VII*, eds. M.A. Bedau, J.S. McCaskill, N.H. Packard & S. Rasmussen, pp. 21–30, MIT Press: Cambridge, MA, 2000.
[150] Langton, C.G., *Artificial Life: An Overview*, MIT Press: Cambridge, MA, 1995.
[151] Adami, C., Ofria, C. & Collier, T.C., Evolution of biological complexity. *Proc. Nat. Acad. Sci. USA*, **97**, pp. 4463–4468, 2000.
[152] Gould, S.J., The meaning of punctuated equilibrium, and its role in validating a hierarchical approach to macroevolution. *Perspectives on Evolution*, ed. R. Milkman, pp. 83–104, Sunderland, MA., 1982.
[153] Koza, J.R., *Genetic Programming: On the Programming of Computers by Means of Natural Selection*, MIT Press: Cambridge, MA, 1992.
[154] Koza, J.R., Artificial life: spontaneous emergence of self-replicating and evolutionary self-improving computer programs. *Artificial Life III*, ed. C.G. Langton, (Vol. XVII of *SFI Studies in the Sciences of Complexity*), Addison-Wesley: Reading, MA, pp. 225–262, 1994.
[155] Reynolds, C.W., Flocks, herds, and schools: a distributed behavioral model. *Computer Graphics*, **21(4)**, pp. 25–34, 1987.
[156] Brooks, R.A., New approaches to robotics. *Science*, **253(5025)**, pp. 1227–1232, 1991.
[157] Minsky, M., *The Society of Mind*, Touchstone–Simon & Schuster: New York, 1988.
[158] Ray, T., *How I Created Life in a Virtual Universe*, http://www.isd.ats.co.jp/∼ray/pubs/nathist/index.html, 2002.
[159] Thearling, K. & Ray, T.S., Evolving multi-cellular artificial life. *Artificial Life IV*, eds. R. Brooks & P. Maes, MIT Press: Cambridge, MA, 1994.
[160] Ray, T.S., An approach to the synthesis of life. *Artificial Life II*, Vol. X of *SFI Studies in the Sciences of Complexity*, eds. C.G. Langton, C. Taylor, J.D. Farmer & S. Rasmussen, pp. 371–408, Addison-Wesley: Redwood City, CA, 1992.
[161] Adami, C., Self-organized criticality in living systems. *Phys. Lett. A*, 203, pp. 29–32, 1995.
[162] Skipper, J., The computer zoo – evolution in a box. *Toward a Practice of Autonomous Systems: Proceedings of the First European Conference on Artificial Life*, eds. F.J. Varela & P. Bourgine, pp. 355–364, MIT Press: Cambridge, MA, 1992.
[163] McMullin, B., *SCL: An Artificial Chemistry in SWARM*, working paper 97-01-002, Santa Fe Institute: Santa Fe, NM 87501, USA, 1997.
[164] McMullin, B. & Varela, F.J., Rediscovering computational autopoiesis. *Proceedings of the Fourth European Conference on Artificial Life*, eds. P. Husbands & I. Harvey, MIT Press: Cambridge, MA, 1997.
[165] McMullin, B., The Holland α-universes revisited. *Toward a Practice of Autonomous Systems: Proceedings of the First European Conference on Artificial Life, Series: Complex*

Adaptive Systems, eds. F.J. Varela & P. Bourgine, MIT Press: Cambridge, MA, pp. 317–326, 1992.

[166] Rucker, R., *Boppers Download*, http://www.mathcs.sjsu.edu/faculty/rucker/boppers.htm, 1996–1999.

[167] Rucker, R., *Artificial Life Lab*, Waite Group Press: Corte Madera, CA, 1993 (partly online at http://www.mathcs.sjsu.edu/faculty/rucker/bopbook.htm).

[168] Bedau, M.A., McCaskill, J.S., Packard, N.H., Rasmussen, S., Adami, C., Green, D.G., Ikegami, T., Kaneko, K. & Ray, T.S., Open problems in Artificial Life. *Artificial Life*, **6**, pp. 363–376, 2000.

[169] Tomalia, D.A., Wang, Z.-G. & Tirrell, M., Experimental self-assembly: the many facets of self-assembly. *Current Opinion in Colloid & Interface Science*, **4**, pp. 3–5, 1999.

[170] Gillyatt, P., & Yarrow, J., (eds), *Under the Hood of a Cellular Transport Machine – Collaboration Sheds Light on Assembly of Transporters Associated with Cholesterol, Breast Cancer, and HIV*, News Release, Harvard Medical School, 1999 (http://www.hms.harvard.edu/news/releases/699clathrin.html; see also video by A. Bruce at http://www.hms.harvard.edu/news/clathrin/index.html).

[171] Ciofalo, M., Collins, M.W. & Hennessy, T.R., *Nanoscale Fluid Dynamics in Physiological Processes: A Review Study*, Wessex Institute of Technology Press – Computational Mechanics Publications, Series *Advances in Computational Biomedicine*, 1999.

[172] Rebek, J. Jr., Synthetic self-replicating molecules. *Scientific American*, **271(1)**, pp. 48–55, 1994.

[173] Solem, J.C., Self-assembling micrites based on the Platonic solids. *Robotics and Autonomous Systems*, **38**, pp. 69–92, 2002.

[174] Drexler, K.E., Forrest, D., Freitas, R.A., Jr., Storrs Hall, J., Jacobstein, N., McKendree, T., Merkle, R. & Peterson, C., *On Physics, Fundamentals, and Nanorobots: A Rebuttal to Smalley's Assertion that Self-replicating Mechanical Nanorobots are Simply not Possible*. Institute for Molecular Manufacturing, 2001 (http://www.imm.org/SciAmDebate2/smalley.html).

[175] Penrose, L.S., Mechanics of self-reproduction. *Ann. Human Genetics*, **23**, pp. 59–72, 1958.

[176] Penrose, L.S., Self-reproducing machines. *Scientific American*, **200(6)**, pp. 105–114, 1959.

[177] Penrose, L.S. & Penrose, R., A self-reproducing analogue. *Nature*, **179(1183)**, 1957.

[178] Jacobson, H., The informational content of mechanisms and circuits. *Information and Control*, **2(3)**, pp. 285–296, 1959.

[179] Morowitz, H.J., A model of reproduction. *American Scientist*, **47**, pp. 261–263, 1959.

[180] Lohn, J.D., The evolution of simplicity: self-replicating systems in artificial life. Invited talk at *International Workshop on Mathematical & Computational Biology: Computational Morphogenesis, Hierarchical Complexity & Digital Evolution*, Aizu, Japan, 1997.

[181] Suthakorn, J., Zhou, Y. & Chirikjian, G., Self-replicating robots for space utilization. *Proceedings of the 2002 Robophere Workshop on Self Sustaining Robotic Ecologies*, NASA Ames Research Center, California, 2002 (http://custer.me.jhu.edu/jackritweb/RoboSphere2002.pdf).

[182] Storrs Hall, J., Utility fog: the stuff that dreams are made of *Nanotechnology – Molecular Speculations on Global Abundance*, ed. B.C. Crandall, pp. 161–184, MIT Press: Cambridge, MA, 1996 (http://discuss.foresight.org/~josh/Ufog.html).

[183] Husbands, P. & Meyer, J.A., *Evolutionary Robotics*, Springer Verlag: Berlin, 1998.

[184] Lipson, H. & Pollack, J.B., Automatic design and manufacture of robotic lifeforms. *Nature*, **406**, pp. 974–978, 2000 (http://helen.cs-i.brandeis.edu/golem/download/naturegolem.pdf).

[185] Skidmore, G.D., Parker, E., Ellis, M., Sarkar, N. & Merkle, R., *Exponential Assembly*, online preprint, 2000 (http://www.zyvex.com/Research/Publications/ papers/exponential GS.html).

[186] Sipper, M., *Evolution of Parallel Cellular Machines: the Cellular Programming Approach*, Springer-Verlag: Heidelberg, 1997.

[187] Sanchez, E., Field-programmable gate array (FPGA) circuits. *Towards Evolvable Hardware*, eds. E. Sanchez & M. Tomassini, (Vol. 1062 of *Lecture Notes in Computer Science*), Springer-Verlag: Heidelberg, pp. 1–18, 1996.

[188] Mange, D. & Stauffer, A., Introduction to embryonics: towards new self-repairing and self-reproducing hardware based on biological-like properties: *Artificial Life and Virtual Reality*, eds. N.M. Thalmann & D. Thalmann, Wiley: Chichester, UK, pp. 61–72, 1994.

[189] Mange, D., Sipper, M., Stauffer, A. & Tempesti, G., Towards robust integrated circuits: the embryonics approach. *Proc. IEEE*, **4**, pp. 516–541, 2000.

[190] von Thiesenhausen, G. & Darbro, W.A., *Self-replicating Systems – A Systems Engineering Approach*, NASA TM-78304, July 1980.

[191] Barrow, J.D. & Tipler, F.J., *The Anthropic Cosmological Principle*, Oxford University Press: Oxford, New York, 1986.

[192] Crichton, M., *Prey: A Novel*, HarperCollins: New York, 2003.

[193] Feynman, R., There's plenty of room at the bottom. *Engineering and Science*, **23**, pp. 22–36, 1960. Reprinted in *Nanotechnology: Research and Perspectives*, Crandall, B. & Lewis, J.B. (eds.), MIT Press, Cambridge, MA, 1992.

[194] Franks, A., Nanotechnology. *J. Phys. E: Sci. Instr.*, **20**, pp. 1442–1451, 1987.

[195] Stix, G., Micron Machination. *Sci. Amer.*, **267(5)**, pp. 106–113, 1992.

[196] Dewdney, A.K., Nanotechnology: wherein molecular computers control tiny circulatory submarines. *Scientific American*, **258(1)**, pp. 88–91, January 1988.

[197] Storrs Hall, J., Architectural considerations for self-replicating manufacturing systems. *Nanotechnology*, **10(3)**, pp. 323–330, 1999 (http://discuss.foresight.org/~josh/selfrepsys/index.html).

[198] Lewis, J.B., Smith, B. & Krummenacker, N., *A Proposed Path from Current Biotechnology to a Replicating Assembler*, Molecubotics online preprint, November 1996 (http://www.molecubotics.com/tech-docs/dgap.html).

[199] Smalley, R.E., Of chemistry, love, and nanobots. *Scientific American*, **285**, pp. 76–77, 2001.

[200] Drexler, K.E., *Engines of Creation: Challenges and Choices of the Last Technological Revolution*, Anchor Press-Doubleday: Garden City, New York, 1986.

[201] Merkle, R., Self replicating systems and molecular manufacturing. *Journal of the British Interplanetary Society*, **45**, pp. 407–413, 1992 (http://www.zyvex.com/nanotech/selfRepJBIS.html).

[202] Merkle, R., Self replicating systems and low cost manufacturing. *The Ultimate Limits of Fabrication and Measurement*, eds. M.E. Welland & J.K. Gimzewski, Kluwer: Dordrecht, pp. 25–32, 1994 (http://www.zyvex.com/nanotech/selfRepNATO.html).

[203] Merkle, R., Design considerations for an assembler. *Nanotechnology*, **7**, pp. 210–215, 1996 (http://www.zyvex.com/nanotech/nano4/merklePaper.html#selfrep).

[204] Denton, M., *Evolution: A Theory in Crisis*, Adler & Adler Pub, 1986.

[205] Laing, R.A., Machines as organisms: an exploration of the relevance of recent results. *Biosystems*, **11**, pp. 201–218, 1979.

[206] Lohn, J.D., *Automatic Discovery of Self-replicating Structures in Cellular Space Automata Models*, Technical Report CS-TR-3677, University of Maryland, College Park, 1996.

[207] Darnell, J., Lodish, H. & Baltimore, D., *Molecular Cell Biology*, Scientific American Books: New York, 1986.

[208] Severin, K., Lee, D.H., Martinez, J.A., Vieth, M. & Ghadiri, M.R., Dynamic error correction in autocatalytic peptide networks. *Angewandte Chemie* (International edn), **37**, pp. 126–128, 1998.

[209] Yao, S., Ghosh, I., Zutshi, R. & Chmielewski, J., Selective amplification by auto- and cross-catalysis in a replicating peptide system. *Nature*, **396**, pp. 447–450, 1998.

[210] New, M.H. & Pohorille, A., An inherited efficiencies model of non-genomic evolution. *Simulation Practice and Theory*, **8**, pp. 99–108, 2000.

[211] Shannon, C.E., A mathematical theory of communication. *The Bell System Technical Journal*, **27**, pp. 379–423, 623–656, 1948 (reprinted *Key Papers in the Development of Information Theory*, ed. D. Slepian, IEEE Press: New York, 1974).

[212] Jacobson, H., The informational capacity of the human ear. *Science*, **112**, pp. 143–144, 1950.

[213] Jacobson, H., The informational capacity of the human eye. *Science* , **113**, pp. 292–293, 1951.

[214] Jacobson, H., On models of reproduction. *American Scientist*, **46**, pp. 255–284, 1958.

[215] Pattee, H.H., Evolving self-reference: matter, symbols, and semantic closure. *Communication and Cognition – Artificial Intelligence*, **12(1-2),** pp. 9–27, 1995 (Special issue *Self-reference in Biological and Cognitive Systems*, ed. L. Rocha).

[216] von Neumann, J., *The Mathematical Foundations of Quantum Mechanics*, Princeton University Press: Princeton, NJ, 1955.

[217] Dennett, D.C., *Darwin's Dangerous Idea: Evolution and the Meanings of Life*, Simon & Schuster: New York, 1995.

[218] Pattee, H.H., Artificial life needs a real epistemology. *Advances in Artificial Life* (*Proc. Third European Conference on Artificial Life*, Granada, Spain, June 4–6, 1995), eds. F. Moran, A. Moreno, J.J. Merelo & P. Chacon, Springer-Verlag: Berlin, pp. 23–38, 1995.

[219] Pattee, H.H., Simulations, realizations, and theories of life. *Artificial Life* (vol. VI of Santa Fe Institute Studies in the Sciences of Complexity), ed. C. Langton, Addison-Wesley: Reading, MA, pp. 63–77, 1988. Reprinted in *The Philosophy of Artificial Life*, ed. M.A. Boden, Oxford University Press: Oxford, 1996.

[220] Gell-Mann, M., *The Quark and the Jaguar: Adventures in the Simple and the Complex*, W.H. Freeman & Co.: New York, 1994.

[221] Drexler, K.E., *Nanosystems: Molecular Machinery, Manufacture, and Computing*, John Wiley & Sons: New York, 1992.

[222] Adami, C., *Introduction to Artificial Life*, Springer-Verlag: New York, 1998.

[223] Etxeberria, A. & Moreno, A., From complexity to simplicity: nature and symbols. *Bio-systems*, **60**, pp. 149–157, 2001 (special issue on *Physics and Evolution of Symbols and Codes*) (www.c3.lanl.gov/∼rocha/pattee/etxeberriamoreno.pdf).

[224] Lohn, J.D., Haith, G.L. & Colombano, S.P., Two electromechanical self-assembling systems. Presented at the *Sixth Foresight Conference on Molecular Nanotechnology*, Santa Clara, CA, Nov. 13–15, 1998, poster session (http://ic.arc.nasa.gov/ic/people/jlohn/Papers/nano1998.pdf).

[225] Myhill, J., The abstract theory of self-reproduction. *Essays on Cellular Automata*, ed. A.W. Burks, University of Illinois Press: Urbana, IL, pp. 206–218, 1970.
[226] Orgel, L.E., Molecular replication. *Nature*, **358**, pp. 203–209, 1992.
[227] Rocha, L.M., Selected self-organization and the semiotics of evolutionary systems. *Evolutionary Systems: The Biological and Epistemological Perspectives on Selection and Self-Organization*, eds. S. Salthe, G. Van de Vijver & M. Delpos, Kluwer Academic Publishers: Amsterdam, pp. 341–358, 1998 (http://www.c3.lanl.gov/~rocha/ises.html).
[228] Stauffer, A. & Sipper, M., On the relationship between cellular automata and L-systems: The self-replication case. *Physica D*, **116**, pp. 71–80, 1998.

Chapter 4

The Human Genome Project

P. Gross & T. Oelgeschläger
Eukaryotic Gene Regulation Laboratory, Marie Curie Research Institute, UK.

Abstract

The Human Genome Project, the most ambitious biological research project to date, was inaugurated in the 1980s with the aim to decipher the precise DNA sequence of the entire human genetic material and culminated in the publication of two human genome sequence drafts in February 2001. These drafts represent a major milestone in biological research as they provide the first panoramic view of the genomic landscape of a vertebrate. Among the results of the sequencing efforts was the surprising finding that only 1.4% of the human DNA encodes instructions for the assembly of proteins. The estimated total number of just over 30,000 human protein-coding genes is insufficient to explain human complexity compared to other organisms simply on the basis of gene number. Additional levels of complexity must exist, both in the coordinated temporal and spatial read-out of genetic information and in the functional interplay of expressed gene products. This insight has ushered in functional genomics, the genome-wide analysis of cell-specific gene expression patterns and protein interaction networks. The human genome sequence project has led to the emergence of novel technologies that are expected to promote research into virtually all aspects of life science. Medical sciences are expected to benefit in particular, as the availability of human genome sequence information has paved the way for the development of novel diagnostics and advanced therapeutics. These may ultimately provide the means for the prevention and targeted treatment of complex genetic disorders such as cancer.

1 Introduction

1.1 Genes

The elucidation of the precise mechanisms underlying heredity, the accurate transfer of biological information from one generation to the next, has been a key problem in biological sciences over several centuries. Gregor Mendel, an Austrian monk, concluded in 1865 from genetic crossing experiments that hereditary information is passed from parents to offspring in discrete packets [1]. Later on, these units of heredity were called genes. The biological nature of genes was not revealed until 1944, when it was demonstrated that genetic information is carried by a biological substance

with the chemical designation deoxyribonucleic acid (DNA) [2]. DNA is a polymer composed of deoxyribonucleotides, each of which typically contains one of four different bases: adenine (A), guanine (G), thymine (T), or cytosine (C). The discovery of the double-helical structure of DNA by James Watson and Francis Crick in 1953, certainly one of the most significant scientific discoveries of the 20th century, provided the first clues on how DNA can be accurately duplicated in order to transmit genomic information from one generation to the next. Within the DNA double helix structure, two DNA strands closely interact via hydrogen bonds between their nucleotide bases to form specific inter-strand base pairs: adenine pairs with thymine and guanine pairs with cytosine [3]. Hence, the two strands of the DNA double helix are complementary to each other and can both serve as templates during replication, the process by which cells copy their genetic information before cell division. Complete replication of a DNA molecule results in two identical DNA duplexes, each consisting of one parental and one newly synthesised DNA strand [4–6].

Genetic information is encoded within the nucleotide sequence of the DNA strands. The read-out of genetic information occurs through a process called transcription, the synthesis of RNA, a single-stranded nucleic acid consisting of ribonucleotides, from a DNA template. Transcription is catalysed by RNA polymerases and results in a RNA molecule with the identical nucleotide sequence as the coding DNA strand of a gene. This RNA transcript can either have biological activity by itself or, as in the case of protein-coding genes, can be used as a template to synthesise a specific polypeptide chain from different amino acid components.

RNA molecules that serve as templates for protein synthesis are called messenger RNAs (mRNAs). The genetic code that is used to translate mRNAs into a protein polypeptide chain with specific amino acid sequence had been mainly worked out by 1964 [5–7]. A triplet of nucleotide bases, a codon, encodes one specific amino acid and the sequence of codons in mRNA is co-linear to the amino acid sequence of the resulting protein. There are $4^3 = 64$ possible codons, 61 of which encode the 20 naturally occurring amino acids in humans. The remaining three codons are used as signals that cause the termination of protein synthesis [4–6].

In 1977, it was discovered that genes could be much longer than would be predicted from the amino acid sequence of their protein products [8, 9]. Additional DNA sequences were found that that were not present in the corresponding mRNA used for protein synthesis, and these sequences were interspersed with coding sequences. Coding regions of a gene present in mRNA were subsequently called exons, and the non-coding regions were called introns. Transcription of a protein-coding gene consisting of exons and introns gives rise to a primary RNA molecule (pre-mRNA) from which the intron regions have to be removed to yield mature mRNA that then can be used for protein synthesis. This process of RNA splicing involves the deletion of intron sequences followed by the precise joining of exons so that the co-linearity of gene and protein is maintained between the individual exons and the corresponding parts of the protein chain. In the majority of human protein-encoding genes, exons are usually shorter than introns and the number of introns per gene can be higher than 10. Depending on their exon/intron structure protein-coding genes can vary considerably in length, ranging from 1 kb to several millions of base pairs (Mb) [5, 6, 10].

Gene expression (i.e. the transcription of a gene and its translation into a specific protein molecule) has to be strictly regulated to ensure the proper functioning of a cell. In multicellular organisms (metazoans), only a subset of the total genetic material is actively expressed at a given time in a given cell type. Perturbations in the coordinated expression of genes in only a single cell can have disastrous consequences for the whole organism, it can, for example, result in the development of cancer. Gene expression is primarily regulated at the level of transcription. The information for the precise control of transcription is encoded in specific DNA sequence elements called gene promoters. Core promoter elements serve to assemble the transcription machinery

and define the start site of transcription. The activity of core promoter elements is controlled by regulatory DNA sequences that are typically recognized by DNA-binding gene- and cell type-specific transcription activators or repressors. Regulatory promoter regions vary considerably in length and typically contain binding sites for several regulatory proteins. They can be found either close to the start site of transcription of a given gene (basal promoter elements) or at a distance of several thousand base pairs (kb) (enhancers) [5, 6].

1.2 Genome organisation

A genome is the entirety of all DNA within an organism. In addition to genomic DNA present in the nucleus, there is also DNA present in the mitochondria, semi-autonomous organelles found in most eukaryotic cells that serve as energy factories. The mitochondrial genome contributes to less than 1% of the total cellular DNA in mammalians and will not be discussed in this article.

In the nuclei of eukaryotic cells, DNA is organised in a set of chromosomes. Human cells contain 22 pairs of autosomes; one chromosome of each pair is inherited from the mother and the other from the father. Since cells contain two copies of each autosome, they also contain two copies of each gene, the so-called alleles. In addition to the autosomes, there are the sex chromosomes X and Y. Females possess two X chromosomes, one from each parent, whereas males possess an X chromosome inherited from their mother and a Y chromosome inherited from their father.

The DNA within the 23 human chromosomes contains a total of 3.2×10^9 nucleotide bases (3.2 Gb) and, if completely stretched, would be approximately 2 m long. To fit the genomic DNA into the cell nucleus, which is only a few micrometres in diameter, an enormous level of compaction has to be achieved. Chromosomes represent the most compact form of nuclear DNA and are only observed during cell division. In non-dividing cells, so-called interphase cells, the DNA material occupies the cell nucleus as chromatin without distinguishable chromosomes. In chromatin, DNA is organised into nucleosomes; $\sim$200 bp of DNA are wrapped around a protein disc containing a histone protein octamer [11]. Nucleosomes are arranged like beads on a string to form a 10 nm chromatin fibre that can be further compacted to a 30 nm fibre by incorporating a specific linker histone protein, H1 [5, 6]. As a 30 nm fibre, a human chromosome would still span the nucleus more than a 100 times. The mechanism(s) underlying further condensation of the 30 nm fibre into higher order structures such as chromosomes are not yet fully understood.

Two major types of chromatin can be distinguished based on the level of DNA condensation. Euchromatin occupies most of the nucleus and the underlying DNA fibres are much less densely packed as compared to heterochromatin, which exhibits a level of DNA compaction comparable to chromosomes. Of the 3.2 Gb of the human genome, 2.95 Gb or 92% are euchromatic and only 0.35 Gb or 8% are heterochromatic [5].

1.3 Genome contents

Much progress has been made in understanding the molecular structure of DNA and its organisation into chromatin. However, the function of the plethora of genomic DNA sequences remains poorly understood. Evidently, the pivotal function of DNA sequences in genes is to provide a blueprint for biologically active RNAs and proteins required for cell growth, differentiation, and development of multicellular organisms. However, the majority of the human genome (53%) consists of repeated DNA sequences of various types that do not confer essential cellular functions. These are sometimes referred to as 'junk' sequences although this implicative negative role is not justifiable a priori.

The largest proportion of repeated sequences, about 45% of the human genome, are parasitic DNA sequences, which can provide valuable clues about evolutionary events and forces [10, 12]. These DNA elements are discussed in detail in Chapter 3.1. Segmental duplications of 10 to 300 kb that have been copied from one region of the genome to another contribute to about 5% of the human genome [10, 12]. Blocks of tandemly repeated sequences are found in centromeres, constricted regions of the chromosome that include the site of attachment of the mitotic spindle, and in telomeres, the DNA regions at the chromosome ends. A different class of repeated DNA elements present in the human genome contains only small stretches of repeated DNA. Depending on the size of the repeat unit these sequence elements are either called microsatellites (1–13 bp) or minisatellites (14–500 bp) [5]. Satellite sequences have been extremely useful in human genetic studies, as there is considerable variation between individuals. The precise mapping of satellite positions on individual chromosomes has provided a comprehensive catalogue of gene markers [13].

Only a very small proportion of the human genome (~1.4%) contains protein-coding genes [10, 12]. However, the simple rule 'one gene one protein' does not always apply and the total number of distinct cellular proteins is significantly greater than the total number of protein-encoding genes. One mechanism by which a single gene can give rise to several distinct protein products is alternative splicing, which involves the differential assignment or usage of intros and exons within pre-mRNAs [5, 6]. Alternative splicing has been found to be more pronounced in humans than in any other species and it has been estimated that two to three alternative splicing products exist for each gene [14, 15]. Another process by which the informational content of RNA can be altered is RNA editing, which involves the introduction of changes at individual nucleotide positions or the addition of nucleotide bases within a mRNA [6]. Furthermore, protein synthesis can start or terminate at different positions within the mature mRNA, giving rise to protein products that are identical in amino acid sequence but different in length and therefore potentially different in function. In addition, proteins are subject to post-translational modifications that add to the complexity of their role and function within the cell. These include chemical modifications (e.g. phosphorylation, acetylation, methylation) or conjugation to other proteins such as ubiquitin or ubiquitin-related proteins [5, 6, 16, 17].

Finally, the human genome contains a class of genes that, although primary transcripts can be found in some cases, do not give rise to products with cellular function. These pseudogenes contain DNA sequences typical for functional genes (i.e. promoter and coding regions) but are rendered inactive by mutations in the DNA sequence that affect transcription, splicing, or translation. Most pseudogenes contain deletions of one or several DNA segments. Since there is no further selection against the accumulation of additional mutations once gene expression is abolished, the time of inactivation can be estimated by comparing the DNA sequences of pseudogenes with those of the original genes. Pseudogenes have been found to be several ten million years old [5].

2 The Human Genome Project

2.1 History of the Human Genome Project

The Human Genome Project was first proposed in the 1980s with the aim to decipher the entire human genetic material and to identify the complete set of human genes [18]. It was evident from the start that a project of this scale would require a communal effort in infrastructure building unlike any other previously attempted biomedical enterprise. The first official programme to sequence the human genome was announced in April 1990 as a joint effort of the Department of Energy and the National Institutes of Health in the US. This programme featured a broad approach that included

the construction of human genetic maps, which identify the relative position of particular genes on a chromosome and which provide starting points for the assembly of genomic DNA sequences. In parallel, efforts were directed to sequence key model organisms such as bacteria, worm, fly, and mouse. Furthermore, research into the ethical, legal and social issues raised by human genome research was intended.

In October 1990, the Human Genome Project was officially launched and genomic centres in the participating countries US, UK, France, Germany, Japan, and China, were created. In addition, the Human Genome Organisation was founded to provide a forum for the coordination of international genomic research. The main sequencing centre in the US was established at the Whitehead Institute in Cambridge, Massachusetts. In the UK, the Wellcome Trust and the MRC opened the Sanger Centre close to Cambridge where later one third of the entire sequencing effort would be taken on. Rapid progress was made. The first human genetic maps were developed and refined [19, 20] and genome sequencing of the first free-living organism, the bacterium *Haemophilus influenzae*, was completed in 1995 [21]. In the following years the genome sequences of two key model organisms in biological research, the yeast *Saccharomyces cerevisiae* [22] and the worm *Caenorhabditis elegans* [23], were released.

In 1998 the international collaboration fell apart over disputes regarding the strategic approach for sequencing the human genome. Craig Venter at the Institute for Genomic Research in Rockville, US, argued that the progress on the human genome could be accelerated considerably by using whole-genome shotgun sequencing. He and his colleagues had successfully used whole-genome shotgun sequencing to determine the *H. influenzae* genome sequence in record time [21]. However, Francis Collins, head of the National Human Genomic Research Institute, insisted on a more methodical and conservative sequencing strategy which he considered to give the highest possible quality of sequence data. Craig Venter finally left the publicly funded consortium to set up a biotechnology company, Celera Genomics. He announced the complete sequencing of the human genome by 2003, 2 years earlier than the completion date projected by the Human Genome Project. While Celera Genomics promised free access to raw DNA sequence data, they proposed to perform analyses of their genome sequence database on a commercial basis. This announcement created an uproar amongst the scientific community who argued that human genome sequence information was fundamental and that it should be freely accessible. Additional concerns regarded the future of the public programme after the long and difficult groundwork that had been done [18]. In response, the leaders of the public consortium announced new goals for the public project in order to beat Celera Genomics to the finish line. The pace of sequencing was to be increased to produce a first 'rough' draft covering $\sim$90% of the human genome by spring 2001. This was the starting point for a race between the two groups that was punctuated by duelling press releases of respective milestones over the following two years [24]. Finally, the international consortium (Human Genome Project) and Celera Genomics jointly announced working drafts of the human genome sequence in a ceremony at the White House in Washington on 26 June 2000. A simultaneous publication of the results, however, collapsed over controversies regarding the amount of data Celera Genomics was willing to publicise and two separate reports on the human genome sequence were released. The publicly funded Human Genome Project published in the British journal *Nature* on 15 February 2001 [10], and Celera Genomics published in the American journal *Science* on 16 February 2001 [25].

2.2 Strategy of the Human Genome Project

The basis of any effort to decipher genomic information is the ability to determine the exact nucleotide sequence of any given DNA segment. Two research groups led by Frederick Sanger

at MRC Laboratory of Molecular Biology in Cambridge, UK [26], and by Walter Gilbert at Harvard University in Boston, US [27], independently published different strategies to sequence DNA segments with high accuracy in 1977. Sanger and Gilbert were awarded the Nobel Price for this groundbreaking work in 1980. Until the mid-80s DNA sequencing technologies were not significantly advanced and even state-of-the-art laboratories could only sequence around 500 nucleotide bases in one experiment. From a technical point of view, sequencing an entire genome seemed a virtually impossible task. This limitation was overcome with the development of automated sequencing machines, first introduced in 1986, that made daily outputs of over one million bases (1 Mb) possible [28].

Very large DNA molecules such as chromosomes cannot be sequenced directly. The DNA material needs to be fragmented into smaller and more manageable pieces that, after sequence determination, have to be re-assembled in the correct order and orientation *in silico*. This approach, called shotgun sequencing, was first introduced in 1981 and later more refined and extended to increase efficiency and accuracy [29]. Two different shotgun sequencing strategies were employed to decipher the human genome: the public consortium (Human Genome Project) used 'hierarchical shotgun sequencing' [10], whereas Celera Genomics utilised 'whole-genome shotgun sequencing' [21].

In hierarchical shotgun sequencing the target genome is first fragmented into pieces of about 100–200 kb in length. These DNA fragments are then randomly inserted into bacteria using bacterial artificial chromosome (BACs) vectors [10]. Bacteria populations carrying only one specific BAC construct, so-called clones, are isolated and propagated to multiply the DNA inserts. In order to ensure a complete representation of the entire genome in the resulting BAC library a very large number of individual BAC clones has to be isolated; the BAC library that was constructed to sequence the human genome consisted of more than 1.5 million different BAC clones. Next, BAC clones with overlapping sequences are identified and the order of their DNA inserts in the target genome is determined. An assembly of BAC clones that covers the entire genome is selected and sequenced. Around 30,000 BAC clones were sequenced by the publicly funded Human Genome Project. It is important to note that DNA fragments in BAC clones are still far too large to be sequenced directly. They are fragmented randomly into small pieces that can be sequenced with high accuracy. The DNA sequences obtained for random DNA fragments are first aligned to reconstruct the sequence of the source BAC clone before the whole sequence of the target genome is assembled.

Whole-genome shotgun sequencing circumvents the construction of a BAC library. The entire target genome is randomly fragmented into DNA pieces between 2 and 50 kb in length, which are sequenced directly [25]. Given the size of the human genome of 3.2 Gb, one can easily apprehend the enormous collection of single fragments required to warrant a complete coverage of the genome. In contrast to the hierarchical shotgun strategy, there is no initial information about the genomic location of individual DNA sequences. Therefore, correct alignment of the vast amounts of DNA fragments is the pivotal point of the whole-genome shotgun sequencing approach. Whole-genome shotgun sequencing assembly can be compared to an immense jigsaw puzzle with millions of pieces that need to be arranged in the right order and orientation. Powerful computers are employed to search for overlapping DNA sequences, to link DNA fragments of known sequence in the right order and orientation, and to compare the resulting sequence assemblies with known sequences or genetic markers in human genetic maps.

Both sequencing strategies described above have limitations. The major problem in deciphering the human genome is its high content of repetitive DNA sequences. This is in stark contrast to the genomes of simpler organisms that had been sequenced previously. Because repetitive DNA sequences do not occupy specific locations within the genome their positioning relative

to unique sequences is very difficult. This problem is more pronounced with the whole-genome sequencing strategy because the genomic origin of small DNA fragments containing repetitive sequences is not known. In the hierarchical shotgun strategy, repetitive DNA sequences can be traced back to their source BAC clones, for which additional information about their genomic vicinity is available. The major problem that resides in the hierarchical shotgun sequencing technique concerns the construction of the BAC library. Some regions of the genome are difficult to clone and will therefore be underrepresented. These cloning biases are difficult to overcome [10]. Finally, both sequencing strategies rely on computational methods to assemble the genome from sequenced DNA fragments. Complex mathematical algorithms have been developed but the computer programs employed for sequence assembly are far from being without errors. During sequence assembly, *in silico* DNA fragments might be misplaced or their orientation might be inverted. Some regions of the genome are less well resolved as others, leaving gaps in the genome sequence. Additional data have to be obtained to confirm these preliminary genomic DNA sequence assemblies, to define problematic regions of the genome, and to close the remaining sequence gaps.

3 The human genome sequence draft

The recently published human genome sequence drafts contain a number of gaps and uncertainties but have, despite their partial preliminary nature, already provided a number of important insights. The competing teams sequenced a comparable number of DNA nucleotide bases, the Human Genome Project sequenced 2.7×10^9 bases, and Celera Genomics sequenced 2.9×10^9 bases. The resulting human genome sequence drafts cover ~90% of the euchromatic portion of the genome and indicate a very similar genome composition. Celera Genomics reported that only 1.1% of the genome consists of exons and that 24% of the genome are introns. The remaining 74% of the genome are intergenic [25]. The Human Genome Project determined a protein-coding content of the genome of 1.1–1.4%. As illustrated in Fig. 1 more than half of the human genome was found to consist of different types of repeated sequence elements [10]. Further sequencing efforts by both the Human Genome Project consortium and Celera Genomics are ongoing and

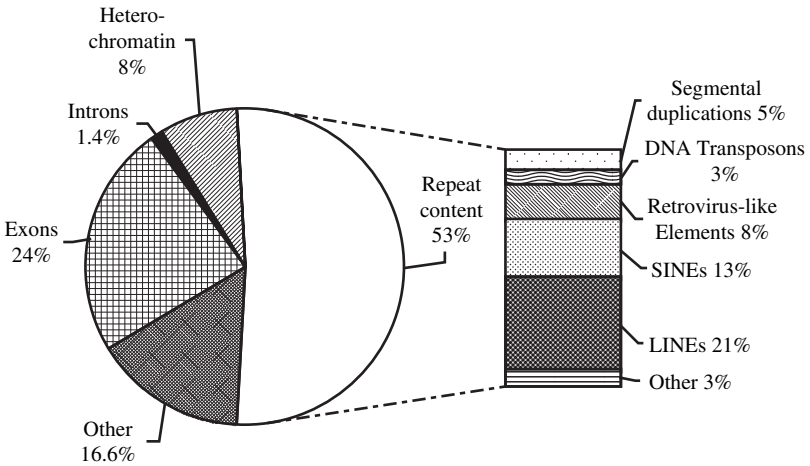


Figure 1: Human genome content.

Table 1: A selection of publicly accessible sequence databases on the World Wide Web.

http://www.ensembl.org	EBI/Sanger Centre; access to DNA and protein sequences
http://www.ncbi.nlm.nih.gov	NCBI; views of chromosomes and maps and loci with links to other NCBI resources
http://www.celera.com	Celera Genomics; central site for public access to data and tools
http://genome.ucsc.edu/	University of California at Santa Cruz; assembly of the draft genome sequence and updates
http:/genome/wustl.edu/gsc/human/Mapping/	Washington University; links to clone and accession maps of human genome
http://hgrep/ims.u-tokyo.ac.jp/	RIKEN and the University of Tokyo; overview over entire human genome structure
http://snp.cshl.org	The SNP Consortium; human SNP maps
http://www.ncbi.nlm.nih.gov/Omim	OMIM, Online Mendelian Inheritance in Man; information on human genes and disease
http://cgap.nci.nih.gov/	Cancer Genome Anatomy Project; annotated index of cancer related genes
http://www.ebi.ac.uk/swissprot/hpi/hpi.html	HPI; annotated human protein data
http://www.nhgri.nih.gov/ELSI/	NHGRI; information, links and articles on a wide range of social, ethical and legal issues

publicly accessible human genome resources are updated on a daily basis (Table 1). The complete human genome sequence is expected to be available by 2003.

3.1 Transposable DNA elements

About 45% of the human genome consists of transposable or mobile DNA elements, the most abundant class of repeat DNA sequences. These sequence elements have been acquired during our palaeontological past and have left a permanent mark in our genome. Transposable DNA elements can be divided into four types: (i) long interspersed elements (LINEs), (ii) short interspersed elements (SINEs), (iii) long terminal repeat (LTR) retrotransposons, and (iv) DNA transposons. In contrast to DNA-transposons, LINEs, SINEs, and LTR retrotransposons are transposed through RNA intermediates [30, 31].

LINEs are one of the most ancient components of the eukaryotic genome. There are about 850,000 LINE copies in the human genome, accounting for 21% of the genome sequence [10]. LINEs are $\sim$6 kb in length and harbour a promoter specific for RNA polymerase II and two open reading frames (i.e. DNA sequences that are potentially translated into proteins). LINE transcripts associate with their own protein products, one of which is a reverse transcriptase, an enzyme that catalyses the reverse transcription of RNA into DNA. Reverse transcription of

LINE RNAs yields complementary DNA that can be integrated back into the genome. LINEs are thought to be responsible for most of the reverse transcription activity in the human genome [31].

SINEs are shorter than LINEs and are typically between 100 and 400 bases in length. SINEs do not contain protein-coding sequences, but instead harbour a promoter specific for RNA polymerase III that is typical for tRNA genes. There are about 1.5 millions SINEs in the human genome accounting for 13% of human genomic DNA [10]. SINEs share sequences elements with LINEs and are thought to make use of the LINE reverse transcriptase machinery for transposition [31].

The most common SINE element is the Alu element, which is also the only known active SINE element in the human genome [31]. Alu elements were initially dismissed as parasitic and non-functional DNA sequences. However, in many species Alu elements are transcribed under conditions of stress. Alu RNAs can specifically bind double-stranded RNA-induced kinase (PKR), which under normal conditions inhibits translation of mRNA into proteins. Thus, binding of Alu RNA to PKR counteracts inhibition of translation and may therefore serve to stimulate protein production under stress. Consistent with the idea of positive Alu element function, Alu elements are preferentially found in regions of the genome that are rich in actively transcribed genes [30].

The third class of transposable elements is LTR retrotransposons that resemble retroviruses, RNA viruses that make use of reverse transcription to integrate into the host genome. There are about 450,000 copies of retrovirus-like elements in the human genome, accounting for 8% of human genomic DNA [10]. LTR retrotransposons encode for a number of proteins, including a reverse transcriptase [31].

The fourth class of transposable elements is DNA transposons, which encode a DNA transposase that mobilises the transposon element by a 'cut-and paste' mechanism [31]. The human genome contains about 300,000 DNA transposons that account for 3% of the genome sequence [10].

The sequence divergence within different classes of mobile DNA elements can be analysed to estimate their approximate age. Most DNA transposons found in the human genome are older than 100 million years and appear to have been inactive during the past 50 million years. LINE1, the predominant member of the LINE class and the only known active human LINE, is estimated to be more than 150 million years old. Alu elements are thought to exist in the human genome for at least 80 million years. LTR retrotransposons appear to be on the brink of extinction and only one LTR retrotransposon family appears to be active since humans diverged from chimpanzees [10].

While the human genome is filled with ancient transposons, the genomes of other organisms have been found to contain transposable elements of a more recent origin. The relatively old age of human mobile DNA elements are testament for an extremely slow rate by which non-functional elements are cleared from vertebrate genomes. The half-life of non-functional DNA elements is estimated to be 800 million years in humans but only 12 million years in the fruitfly (*Drosophila*) [10].

DNA sequence comparisons between the human and mouse genomes have led to some striking observations. The distribution of various classes of transposons is similar in both genomes. However, the activity of transposons in the mouse has not declined in the same way as in humans. The mouse genome contains a number of active transposon families that might contribute to a higher mutation rate compared to humans. It is estimated that 10% of all mutations found in the mouse genome are due to DNA transpositions into genes and this number appears to be 60 times lower in humans. Thus, evolutionary forces appear to affect the persistence of active transposable elements in humans and mice differentially [10].

Clues to the possible role or function of transposable DNA elements can be gained from their relative distribution within genomes. LINEs accumulate in regions of the genome that are rich in the nucleotide bases A and T, whereas SINEs, and especially Alu elements, are enriched in G/C-rich regions [30]. G/C-rich DNA sequences coincide with a high density of genes but their exact

biological functions have yet to be determined [32]. The preference of LINEs for A/T-rich regions that have a low gene content may reflect a mechanism by which parasitic DNA elements can persist in a genome without causing damage to the host organism. In this regard, the accumulation of SINEs in G/C-rich genome regions is quite puzzling, especially since SINEs rely on the LINE machinery for their transposition. A possible explanation might be that SINEs initially insert in A/T-rich regions with the same preference as LINEs but that evolutionary forces subsequently change the relative distribution of SINEs and LINEs within the genome. In light of their proposed positive function in stimulating protein production under stress, Alu elements may be regarded as genomic symbionts [30]. This may help to explain the preferential positioning of Alu sequences close to actively transcribed genes in both G/C-rich and A/T-rich regions of the genome.

Mobile DNAs are also responsible for innovations in the host genome, e.g. by introducing novel regulatory DNA elements or even novel genes. An example for acquired regulatory DNA elements are transcription terminator regions that are thought to originate from LTR retrotransposons. Twenty human genes are currently recognised to originate from transposons [30]. These include the genes *RAG1* and *RAG2* that encode the lymphocyte-specific proteins of the V(D)J recombination system responsible for antigen-specific immunity. Our immune system therefore may have originated from an ancient transposon insertion [33]. Another example for the utilisation of a transposable element is the *BC200* gene. BC200 is a brain-specific RNA located in the dendrites of all higher primates and most likely derived from an Alu element about 50 million years ago [30]. Further examples of human genes derived from transposable elements include the gene for telomerase, the enzyme responsible for the proper maintenance of chromosome ends [34], and CENPB, the major centromere-binding protein [35]. Preliminary analyses of the human genome sequence revealed 27 additional candidate genes that are suspected to originate from mobile elements and that will have to be characterised in the future [10].

Side effects of transposon activity are also observed. For example, reverse transcription of genic mRNAs by LINEs, which generally results in non-functional pseudogenes, can occasionally give rise to functional processed genes. It is believed that many intron-less genes have been created in this manner [10]. In addition, active transposable elements have been recognised as the cause of human disease. Haemophilia A, a disorder of blood coagulation that is linked to the X chromosome, is caused by disruption of the gene for a protein called factor VIII through insertion of LINE1 transposable elements. Examples for Alu elements involved in human disease include insertions into the factor IX gene as cause for haemophilia, insertions into the *APC* gene as cause for desmoid tumours, and insertions into the *BRCA*2 gene as cause for breast cancer [36].

3.2 Gene content

The ultimate aim in deciphering the human genome sequence was to compile a list of all human genes. This is a daunting task because genes represent only a very small proportion of the human genome and computer programs employed in gene finding are confronted with a significant signal-to-noise problem. Gene finding and gene prediction employs three basic approaches [10, 25]: (i) the identification of genes on the basis of known mRNAs [37], (ii) the identification of genes on the basis of sequence similarities to previously identified genes or proteins in other species [38], and (iii) ab initio recognition using DNA sequence databases and statistical information [25, 39–41]. All approaches to find or predict genes hold sources of errors. On one hand, genes might be missed because they are expressed only in a subset of cells or at very low levels and their mRNAs are thus undetectable [10]. On the other hand, the total number of genes in a genome tends to be overestimated because the different parts of long and complex genes can be misinterpreted as several distinct genes. Furthermore, the set of predicted genes identified solely on the basis of DNA

sequences characteristics typical for active genes might include pseudogenes. The development of advanced computer algorithms that allow gene finding with high accuracy will require a far more detailed understanding of the cellular mechanisms by which genes are recognised within the bulk of nuclear DNA.

The Human Genome Project reported the identification of about 15,000 known genes and predicted the existence of another 17,000 genes [10]. Celera Genomics reported strong evidence for about 26,000 human genes and weak evidence for 12,000 additional genes [25]. In view of earlier predictions ranging from 60,000 to 100,000 human genes, these numbers are unexpectedly low [42]. By comparing 30,000–40,000 human genes to 13,000 genes in the fruitfly *Drosophila melanogaster*, 18,000 genes in the primitive worm *C. elegans*, and 26,000 genes in the mustard weed *Arabidopsis thaliana*, it becomes evident that increased complexity of multicellular organisms is not simply achieved by using many more genes [43]. The human genome is the first vertebrate genome sequenced and more suitable sequence comparisons will be possible once the genomes of more closely related species become available. Of particular interest is the genome sequence of the chimpanzee, our closest relative, since it may hold the answer to the question whether the most significant advancements in humans, such as the origin of speech and the ability of abstract reasoning, are actually manifested in the genome sequence itself or whether they evolved from more subtle changes, for example, in specific gene expression patterns [44].

There is considerable variation in the size of the exons and introns and, consequently, in the size of protein-coding genes. On average, human genes are ~27 kb long, but many genes exceed 100 kb in length. The longest known human gene is the *dystrophin* gene (DMD) which spans over 2.4 Mb. The *titin* gene contains the largest number of exons, 178 in the Human Genome Project draft [10], and 234 in the Celera Genomics draft [25]. It also contains the longest known coding sequence at 80,780 bases and the longest single exon at 17,106 bases [10]. The average length of exons in humans is comparable to those found in the fruitfly or in the worm. Most exons in human genes are between 50 and 200 bases long. However, the intron size distribution differs significantly between fly, worm and humans. In humans, introns tend to be much longer with an average size of 3.3 kb and this variation in intron length results in a larger average gene size [10].

The distribution of the four different DNA nucleotide bases A, G, C, and T over the human genome is not uniform. Large genome segments with either high or low G/C content can be distinguished. Earlier studies had indicated that G/C-rich genome regions contain a higher gene density compared to regions that are low in G/C [32]. Results of the detailed analysis of the human genome sequence draft are broadly consistent with these observations. However, the actual proportion of genes located in genome segments that are relatively poor in G/C was found to be significantly higher than previously predicted [25].

The human genome can be described as oases of genes in a desert of non-coding DNA sequences. About 20% of the genome consist of very long gene-less DNA segments. The distribution of these gene deserts varies across the genome. Chromosomes 17, 19 and 22 contain the highest density of genes and only a small percentage of DNA sequences reside in gene deserts. The situation is reversed on chromosomes 4, 13, 18, and the sex chromosomes that contain a low gene density and a high percentage of gene deserts. It is noteworthy that genes deserts are not necessarily devoid of biological function.

Of particular interest is the distribution of CpG islands. The dinucleotide 5′-CpG-3′ ('p' designates the phosphate residue that connects the nucleotide residues) is underrepresented in the human genome because most CpGs become methylated at the cytosine base. This results in a spontaneous chemical reaction that ultimately leads to mutation of the CpG dinucleotide to TpG. CpG islands are regions of the genome containing unmethylated, and therefore stable, CpG dinucleotides. CpG islands are often associated with active genes, in particular with DNA regions

close to the start site of transcription [45]. Consistent with this observation, CpG islands have been shown to play important roles in the regulation of gene transcription and in gene imprinting, a process that determines gene activity in cell lineages during development and differentiation [46]. The Human Genome Project reports 28,890 CpG islands, and Celera Genomics counts a total of 28,519 CpG islands [10]. Both numbers are remarkably close to the number of predicted genes.

3.3 Single nucleotide polymorphism

Only 0.1% of the genome contribute to phenotypic variation amongst humans and, with the exception of identical twins, the genomes of two individuals are about 99.9% identical. Most of the DNA sequence variation between humans can be attributed to changes in DNA sequences at a single base pair, so-called single nucleotide polymorphisms (SNPs). SNPs occur on average every 1.9 kb when two chromosomes are compared. The International SNP Map Working Group presented a map of 1.42 million SNPs distributed throughout the entire genome [47]. Celera Genomics assigned 2.1 million SNPs to the genome [25]. Both groups reported that the genomic distribution of SNPs is markedly heterogeneous. About 75% of all SNPs are found outside genes. Within genes, the SNP rate is highest in introns and less than 1% of all SNPs are found in DNA sequence regions coding for proteins and therefore potentially affect protein structure and/or function. Nevertheless, this low percentage comprises thousands of candidate SNPs that may significantly contribute to the diversity of human proteins.

The identification of specific SNPs and their functional consequences is one of the major objectives of future genetic studies. It is well established that genetic variations affect the susceptibility to disease, the age of disease onset, the severity of illnesses, and the way the human body responds to medical treatment [48–50]. For example, single base differences in the *ApoE* gene have been implicated in Alzheimer's disease. The *ApoE* gene exists in three variants, *ApoE2*, *ApoE3*, and *ApoE4*, and individuals carrying the *ApoE4* version of the gene are the most likely to develop Alzheimer's disease [51]. Large-scale studies of SNP patterns in patients and healthy individuals will help to identify the molecular basis of many other diseases in the future.

A complete map of SNPs will also be prerequisite for detailed studies into the molecular basis for human phenotypic variation. SNP patterns entail a snapshot of the actions of evolutionary forces that are operative in human population genetics. For example, it could be demonstrated that our genes carry the signature of an expansion from Africa within the past 150,000 years [52]. A complete map of human SNPs is expected to fuel future research aimed to explore our evolutionary past and to discover the origin of our present diversity.

3.4 The human proteome

Analogous to the genome, the proteome represents the complete set of all proteins within an organism. Proteins typically consist of several discrete structural or functional domains that are conserved during evolution. More than 90% of protein domains in humans have counterparts in the fruitfly or the worm. About 60% of the human proteins that are predicted to exist based on the human genome sequence draft show sequence similarities to other organisms whose genomes have been sequenced. Also, 61% of the fly proteome, 41% of the worm proteome, and 46% of the yeast proteome have sequence similarities to predicted human proteins [53].

The draft of the human genome brought to light about 1200 protein families. Only 94 protein families, or 7%, appear to be vertebrate-specific suggesting that only a small number of novel protein domains were introduced into the vertebrate lineage. Vertebrate-specific protein families

reflect important physiological differences between vertebrates and other metazoans. A large proportion of these proteins exhibit functions in the immune and nervous systems [10].

Although there is only a small number of novel human protein families, there is substantial innovation in the creation of new proteins. New proteins can be created by rearrangement, insertion, or deletion of protein domains, resulting in new domain architectures. This mechanism is especially prominent in human proteins involved in extracellular structures and transmembrane structures where the total number of human domain architectures is more than twice of those found in the worm or the fruitfly [10]. A genome-wide analysis of domain architectures will be extremely helpful in resolving the evolutionary history of different species. About 40% of proteins predicted by the human genome sequence draft are of unknown function and cannot be assigned to known protein families [25]. A large proportion of proteins with known functions are either enzymes that play a crucial role in the cell metabolism, or proteins involved in signal transduction processes that are essential for intra- or inter-cellular communication.

The most common molecular function of human proteins is nucleic acid binding, employing 13.5% of the human proteome [25]. Nucleic acid-binding proteins include sequence-specific DNA-binding factors responsible for the regulation of gene transcription, and enzymes that participate in nucleic acid metabolism. Given the crucial importance of establishment and maintenance of cell type-specific gene expression patterns in multicellular organisms, it is not surprising that are a significant part of the proteome is engaged in gene regulation. A search of the human genome sequence revealed more than 2000 hypothetical genes that encode potential transcriptional activators [54]. These transcription factors need to be verified, biochemically characterised, and their target genes identified, before mechanisms of genome-wide transcription regulation processes can be fully elucidated.

Remarkably, a set of around 200 human proteins has significant amino acid sequence similarities to bacterial proteins, but not to any proteins found in yeast, worm, or the fruitfly [10]. These proteins appear to be of bacterial origin and were possibly acquired by gene transfer. Some of these genes are involved in metabolism and stress response, suggesting that they may have provided a selective advantage for the host organism during evolution.

The greater complexity of the human proteome as compared to the worm or the fruitfly is achieved only in part by the invention of novel proteins and novel protein architectures on the DNA level [10]. Additional levels of complexity arise from mechanisms such as alternative splicing and post-translational modifications. In addition, there are a bewildering number of potential interactions between individual cellular proteins that might affect their activity or function. The regulation of protein–protein interactions within the cell is considered to contribute significantly to the functional complexity of the human proteome.

A major objective of future research will be to decipher the human proteome and to ultimately identify the protein networks and functional pathways that give rise to complex multicellular organisms such as ourselves.

4 Functional genomics: assigning function to the genome

Biological functions are generally not evident from raw genome sequence data. For about 40% of the human genes, DNA sequence analysis has led to no prediction of function. In addition, the inferred functions of most of the remaining genes have yet to be confirmed [10, 25]. Functional genomics aims to determine the biological function(s) of individual genes or genome segments and comprises different areas including comparative genomics, gene expression studies, proteomics, and structural genomics. The combined data obtained from these areas will be required

to understand both the individual and collective function of genes, and will be prerequisite for a complete biochemical comprehension of cell biology.

4.1 Comparative genomics

Genome sequences comparisons between species are extremely valuable to elucidate innovations during evolution and to determine the timing of the divergence of species. DNA segments with important functions are more likely to retain their sequences during evolution than non-functional segments. Thus, conserved sequences between species are likely to point to important functions in key biological processes. Biological studies over the last century have made use of a number of key model organisms, including protozoans such as the yeasts *S. cervisiae* and *Schizosaccharomyces pombe*, metazoans such as the fruitfly (*D. melanogaster*) and the worm (*C. elegans*), and vertebrates such as zebrafish (*Brachydanio rerio*) and mouse (*Mus musculus*). There are two principal experimental approaches to identify and functionally characterise genes in animal models: forward and reverse genetics. Forward genetics starts with a mutant phenotype, which identifies the function of a gene. However, the identification of the corresponding DNA sequence within the genome by conventional gene-mapping techniques is a very time-consuming and laborious process. In contrast, reverse genetics starts from the DNA sequence of a known or predicted gene and attempts to gain insights into its function by obtaining phenotypic changes in model organisms upon gene mutation or gene deletion (knock-out). Complementary to both strategies, genetic crossing experiments are employed to examine the functional interplay of different genes. These studies were instrumental in the dissection of a number of fundamental metabolic and signalling pathways that are evolutionary conserved between species [55].

Genetically well characterised organisms such as the yeast *S. cerevisiae*, the fruitfly *D. melanogaster*, and the worm *C. elegans* were initially chosen for complete genome sequencing [22]. The genome sequences of these organisms proved to be invaluable both for the identification of human genes and the assignment of human gene function. In addition, detailed comparisons of the human genome sequence draft with the genome sequences of these distantly related organisms led to the identification of a number of vertebrate-specific candidate genes. However, confirmation of the vertebrate-specific nature of these genes and further elucidation of their function will require genome sequence comparisons with more closely related species such as the mouse. The mouse genome sequencing is well under way and is carried out by the publicly funded Mouse Sequencing Consortium and by Celera Genomics. Databases containing draft sequences obtained by the Mouse Sequencing Consortium are already accessible and completion of the mouse genome project is imminent. The available data have already revealed striking similarities between the human and mouse genomes. More than 180 cases of synteny, the presence of conserved DNA sequence segments that contain the same genes in the same order, have been found. Almost all genes on human chromosome 17 are found on mouse chromosome 4, and human chromosome 20 appears to be almost completely orthologous to mouse chromosome 2. The average length of the conserved segments is 14.5 Mb. The largest contiguous conserved segment found so far spans about 90 Mb on human chromosome 4 and corresponds to mouse chromosome 5 [10]. The completion of the mouse genome project is greatly anticipated and will further enhance our understanding of gene function in fundamental biological processes. Many human diseases with complex genetic background have counterparts in the mouse or rat. Therefore, knowledge of the mouse genome sequence will be instrumental for the diagnosis, prevention and treatment of human disease. In a number of instances, a conserved genome region containing a locus that contributes to a complex genetic disease involving several genes, a so-called quantitative trait

locus (QTL), could already be identified. Prominent examples for known human disease QTLs identified in animal models are cardiovascular disorders such as hypertension and atherosclerosis [56–58]. Large-scale genome-wide mutagenesis projects in mice have been set up in the UK [59] and in Germany [60] with the aim to screen thousands of mice mutants for links between DNA sequences and function.

To understand the controlled and coordinated read-out of genetic information, the identification of regulatory DNA sequences is of crucial importance. However, the identification of regulatory DNA sequences within the complex genomes of higher eukaryotes is extremely difficult. Transgenic studies have shown that human genes when introduced in mice are expressed in a manner that mimics their expression in their natural host. This observation suggests that the instructions for regulated gene transcription are evolutionary conserved [61]. Inter-species DNA sequence comparisons will therefore greatly facilitate the identification and functional characterisation of regulatory DNA elements.

Understanding coordinate gene regulation at the genome-wide level requires the identification of gene regulatory networks. Co-expression of certain genes may reflect their regulation by common sequence-specific transcription factors. Sequence comparisons between genes can therefore be used to identify common DNA elements that might be responsible for their coordinated expression. This approach is becoming increasingly feasible with the availability of genome-wide expression profiling data, in particular for smaller genomes. Genome-wide expression profiling in yeast has already been successfully used to identify regulatory networks involved in the coordinate expression of gene clusters during sporulation [62] and cell cycle progression [63]. Similar approaches to identify regulatory DNA sequence elements in mammals face challenges that do not exist in simpler organisms such as yeast. While regulatory sequences in yeast are typically located in close proximity to the transcription start site, regulatory elements in mammals are frequently found in great distances from their target genes. The size and complexity of mammalian genomes and their high content of non-coding sequences further complicates the identification of regulatory elements. Detailed DNA sequence analyses between closely related species may help to overcome these obstacles. Indeed, a comparison of the DNA sequences of mouse and human genes that are up-regulated in skeletal muscle revealed novel muscle-specific regulatory elements [64].

4.2 Proteomics

Comprehension of the genome at the proteome level will be prerequisite for a complete understanding of the functioning of a human cell. As outlined above, mechanisms such as alternative splicing of primary gene transcripts and post-translational modifications of gene products can considerably increase the complexity of the proteome over the genome. Indeed, the total number of different protein molecules expressed by the ∼30,000 genes in the human genome is estimated to be in the order of 10^6. In addition, the activity and/or function of individual proteins may be subject to regulation, e.g. by specific protein–protein interactions, by targeting specific cellular compartments, by covalent modifications, and by protein degradation. This notion has led to independent efforts in proteomics, the study of protein function, subcellular localisation, and protein–protein interactions on a genome-wide scale [65]. Combinations of well-established biochemical and molecular biology methods, e.g. the combination of two-dimensional protein electrophoresis and mass spectrometry, are employed.

In order to combine and extract as much relevant information as possible a number of database resources were integrated into the human proteomics initiative (HPI), a joint effort between the Swiss Institute of Bioinformatics and the European Bioinformatics Institute. The HPI has

two phases. The first phase aims to annotate the protein products of all known human genes. The second phase is a long-term commitment to rapidly incorporate well-annotated protein data from ongoing research efforts and to provide the scientific community with free access to continuously updated databases. The HPI database contains currently over 6000 annotated human sequences as well as relevant information such as literature references, predicted post-translational modifications, splice variants, and known SNPs. Databases are being developed in order to allow automated annotations of predicted proteins from genome DNA sequences, and to facilitate a classification of proteins into cell type-specific proteome subsets [66]. These computational methods will provide meaningful tools to exploit the full potential of human genome sequence for basic and medical research by integrating biological and human genome data.

4.3 Structural genomics

The results of genome sequencing and recent advances in structure determination have ushered in structural genomics, a new field focused on the large-scale analysis of protein structures and functions. Three-dimensional high-resolution structures of proteins are required for understanding the molecular chemistry underlying their biological action. Protein structure determination can be a difficult and time-consuming process and the two major techniques used, X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy, possess certain drawbacks. X-ray crystallography requires the preparation of stable protein crystals, generally a laborious task, and NMR spectroscopy is limited to small and medium-sized molecules. However, technical advances are made and speed and productivity of macromolecular structure determination is continuously improved [67].

Structural genomics aims to provide structural information for all known cellular proteins by employing automated high-throughput methods. To achieve this goal, proteins are grouped into protein families based on their amino acid sequences and the structure of at least one family member is solved. Structure predictions for the remaining family members can then be made based on the known structures. Estimates for the total number of protein structure families range from 30,000 to 50,000, several orders of magnitude below the estimated total number of proteins in nature. Obviously, an initiative of this scale requires coordinated efforts of different disciplines such as biology, chemistry, physics, and bioinformatics.

The next pivotal step is to assign biological function to particular protein structures. High-resolution structures rarely give immediate insight into the function of a protein, but a rigorous analysis, for example, of key residues or protein surfaces may give functional clues. In addition, proteins with a low degree of amino acid sequence similarity may reveal unanticipated structural similarities. This may provide evidence for evolutionary links between protein families unrelated by amino acid sequence. The accuracy of the prediction of protein function based on structural information is expected to increase proportionally with the number of high-resolution protein structures available. However, the biological relevance of functional predictions based on structure has to be validated by independent biochemical or genetic studies.

There are a number of ongoing structural genomics projects. Japan was one of the first countries to recognise the importance of structural studies on a genome-wide scale and founded the RIKEN Structural Genomics Initiative in 1995 [68]. In the US, the National Institute of General Medical Sciences inaugurated its 'Protein Structure Initiative' in 1998, which is designed to organise a large cooperative effort in structural genomics and to produce a database resource which links sequence, structural, and functional information. Pilot projects in the US and in Canada have been set up to determine the structure of 10,000 proteins within the next 10 years [69]. Each project consists of a collaboration between different research groups and focuses on proteins

from different species or from different proteins families. There is at present little coordination between ongoing national efforts in Europe [70]. However, the Wellcome Trust in the UK is considering setting up a Structural Genomics Consortium, modelled on the highly successful SNP Consortium, including publicly accessible up-to-date protein structure/function databases.

5 Applications of the human genome sequence in medical sciences

Acquisition of the human genome sequence has led to the emergence of new technologies that promise to accelerate the identification and functional characterisation of genes involved in human disease and to facilitate major steps of drug development including drug target identification and validation, optimising drug efficacy, and reducing drug toxicity.

5.1 Genes and human disease

Many diseases have their origin in gene mutations. A prominent example is Haemophilia A, an inherited disease that affected the male lines of royal dynasties in Europe for several centuries. Haemophilia A is inherited recessively on the X chromosome. Heterozygotic females carrying only one disease variant of the gene are healthy, whereas males who carry the disease gene on their only X chromosome are haemophilic [71]. At present, more than 12,000 mutations associated with human diseases have been mapped to specific loci in the genome. The catalogue of these disorders is listed in a public database, Online Mendelian Inheritance in Man, (OMIM, http://www.ncbi.nlm.nih.gov/omim; see also Table 1). OMIM is updated almost daily, with new genes and mutations being discovered at a rapid pace.

Genetic disorders are classified by the number of mutated genes involved. Monogenic diseases are caused by mutations in a single gene, whereas polygenic diseases involve several mutated genes. Monogenic diseases are very rare compared to more complex polygenic diseases that affect several million people worldwide. Prominent examples of monogenic genetic diseases are cystic fibrosis (CF) and Huntington's disease (HD). CF is the most common potentially lethal genetic disease, affecting one in 2500 newborns in Northern Europe. The predominant symptoms of CF are obstructions of the airways by viscous mucus and, as a consequence, frequent bacterial infections of the respiratory tract. CF is inherited in a recessive fashion; heterozygote children carrying one mutant allele and one wild-type allele of the CF gene are healthy, whereas homozygote children carrying two mutant CF alleles develop CF [72]. The gene involved in the occurrence of CF was identified in 1989 by positional cloning [73–75]. Initially, the CF locus was mapped to chromosome 7 [73]. In an intense collaborative effort, sequential refinement of the genetic analysis yielded finally a candidate gene encoding a protein product that was called CFTR, for cystic fibrosis transmembrane conductance regulator [74]. It was subsequently found that 70% of all CF cases carried a three base pair deletion in the *CFTR* gene that caused the loss of a single amino acid residue in the CFTR protein product, resulting in its malfunction [75]. Cloning of the *CFTR* gene made biochemical studies possible and raised hopes for a gene therapy approach to CF [72]. The concept and strategies implicated in gene therapy are discussed in Section 5.6. HD is a neurodegenerative disease which typically affects individuals about 40 years of age and which progresses inexorably to death within 15 to 20 years [76]. Symptoms include alterations in mood, loss of memory, and involuntary movements. HD is caused by the insertion of CAG triplets into the gene *Huntingtin* (*Htt*). These additional DNA sequences lead to an extended stretch of glutamine amino acid residues in the corresponding protein product (Htt^{ex}). HD is inherited as a dominant trait; inheritance of only one Htt^{ex} allele is sufficient for the development of the disease.

It is not yet understood how the Htt^{ex} protein causes the degeneration of neurons, but aggregation of Htt^{ex} within cell nuclei appears to be, at least in part, responsible for its toxicity. Clearance of the mutant Htt^{ex} aggregates is therefore viewed as the most promising therapeutic strategy and medical research on HD is focussing on this approach [76].

Polygenic diseases represent traits that are caused by interactions between different genes, but are also frequently influenced by environmental factors such as smoking, the type of diet, exercise habits, and childhood exposure to infections [55]. Prior to the past decade, a connection between complex diseases and genetic factors was often not considered because genetic traits in patients did not follow an obvious pattern of inheritance. Only recently, a number of complex disorders with hitherto unknown origin could be linked to inherited genetic variations. While some progress has been made in identifying genes that contribute to complex diseases, the elucidation of underlying molecular mechanisms remains extremely difficult. The following findings exemplify earlier genomic research that has provided important insights into the pathogenesis of complex genetic diseases and that has paved the way for the development of novel therapeutics.

Hypertension (elevated blood pressure) has been associated with mutations in eight different genes and nine specific genes were found to contribute to hypotension (lowered blood pressure). The corresponding gene products act in a common regulatory pathway operative in the kidney that controls the net salt balance in the human body [77]. Heart failure is a predominant health-problem in industrialised countries and affects more than four million people in the United States alone. A common clinical type of heart failure is hypertrophic cardiomyopathy in which the mass of the heart is increased by a thickening of the left heart chamber wall. Ten different genes have been identified that, when mutated, can cause hypertrophic cardiomyopathy. The protein products of these genes are involved in different steps of the heart muscle contraction process [78]. Finally, cardiac arrhythmia is the failure of the heart to sustain a precise rhythm. It affects more than 300,000 individuals per year in the United States alone and can lead to sudden death. In recent years, six arrhythmia susceptibility genes have been discovered. Mutations in these genes affect ion channel proteins of heart cells [79].

5.2 Genetic basis of cancer

Cancer affects one in three people in the western world and is responsible for 20% of all deaths [80]. Throughout life, the DNA in human cells is exposed to mutagens, such as the ultraviolet component of sunlight and ionising radiation. This causes a progressive decay of genetic information that can occasionally lead to functional alterations in genes critical for cell proliferation. Gene mutations in only a single cell can be sufficient to give rise to cancer – the emergence of a cell population in which the mechanisms that control normal cell division and cell death are suspended. In addition, cancer cells can also acquire the ability to invade different tissues and to metastasise [81].

Tumorigenesis, the development of cancer, is a multistep process. It has been proposed that four to seven distinct events are required for the development of common epithelial cancer [82]. These events reflect genetic alterations that drive a progressive conversion from normal human cells into cancer cells. Events leading to tumorigenesis can either be the result of environmental influences (somatic), or inherited. Patients with inherited forms of cancer carry transforming mutations in every cell of their body, whereas in patients with somatic cancers, these mutations are found only in tumour cells. The vast majority of mutations in cancer are somatic and inherited forms of cancers contribute only to about one percent of all cancer cases [83]. Family history is the cardinal feature by which an inherited predisposition for cancer is recognised. However, inherited mutations in a cancer gene do not necessarily indicate similar probabilities for different

individuals to develop cancer; strong and weak predisposition can be distinguished. For certain cancers, e.g. breast cancer, mutations that result in a high cancer risk have been identified [84]. Women carrying mutations in the genes *BRCA-1* and *BRCA-2* have a lifetime breast cancer risk of 60–80%. However, only 15–20% of all inherited breast cancer cases are attributable to mutations in either gene. For the remaining 80%, mutations in different genes must be responsible for increased breast cancer susceptibility. A number of additional candidate genes have been proposed but require conformation by independent analyses [85].

Patterns of genetic alteration differ between different cancer types, probably because different tissues impose different constraints that need to be overcome. However, genetic alterations leading to cancer are not random, suggesting that cancers evolve along particular pathways [84]. This notion has lead to the optimistic view that determining the genetic alterations specific for a particular type of tumour cells, so-called molecular profiling, will provide information of clinical value, such as the future malignancy potential. It is hoped that this information can be used to tailor therapeutic approaches to individual cases [86]. More than 100 distinct types of human cancers with subtypes have been found within specific organs and all cancer types can be attributed to six essential alterations in cell physiology that collectively dictate malignant growth. Among these cancer-induced physiological changes are the capacity to sustain prolonged cell growth, the evasion of controlled cell death (apoptosis), and the ability to invade different tissues. Each acquisition of a cancer capability successfully breaches an anticancer defence mechanism hardwired into cells and tissues. However, the paths that different cells can take on their way to malignancy are highly variable and, in a given cancer type, specific gene mutations may be found only in a subset of tumours. Furthermore, mutations in critical genes resulting in the acquisition of capabilities that override normal cellular mechanisms may occur at different times during cancer development. Finally, genetic changes may affect various tumours differentially. In some tumours, a specific mutation might lead to the acquisition of only one cancer capability, whereas in other tumours the same genetic event could facilitate the simultaneous acquisition of several distinct capabilities. Despite the evident complexity of these pathways, it is believed that all tumours ultimately reach common biological endpoints, the acquisition of cancer capabilities [81].

Two major types of cancer genes can be distinguished. Oncogenes typically harbour mutations in their DNA sequence and the corresponding mutated protein products can cause the development of cancer. Oncogenes were first isolated from viruses capable to transform cultured human cells into cancer cells. Subsequently, it was discovered that human cells contain homologues to viral oncogenes that are involved in normal cellular functions [87]. These cellular genes were termed proto-oncogenes and their mutation or aberrant activation can promote transformation of normal cells into cancer cells [88]. A number of proto-oncogenes, e.g. oncogenes of the *Ras* family, could be identified in studies in which 'normal' cells were transfected with DNA isolated from animal tumours.

Tumour-suppressor genes serve to prevent the formation of cancer and their inactivation contributes to tumorigenesis [89]. Tumour suppressor genes were first proposed in studies of cancer in children [90]. Retinoblastoma, a cancer affecting one or both eyes in young children is in 35–40% of all cases inherited. Statistical analysis of inherited retinoblastoma cases revealed a requirement for one transforming event that occurred with constant probability over time. In contrast, the appearance of the sporadic form of retinoblastoma, which occurs usually much later in life, was consistent with a requirement for two such events. These observations suggested that two events were necessary for both the somatic and inherited forms of retinoblastoma and that in the case of the inherited form one event was already present in the germ line. Subsequently, a deletion in a region of chromosome 13 was found in some cases of inherited retinoblastomas and further studies confirmed that tumorigenesis required the loss of function of both copies of this specific

chromosomal region [91]. The *Rb* gene was finally identified to be responsible for retinoblastomas and is now regarded as the prototype of tumour-suppressor genes. Many more tumour-suppressor genes have been identified since, especially in inherited forms of cancer. Oncogenic mutations in inherited cancers are rare, presumably because they promote cancer during early stages in development and are therefore embryonic lethal [84]. Tumour-suppressor genes can be classified into different types. 'Classical' tumour-suppressor genes such as *Rb*, of which loss-of-function is required for cancer development, are termed 'gatekeepers'. Another class of tumour suppressor genes termed 'caretakers' contains genes involved in functions outside the actual pathway of cancer development. These genes are important for normal DNA repair and genome integrity and their mutation accelerates the acquisition of cancer capabilities [92].

A particularly complex mechanism by which loss-of-function or gain-of-function of genes can be acquired is the perturbation of epigenetic regulation. Epigenetic regulation of gene expression is an important process, which insures the expression of certain genes only at specific stages in development. In many cases, developmental genes are permanently switched off after they have fulfilled their function. Permanent silencing involves an inheritable marking of genes, e.g. the methylation of cytosine residues within promoter regions [93]. Perturbations in epigenetic gene regulation can lead to either expression of genes that need to be silenced or, conversely, to the silencing of genes, such as tumour-suppressor genes, that need to be active. It is not clear whether silencing of particular genes in cancer occurs through a stochastic process or whether certain promoters are predisposed. It is also unclear what exactly determines the particular molecular mechanism by which loss-of-function events occur since the frequency of distinct mechanisms can differ considerably between tumour types [94].

5.3 Identification of disease genes and disease pathways

Application of the human genome sequence is anticipated to accelerate both medical genetics and its application, the targeted treatment of genetic diseases. The identification of candidate genes involved in human disease had been extremely laborious and time-consuming. For example, identification and characterisation of the human CTFR gene involved in CF took several years [73]. In contrast, about 30 disease genes could be identified and their chromosomal location determined during the period in which the human genome was sequenced [10]. The human sequence database is further used to identify paralogues, genes that arose as a consequence of gene duplication within the genome and therefore contain closely related DNA sequences. Finding paralogues of disease genes is important for two reasons. First, paralogues can give rise to related genetic diseases. For example, the *CNBG3* gene has been identified as a paralogue for the *CNGA3* gene that, when mutated, causes colour blindness [95, 96]. Another example is the discovery of the *Presenilin-1* and *Presinilin-2* genes that can cause the early onset of Alzheimer's disease [97]. Second, paralogues may provide the means for therapeutic intervention, as exemplified by attempts to reactivate foetally expressed haemoglobin genes in individuals suffering from sickle cell disease or β-thalassaemia, caused by mutations in the *β-Globin* gene [98]. A complete scan of the human genome sequence revealed more than 200 potential paralogues that will have to be experimentally confirmed and characterised [10].

The identification of disease genes, and in particular the identification of sets of genes underlying complex polygenic disorders, by genome-wide DNA sequence comparisons between healthy individuals and patients is extremely complex due to the large number of single nucleotide polymorphisms present within the human genome sequence (see Section 3.3) [47]. SNPs responsible for human disease have to be distinguished from sequence variations that have no apparent detrimental effect. In addition, some SNPs may not affect disease immediately but may nevertheless

increase the susceptibility of individuals towards specific diseases. A large number of genomic sequences from healthy individuals and patients will have to be compared to establish a comprehensive catalogue of SNPs.

Alterations in cellular gene expression patterns that coincide with the manifestation of disease can be analysed using DNA microarray (DNA chip) technologies. DNA microarrays containing many thousand DNA molecules, each specific for a single gene, can be used to measure the levels of individual RNA species throughout the entire cellular RNA population [99, 100]. By comparing cellular gene expression profiles of healthy individuals with those of patients on a genome-wide level, specific genes that may be misregulated, and therefore contribute to the disease, may be identified. In addition, specific alterations in cellular gene expression patterns can be used to diagnose a particular disease. In addition to its application in genome-wide expression profiling, DNA microarray technology is currently developed to identify genetic variations (e.g. SNPs) on a genome-wide scale.

Once candidate disease genes have been identified, their protein products need to be placed within the specific cellular pathways in which they exert their function. The human genome sequence allows for the first time a comprehensive analysis of cellular protein networks, hitherto accessible only through classical biochemical and genetic studies. Information on known genetic networks and cellular pathways in well-characterised model organisms can be used as a starting point to identify protein orthologues in humans. Searching the human genome sequence database may in addition reveal the existence of protein paralogues that may function in identical or related cellular pathways. Furthermore, the presence of common DNA regulatory elements and common expression behaviour under a range of conditions are taken as good indicators for genes that operate in networks. In this regard, DNA microarray technology has been proven an extremely powerful tool [55, 100–102].

Two large-scale projects that make use of the Human Genome Project sequence data have been initiated to complete a catalogue of genetic changes related to cancer. The Cancer Genome Anatomy Project (CGAP) was launched in 1997 by the National Cancer Institute in the US as a collaborative network providing an up-to-date and publicly accessible annotated index of genes linked to cancer. CGAP databases are available through a series of web sites (http://cgap.nci.nih.gov/; [103]). A tumour gene index has been established that contains cancer genes arranged by tissues, stages of cancer, and specificity of gene expression. Furthermore, a database for chromosomal rearrangements related to cancer was created using more than 30,000 different patient cases with different types of tumours. The CGAP has developed into an important genetics and genomics tool and studies performed with the aid of CGAP have already started to yield important results. For example, new tumour markers, genes that are specifically expressed in tumour endothelium cancers, ovarian cancers, and glioblastomas, were identified [103, 104]. A large-scale project with a similar aim, the Cancer Genome Project, funded by the Wellcome Trust, was launched at the Sanger Centre in the UK. DNA microarray technology has been successfully employed to identify cancer genes and to decipher cancer pathways and several studies indicate that genome-wide expression profiling can be used both for cancer diagnosis and prognosis. Genome-wide profiling of gene transcription has been obtained for patients with leukaemia [105] and breast cancer [106]. In addition, three genes involved in metastasis, the most disastrous attribute of cancer, could be identified. At least one candidate gene, *Rho C*, might represent an attractive drug target to prevent the spread of cancer [107]. DNA microarray technology further allowed the characterisation of two different classes of B-cell lymphoma that are indistinguishable by tumour histology. Importantly, only one tumour type reacted to chemotherapy [108]. Finally, multidrug resistance (MDR) is a major problem associated with cancer treatment [109, 110]. Because the genes and their corresponding proteins responsible for MDR are known, DNA microarray technology might be used

to screen patient samples for appropriate cancer drugs. These results have important implications for the development of customised cancer therapies.

Another priority is the development of tumour cell-specific drugs. Most cancer drugs lack specificity and affect normal cells as well as cancer cells. This causes serious side effects and toxicities and thus limits their therapeutic value. The first example of a drug tailored for a specific type of cancer is Glivec® (Novartis Pharma AG) which has been approved for use in the United States and in Switzerland in 2001. Glivec® specifically targets chronic myeloid leukaemia (CML), a blood cancer caused when part of chromosome 9 is exchanged with chromosome 22 in a white blood cell. This rearrangement forms a new structure, the Philadelphia chromosome, and creates an active *Bcr-Abl* gene by gene fusion. The Bcr-Abl protein product is a protein kinase that affects cell growth and cell differentiation [111]. Glivec® specifically blocks the action of the Bcr-Abl kinase and therefore specifically inhibits the growth of CML cells whereas normal white blood cells are unaffected. The process of drug discovery is significantly enhanced by the use of human genome sequence data and further targeted drug therapies are within scope. Using automated systems, up to 100,000 substances per day can be screened for their ability to affect expression of a particular gene [112]. Once gene mutations responsible for cancer predisposition are identified, drugs can be designed that prevent tumour formation. For example, a combination of drugs directed to different cancer pathways has been shown to reduce tumour formation in a specific strain of mice that has increased susceptibility to colon cancer [113]. These drugs are now in clinical trials. It is anticipated that additional anti-cancer drugs will be identified within a short period of time. However, because these drugs have to be administered over long periods of time, extensive clinical trials involving a large number of individuals have to be carried out in order to guarantee their safe application [84]. For this reason, the full impact of the human genome sequence on the development of novel cancer therapies will not be seen for many years.

5.4 Drug target identification

A recent survey lists 483 drug targets employed by the pharmaceutical industry that account for virtually every drug on the market. Of these drug targets 45% are G protein-coupled cell membrane receptors and 28% are enzymes [114]. Knowledge of the entire set of human protein-coding genes increases the number of potential drug targets to the order of thousands and this prospect has led to a massive expansion of genomic research towards pharmaceutical applications. One important example for potential drug targets is the significant number of ligand-binding domains in proteins [115].

Access to the human genome sequence draft has already led to the proposal of a number of new candidate drug targets. The authors of the human genome sequence draft discussed potential drug targets for molecular pathways important in schizophrenia and in mood disorders [10]. They further reported the discovery of a new receptor molecule that may constitute a promising target for the development of drugs against asthma. Finally, the initial search of the human genome sequence for paralogues of classic drug targets has led to the identification of 16 novel drug target candidates. These include paralogue genes coding for receptors of neurotransmitters and of insulin-like growth factors [10].

Once proteins important for a particular biochemical pathway are identified, demonstrating that affecting their function has therapeutic utility is the pivotal point [50]. This process has been time-consuming and expensive in the past. Post-genomic technologies such as genome-wide gene expression profiling in cultured human cells promise considerable improvements in cost-effectiveness and accuracy of drug target validation. The application of functional genomics

in pharmaceutics will significantly facilitate the identification of appropriate candidate drugs, thereby reducing failures in the clinical phase of drug development [50].

5.5 Pharmacogenetics

Pharmacogenetics investigates how genetic variations in patients affect the therapeutic value of a particular drug [50, 116]. Adverse reactions of patients to a particular drug have already been correlated with amino acid variations in drug-metabolising enzymes, such as plasma cholinesterase and glucose 6-phosphate dehydrogenase, more than 50 years ago [117]. Since then, SNPs in more than 30 drug-metabolising enzymes as well as in a number of drug transporters have been linked to compromised levels of drug efficacy or drug safety. This information is already been used to prevent drug toxicities by screening patients for specific SNPs prior to drug treatment [116].

Currently, only a small percentage of drug toxicities can be explained with genetic variation in specific genes. A systematic research into the genetic basis of adverse drug reaction has been hampered by the fact that severe reactions of patients are rare and difficult to trace. With the availability of the human genome sequence, a genomic approach to pharmacogenetics is the method of choice to establish a comprehensive catalogue of gene products involved in the binding, metabolism, or transport of specific drugs [116]. DNA samples from patients with poor or adverse effects to a specific drug can be compared to samples from patients that respond well to the treatment using as markers the set of SNPs available in the human genome database. Complementary to this approach, the human sequence database can be used to identify paralogues of genes encoding known regulators of drug kinetics or drug dynamics. Finally, human genome sequence information may also be used to conduct pharmacogenetic research retrospectively [118]. Comparative genome-wide analysis of stored patient DNA samples may be undertaken after the completion of clinical trials and even after a drug has been introduced into the market. The ultimate vision is to prevent drug prescription by trial and error and to match appropriate therapies to the specific constitution of individual patients.

5.6 Gene therapy

Gene therapy involves the targeted delivery of genes in order to replace or to compensate for malfunctioning genes responsible for genetic diseases. Genetic disorders that include enzyme deficiencies such as cystic fibrosis require long-term and regulated expression of the transgene. Treatment of other genetic diseases such as cancer may require the delivery and expression of a transgene only during a short-term period, e.g. to induce cell death. For clinical applications that require only a short-term presence of the therapeutic gene product, the possibility of protein transduction, the delivery of the gene product instead of the gene, has been raised recently [119].

The first human genetic engineering project was initiated in 1989 in the US. Tumour-infiltrating lymphocytes that contained certain marked genes were transferred into patients with advanced cancer. These experiments had two major objectives, to demonstrate safe delivery of a transgene into patients and to demonstrate its presence in patient cells [120]. Subsequently, additional clinical trials were initiated that addressed different genetic disorders such as malignant melanomas, neuroblastomas, haemophilia B, and cystic fibrosis. Cystic fibrosis (CF) has long been seen as the most promising candidate for human gene therapy [72]. First, the *CFTR* gene mutated in CF patients has been intensively characterised. Second, a successful gene transfer of the *CFTR* gene into cultured cells was demonstrated. Finally, gene therapy of CF appeared to be particularly

feasible since the affected lung tissue is readily accessible through the airways, allowing the use of aerosols for targeted gene delivery.

Most gene delivery systems rely on modified viruses that release their genome containing the transgene upon cell infection. Different virus types have been used in human therapy trials [121]. Adenovirus, a naturally occurring pathogen that causes mild infections in human airways and eyes, is an attractive vector system for short-term gene delivery [122]. The biology of adenovirus is well understood and foreign genes can be readily inserted into the virus genome. In addition, adenovirus infects both dividing and non-dividing cells and adenovirus gene expression does not require integration into the host cell genome. Finally, adenovirus infection does not pose a cancer risk as opposed to other viruses that are known to induce tumours. Inflammatory reactions to viral gene products can be circumvented by using modified viruses lacking the corresponding genes. These 'stealth viruses' are not readily detected by the immune system and therefore have increased chances to deliver the therapeutic gene to the target tissue.

For long-term gene delivery retroviruses, which integrate their genetic material into the host cell's genome, are the preferred vectors. For example, a Moloney retrovirus vector was successfully used in gene therapy of severe combined immunodeficiency-X1 (SCID-X1), a lethal immune disorder [123]. As an alternative to viral vectors, non-viral delivery systems that make use of liposomes or protein-DNA complexes are being explored [124].

Over the last decade, more than 4000 patients have been enrolled in gene transfer experiments, but only a few unambiguous results have been produced [121]. Documented cases for successful gene therapy include the transfer of the gene for the blood coagulation factor IX into three patients with haemophilia B in the US in March 2000 [125] and the treatment of two children with SCID-X1 in France in April 2000 [123]. Tragically, a young volunteer, Jesse Gelsinger, died in September 1999 from a massive immune response to adenovirus during a gene therapy trial in the US. His death led to a temporary halt of gene therapy trials and spurred a thorough investigation into the safety of viral gene delivery systems. Thus, while the recently acquired knowledge of our entire library of genetic information significantly increases the prospects for a cure of inherited disorders, the general applicability of gene therapy remains to be demonstrated.

6 Concluding remarks

The human genome sequence draft is a remarkable achievement that has provided important novel insights into the structure and function of our genome. However, as pointed out by many during the course of the human sequencing project: 'the sequence is just the beginning'. Over the next few years, we will certainly witness a number of important developments based on the published human genome sequence drafts.

Completion of the human genome sequence projected for 2003, together with improved analyses of genomic data, will help to finally answer the crucial question of the exact number of genes and proteins and will provide the basis for a comprehensive mapping of all genetic variations such as SNPs. This information will ultimately allow a complete elucidation of cellular pathways, both on the genomic and on the proteomic level, and will help to identify all genes involved in human disease. Functional genomics will provide valuable information regarding the cell type-specific expression and function of specific genes and place them into complex regulatory networks. Medical and pharmaceutical sciences will benefit greatly from novel insights into the molecular basis of disease. Additional genome sequences of a variety of species will become available and will add to our understanding of evolutionary forces and mechanisms. The comparison of human and chimpanzee genome sequences is of particular interest; it may reveal if and to what extend

specific human features such as conscious thinking and speech are manifested in the genome. The daunting task of biology in the post-genomic era is to understand how genes orchestrate and maintain life both in single cells and in complicated organisms. As eloquently pointed out by Craig Venter and colleagues: 'the real challenge will lie ahead as we seek to explain how our minds have come to organise thoughts sufficiently well to investigate our own existence' [25].

References

[1] Mendel, G., Experiments in plant hybridization. *Verh. Naturforsch. Ver. Brünn*, **4**, pp. 3–47, 1865.

[2] Avery, O.T., MacLeod, M.C. & McCarthy, M., Studies on the chemical nature of the substance inducing transformation of pneumococcal types. *J. Exp. Med.*, **98**, pp. 451–460, 1944.

[3] Watson, J.D. & Crick, F.H., A structure for DNA. *Nature*, **171**, pp. 737–738, 1953.

[4] Saenger, W., *Principles of Nucleic Acid Structure*, Springer Verlag: New York, 1984.

[5] Lewin, B., *Genes*, Oxford University Press: Oxford, 1997.

[6] Alberts, B. *et al.*, *Molecular Biology of the Cell*, Garland Publishing Inc.: New York & London, 1994.

[7] Nirenberg, M. & Leder, P., RNA codewords and protein synthesis. *Science*, **145**, pp. 1399–1407, 1964.

[8] Berget, S.M., Morre, C.S. & Sharp, P., Spliced segments at the 5′ terminus of adenovirus 2 late mRNA. *Proc. Nat. Acad. Sci. USA*, **74**, pp. 3171–3175, 1977.

[9] Chow, L.T., Gelinas, R.E., Broker, T.R. & Roberts, R.J., An amazing sequence arrangement at the 5′ end of adenovirus 2 mRNA. *Cell*, **12**, pp. 1–8, 1977.

[10] International Human Genome Sequencing Consortium, Initial sequencing and analysis of the human genome. *Nature*, **409**, pp. 860–921, 2001.

[11] Richmond, T.J., Finch, J.T., Rushton, B., Rhodes, D. & Klug, A., Structure of the nucleosome core particle at 7 Å resolution. *Nature*, **311**, pp. 532–537, 1984.

[12] Li, W.-H., Gu, Z., Wang, H. & Nekrutenko, A., Evolutionary analyses of the human genome. *Nature*, **409**, pp. 847–849, 2001.

[13] Dib, C. *et al.*, A comprehensive genetic map of the human genome based on 5,264 microsatellites. *Nature*, **380**, pp. 152–154, 1996.

[14] Brett, D. *et al.*, EST comparison indicates 38% of human mRNAs contain possible alternative splice forms. *FEBS Lett.*, **474**, pp. 83–86, 2000.

[15] Galas, D.J., Making Sense of the Sequence. *Science*, **291**, pp. 1257–1260, 2001.

[16] Hershko, A. & Ciechanover, A., The ubiquitin system. *Annu. Rev. Biochem.*, **67**, pp. 425–479, 1998.

[17] Hochstrasser, M., Evolution and function of ubiquitin-like protein-conjugation systems. *Nat. Cell Biol.*, **2**, pp. E153–157, 2000.

[18] Roberts, L., Controversial from the start. *Science*, **291**, pp. 1182–1188, 2001.

[19] Gyapay, G. *et al.*, The 1993–94 Genethon human genetic linkage map. *Nat. Genet.*, **7**, pp. 246–339, 1994.

[20] Hudson, T.J. *et al.*, An STS-based map of the human genome. *Science*, **270**, pp. 1945–1954, 1995.

[21] Fleischmann, R.D. *et al.*, Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*, **269**, pp. 496–512, 1995.

[22] Mewes, H.W. *et al.*, Overview of the yeast genome. *Nature*, **387(suppl.)**, pp. 7–65, 1997.

[23] The *C. elegans* Sequence Consortium, Genome sequence of the nematode *C. elegans*: a platform for investigating biology. *Science*, **282**, pp. 2012–2018, 1998.
[24] Marshall, E., Sharing the glory, not the credit. *Science*, **291**, pp. 1189–1193, 2001.
[25] Venter, J.C. *et al.*, The sequence of the human genome. *Science*, **291**, pp. 1304–1351, 2001.
[26] Sanger, F., Nicklen, S. & Coulson, A.R., DNA sequencing with chain-terminating inhibitors. *Proc. Nat. Acad. Sci. USA*, **74**, pp. 5463–5467, 1977.
[27] Maxam, A. & Gilbert, W., A new method for sequencing DNA. *Proc. Nat. Acad. Sci. USA*, **74**, pp. 560–564, 1977.
[28] Roberts, L., A history of the Human Genome Project. *Science*, **291**, pp. 1195–1200, 2001.
[29] Edwards, A. *et al.*, Automated DNA sequencing of the human HPRT locus. *Genomics*, **6**, pp. 593–608, 1990.
[30] Smit, A.F.A., Interspersed repeats and other mementos of transposable elements in mammalian genomes. *Curr. Opin. Genet. Dev.*, **9**, pp. 657–663, 1999.
[31] Smit, A.F.A., The origin of interspersed repeats in the human genome. *Curr. Opin. Genet. Dev.*, **6**, pp. 743–748, 1996.
[32] Gardiner, K., Base composition and gene distribution: critical pattern in mammalian genome organisation. *Trends Genet.*, **12**, pp. 519–524, 1996.
[33] Agrawal, A., Eastman, Q.M. & Schatz, D.G., Transposition mediated by RAG1 and RAG2 and its implications for the evolution of the immune system. *Nature*, **394**, pp. 744–751, 1998.
[34] Malik, H.S., Burke, W.D. & Eickbush, H., The age and evolution of non-LTR retrotransposable elements. *Mol. Biol. Evol.*, **16**, pp. 793–805, 1999.
[35] Csink, A.K. & Henikoff, S., Something from nothing: the evolution and utility of satellite repeats. *Trends Genet.*, **14**, pp. 200–204, 1998.
[36] Kazazian Jr, H.H. & Moran, J.V., The impact of L1 retrotransposons on the human genome. *Nat. Genet.*, **19**, pp. 19–24, 1998.
[37] Ewing, B. & Green, P., Analysis of expressed sequence tags indicates 35,000 human genes. *Nat. Genet.*, **25**, pp. 232–234, 2000.
[38] Pruitt, K.D. & Maglott, D.R., RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, pp. 137–140, 2001.
[39] Burge, C. & Karlin, S., Prediction of complex gene structures in human genomic DNA. *J. Mol. Biol.*, **268**, pp. 78–94, 1997.
[40] Kulp, D., Haussler, D., Reese, M.G. & Eckmann, F.H., A generalized hidden Markov model for the recognition of human genes in DNA. *ISMB*, **4**, pp. 134–142, 1996.
[41] Guigo, R., Agrawal, P., Abril, J.F., Burset, M. & Fickett, J.W., An assessment of gene prediction accuracy in large DNA sequences. *Genome Res.*, **10**, pp. 1631–1642, 2000.
[42] Fields, C., Adams, M.D., White, O. & Venter, J.C., How many genes in the human genome? *Nat. Genet.*, **7**, pp. 345–346, 1994.
[43] Baltimore, D., Our genome unveiled. *Nature*, **409**, pp. 814–816, 2001.
[44] Pääbo, S., The human genome and our view of ourselves. *Science*, **291**, pp. 1219–1220, 2001.
[45] Bird, A., CpG islands as gene markers in the vertebrate nucleus. *Trends Genet.*, **3**, pp. 342–347, 1987.
[46] Mann, J.R., Szabo, P.E., Reed, M.R. & Singer-Sam, J., Methylated DNA sequences in genomic imprinting. *Crit. Rev. Eukaryot. Gene Expr.*, **10**, pp. 241–247, 2000.

[47] The International SNP Map Working Group, A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature*, **409**, pp. 928–933, 2001.
[48] Brookes, A.J., The essence of SNPs. *Gene*, **234**, pp. 177–186, 1999.
[49] Lander, E.S., The new genomics: global views of biology. *Science*, **274**, pp. 536–539, 1996.
[50] Roses, A.D., Pharmacogenetics and the practice of medicine. *Nature*, **405**, pp. 857–865, 2000.
[51] Roses, A.D., Apoliprotein E affects the rate of Alzheimer disease expression: beta-amyloid burden is a secondary consequence dependent on APOE genotype and duration of the disease. *J. Neuropathol. Exp. Neurol.*, **53**, pp. 429–437, 1994.
[52] Stoneking, M., From the evolutionary past ... . *Nature*, **409**, pp. 821–822, 2001.
[53] Rubin, G.M., Comparing species. *Nature*, **409**, pp. 820–821, 2001.
[54] Tupler, R., Perini, G. & Green, M., Expressing the human genome. *Nature*, **409**, pp. 832–833, 2001.
[55] Peltonen, L. & McKusick, V.A., Dissecting human disease in the postgenomic era. *Science*, **291**, pp. 1224–1229, 2001.
[56] Stoll, M. *et al.*, New target regions for human hypertension via comparative genomics. *Genome Res.*, **10**, pp. 473–482, 2000.
[57] Lusis, A.J., Atherosclerosis. *Nature*, **407**, pp. 233–241, 2000.
[58] Kreutz, R. *et al.*, Dissection of a quantitative trait locus for genetic hypertension on rat chromosome 10. *Proc. Nat. Acad. Sci. USA*, **92**, pp. 8778–8782, 1995.
[59] Nolan, P.M. *et al.*, A systematic, genome-wide, phenotype-driven mutagenesis programme for gene function studies in the mouse. *Nat. Genet.*, **25**, pp. 440–443, 2000.
[60] Hrabe de Angelis, M. *et al.*, Genome-wide, large-scale production of mutant mice by ENU mutagenesis. *Nat. Genet.*, **25**, pp. 444–447, 2000.
[61] Pennacchio, L.A. & Rubin, E.M., Genomic strategies to identify mammalian regulatory sequences. *Nat. Rev. Genet.*, **2**, pp. 100–109, 2001.
[62] Chu, S. *et al.*, The transcriptional program of sporulation in budding yeast. *Science*, **282**, pp. 699–705, 1998.
[63] Tavazoie, S., Hughes, J.D., Campbell, M.J., Cho, R.J. & Church, M.J., Systematic determination of genetic network architecture. *Nat. Genet.*, **22**, pp. 281–285, 1999.
[64] Wassermann, W.W., Palumbo, M., Thompson, W., Fickett, J.W. & Lawrence, C.E., Human-mouse genome comparisons to locate regulatory sites. *Nat. Genet.*, **26**, pp. 225–228, 2000.
[65] Pandey, A. & Mann, M., Proteomics to study genes and genomes. *Nature*, **405**, pp. 837–846, 2000.
[66] O'Donovan, C., Apweiler, R. & Bairoch, A., The human proteomics initiative (HPI). *Trends Biotechnol.*, **19**, pp. 178–181, 2001.
[67] Russell, R.B. & Eggleston, D.S., New roles for structure in biology and drug discovery. *Nat. Struct. Biol.*, **7**, pp. 928–930, 2000.
[68] Yokoyama, S. *et al.*, Structural genomics in Japan. *Nat. Struct. Biol.*, **7**, pp. 943–945, 2000.
[69] Terwillinger, T.C., Structural genomics in North America. *Nat. Struct. Biol.*, **7**, pp. 935–939, 2000.
[70] Heinemann, U., Structural genomics in Europe: slow start, strong finish? *Nat. Struct. Biol.*, **7**, pp. 940–942, 2000.

[71] Gitschier, J. *et al.*, Characterization of the human factor VIII gene. *Nature*, **312**, pp. 326–330, 1984.
[72] Collins, F.S., Cystic fibrosis: molecular biology and therapeutic implications. *Science*, **256**, pp. 774–779, 1992.
[73] Rommens, J.M. *et al.*, Identification of the cystic fibrosis gene: chromosome walking and jumping. *Science*, **245**, pp. 1059–1065, 1989.
[74] Riordan, J.R. *et al.*, Identification of the cystic fibrosis gene: cloning and characterization of complimentary DNA. *Science*, **245**, pp. 1066–1073, 1989.
[75] Kerem, B. *et al.*, Identification of the cystic fibrosis gene: genetic analysis. *Science*, **245**, pp. 1073–1080, 1989.
[76] Tobin, A.J. & Signer, E.R., Huntington's disease: the challenge for cell biologists. *Trends Cell Biol.*, **10**, pp. 531–536, 2000.
[77] Lifton, R.P., Gharavi, A.G. & Geller, D.S., Molecular mechanisms of human hypertension. *Cell*, **104**, pp. 545–556, 2001.
[78] Seidman, J.G. & Seidman, C., The genetic basis for cardiomyopathy: from mutation identification to mechanistic paradigms. *Cell*, **104**, pp. 557–567, 2001.
[79] Keating, M. & Sanguinetti, M.C., Molecular and cellular mechanisms of cardiac arrhythmia. *Cell*, **104**, pp. 569–580, 2001.
[80] Futreal, P.A. *et al.*, Cancer and genomics. *Nature*, **409**, pp. 850–852, 2001.
[81] Hanahan, D. & Weinberg, R.A., The hallmarks of cancer. *Cell*, **100**, pp. 57–70, 2000.
[82] Renan, M.J., How many mutations are required for tumorigenesis? Implications from human cancer data. *Mol. Carcinogenesis*, **7**, pp. 139–146, 1993.
[83] Fearon, E.R., Human cancer syndromes: clues to the origin and nature of cancer. *Science*, **278**, pp. 1043–1050, 1997.
[84] Ponder, B.A.J., Cancer genetics. *Nature*, **411**, pp. 336–341, 2001.
[85] Nathanson, K.N., Wooster, R. & Weber, B.L., Breast cancer genetics: what we know and what we need. *Nat. Med.*, **7**, pp. 552–556, 2001.
[86] Liotta, L. & Petricoin, E., Molecular profiling of human cancer. *Nat. Rev. Genet.*, **1**, pp. 48–56, 2000.
[87] Parada, L.P., Tabin, C.J., Shih, C. & Weinberg, R.A., Human EJ bladder carcinoma oncogene is a homologue of Harvey Sarcoma virus ras gene. *Nature*, **297**, pp. 474–477, 1982.
[88] Bishop, J.M., Enemies within: the genesis of retrovirus oncogenes. *Cell*, **23**, pp. 5–6, 1981.
[89] Weinberg, R.A., Tumor suppressor genes. *Science*, **254**, pp. 1138–1145, 1991.
[90] Knudson, A.G., Mutation and cancer: statistical study of retinoblastoma. *Proc. Nat. Acad. Sci. USA*, **68**, pp. 820–823, 1971.
[91] Cavenee, W.K. *et al.*, Expression of recessive alleles by chromosomal mechanisms in retinoblastoma. *Nature*, **305**, pp. 779–784, 1983.
[92] Kinzler, K.W. & Vogelstein, B., Gatekeepers and caretakers. *Nature*, **386**, pp. 761–763, 1997.
[93] Surani, M.A., Imprinting and the initiation of gene silencing in the germ line. *Cell*, **93**, pp. 309–312, 1998.
[94] Baylin, S.B. & Herman, J.G., DNA hypermethylation in tumorigenesis. *Trends Genet.*, **16**, pp. 168–174, 2000.
[95] Kohl, S. *et al.*, Mutations in the CNGB3 gene encoding the beta-subunit of the cone photoreceptor cGMP-gated channel are responsible for achromatopsia (ACHM3) linked to chromosome 8q21. *Hum. Mol. Genet.*, **9**, pp. 2107–2116, 2000.

[96] Sundin, O.H. *et al.*, Genetic basis of total colourblindness among the Pingelapese islanders. *Nat. Genet.*, **25**, pp. 289–293, 2000.
[97] Sherrington, R. *et al.*, Cloning of a gene bearing missense mutations in early-onset familial Alzheimer's disease. *Nature*, **375**, pp. 754–760, 1995.
[98] Olivieri, N.F. & Wheatherall, D.J., The therapeutic reactivation of fetal haemoglobin. *Hum. Mol. Genet.*, **7**, pp. 1655–1658, 1998.
[99] Lockhart, D.J. & Winzeler, E.A., Genomics, gene expression and DNA arrays. *Nature*, **405**, pp. 827–836, 2000.
[100] Young, R., Biomedical discovery with DNA arrays. *Cell*, **102**, pp. 9–15, 2000.
[101] Rubin, E.M. & Tall, A., Perspectives for vascular genomics. *Nature*, **407**, pp. 265–269, 2000.
[102] Friddle, C.J., Koga, T., Rubin, E.M. & Bristow, J., Expression profiling reveals distinct sets of genes altered during induction and regression of cardiac hypertrophy. *Proc. Nat. Acad. Sci. USA*, **97**, pp. 6745–6750, 2000.
[103] Strausberg, R.L., Buetow, K.H., Emmert-Buck, M.R. & Klausner, R.D., The Cancer Genome Anatomy Project. *Trends Genet.*, **16**, pp. 103–106, 2000.
[104] Riggins, G.J. & Strausberg, R.L., Genome and genetic resources from the Cancer Genome Anatomy Project. *Hum. Mol. Genet.*, **10**, pp. 663–667, 2001.
[105] Golub, T.R. *et al.*, Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, **286**, pp. 531–537, 1999.
[106] Perou, C.M. *et al.*, Molecular portraits of human breast cancer tumours. *Nature,* **406**, pp. 747–752, 2000.
[107] Clark, E.A., Golub, T.R., Lander, E.S. & Hynes, R.O., Genomic analysis of metastasis reveals an essential role for Rho C. *Nature*, **406**, pp. 532–535, 2000.
[108] Alizadeh, A.A. *et al.*, Distinct types of diffuse B-cell lymphoma identified by gene expression profiling. *Nature*, **403**, pp. 503–511, 2000.
[109] Liscovitch, M. & Lavie, Y., Multidrug resistance: a role for cholesterol efflux pathways? *Trends Biochem. Sci.*, **25**, pp. 530–534, 2000.
[110] Szabo, D., Keyzer, H., Kaiser, H.E. & Molnar, J., Reversal of multidrug resistance of tumour cells. *Anticancer Res.*, **20**, pp. 4261–4274, 2000.
[111] Thijsen, S., Schuurhuis, G., van Oostveen, J. & Ossenkoppele, G., Chronic myeloid leukemia from basics to bedside. *Leukemia*, **13**, pp. 1646–1674, 1999.
[112] Marchant, J., Know your enemy. *New Scientist*, pp. 46–50, 2000.
[113] Torrance, C.J. *et al.*, Combinatorial chemoprevention of intestinal neoplasia. *Nat. Med.*, **6**, pp. 1024–1028, 2000.
[114] Drews, J., Drug discovery: a historical perspective. *Science*, **287**, pp. 1960–1964, 2000.
[115] Bailey, D., Zanders, E. & Dean, P., The end of the beginning for genomic medicine. *Nat. Biotechnol.*, **19**, pp. 207–211, 2001.
[116] Rothberg, B.E.G., Mapping a role for SNPs in drug development. *Nat. Biotechnol.*, **19**, pp. 209–211, 2001.
[117] Evans, W.E. & Relling, M.V., Pharmacogenomics: translating functional genomics into rational therapeutics. *Science*, **286**, pp. 487–491, 1999.
[118] McCarthy, A., Pharmacogenetics. *BMJ*, **322**, pp. 1007–1008, 2001.
[119] Ford, K.G., Souberbielle, B.E., Darling, D. & Farzaneh, F., Protein transduction: an alternative to genetic intervention? *Gene Therapy*, **8**, pp. 1–4, 2001.
[120] Anderson, W.F., Human gene therapy. *Science*, **256**, pp. 808–813, 1992.
[121] Marshall, E., Gene therapy on trial. *Science*, **288**, pp. 951–957, 2000.

[122] Benihoud, K., Yeth, P. & Perricaudet, M., Adenovirus vectors for gene delivery. *Curr. Opin. Biotechnol.*, **10**, pp. 440–447, 1999.
[123] Cavazzana-Calvo, M. *et al.*, Gene therapy of human severe combined immunodeficiency (SCID)-X1 disease. *Science*, **288**, pp. 669–672, 2000.
[124] Cristiano, R.J. *et al.*, Viral and nonviral gene delivery vectors for cancer gene therapy. *Cancer Detect. Prev.*, **22**, pp. 445–454, 1998.
[125] Kay, M.A. *et al.*, Evidence for gene transfer and expression of factor IX in haemophilia B patients treated with an AAV vector. *Nat. Genet.*, **24**, pp. 257–261, 2000.

Chapter 5

The laws of thermodynamics: entropy, free energy, information and complexity

M.W. Collins[1], J.A. Stasiek[2] & J. Mikielewicz[3]
[1]*School of Engineering and Design, Brunel University, Uxbridge, Middlesex, UK.*
[2]*Faculty of Mechanical Engineering, Gdansk University of Technology, Narutowicza, Gdansk, Poland.*
[3]*The Szewalski Institute of Fluid–Flow Machinery, Polish Academy of Sciences, Fiszera, Gdansk, Poland.*

Abstract

The laws of thermodynamics have a universality of relevance; they encompass widely diverse fields of study that include biology. Moreover the concept of information-based entropy connects energy with complexity. The latter is of considerable current interest in science in general. In the companion chapter in Volume 1 of this series the laws of thermodynamics are introduced, and applied to parallel considerations of energy in engineering and biology. Here the second law and entropy are addressed more fully, focusing on the above issues. The thermodynamic property free energy/exergy is fully explained in the context of examples in science, engineering and biology. Free energy, expressing the amount of energy which is usefully available to an organism, is seen to be a key concept in biology. It appears throughout the chapter. A careful study is also made of the information-oriented 'Shannon entropy' concept. It is seen that Shannon information may be more correctly interpreted as 'complexity' rather than 'entropy'. We find that Darwinian evolution is now being viewed as part of a general thermodynamics-based cosmic process. The history of the universe since the Big Bang, the evolution of the biosphere in general and of biological species in particular are all subject to the operation of the second law of thermodynamics. Our conclusion is that, in contrast to the rather poor 19th century relationship between thermodynamics and biology, a mainstream reconciliation of the two disciplines is now emerging.

1 Introduction

1.1 General

The Industrial Revolution in Britain at the end of the 18th century stemmed from the invention of the steam engine by Watt. The theory of this engine was elaborated half a century later by the

French scientist Sadi Carnot who was the first to formulate the second law of thermodynamics. From then on thermodynamics became a new philosophy in physics, developing over a parallel timescale to that of evolutionary biology.

In the science of the Renaissance *time* did not feature in the description of phenomena, the laws of physics having a reversible character. Notably Galileo and Newton did not consider the direction of transformations: their mechanistic world was governed by simple reversible principles. Thermodynamics was to change this. While physics at the time of Newton was concerned with optics, mechanics and electrostatics, at the end of the 18th and through the 19th centuries (i.e. because of the Industrial Revolution) the theory of heat became equally important. Relationships between thermal and mechanical forms of energy began to attract attention, which stimulated the development of thermodynamics. Scientists soon came to the conclusion that while mechanical energy can easily be converted into heat, thermal energy cannot be wholly converted into mechanical energy. Carnot (1824) dealing with the maximisation of conversion of thermal into mechanical energy found that there is a limit to the process of energy conversion. This limit was eventually expressed by Clausius's formulation of the second law of thermodynamics. The second law provides a constraint on the use of energy, in that although the total energy does not change, its ability to generate work depends on a special feature of the energy. This feature was termed 'entropy' by Clausius and defined as the quotient of 'heat' and 'temperature on the absolute scale'. Further, Clausius postulated that the entropy of any isolated system would as a general rule increase and entropy began to be interpreted as disorder. Finally, the realisation that this also would hold for the universe gave a *direction* of change for the universe. The second law introduced the importance of time into science [1].

Unlike, say, energy entropy is not an immediately understandable scientific variable. This is partly because it has a number of different aspects. There is the original phenomenological or thermal meaning, then the microscopic or statistical meaning from the work of Boltzmann, and most recently the information meaning, related to Shannon's communication theory. Many scientists think that these three formulations of entropy are equivalent, although the third form does require rather careful understanding. This opens new possibilities for the application of thermodynamics to such diverse fields as medicine, history and social processes [2].

Classical physics and chemistry deal with *isolated* and *closed systems.* Recently, however, science and engineering have become more interested in *open systems* [1,2]. Apart from engineering applications, open systems are common in biology, medicine, economics, sociology and history. With closed systems their entropy can increase and the systems finally reach a state of equilibrium. Open systems, on the other hand, and living organisms in particular, can exist in a far-from-equilibrium steady state under conditions of a continuous import of matter and energy from the surroundings [3, 4]. At this point we stress that these latter are accepted characteristics of living systems, stemming from the description given by Schrödinger in his What is Life? lectures of 1944. 'The living organism ... keeps alive ... by continually drawing from its environment negative entropy' ([5], p. 71; [6], p. 88). The Nobel prize-winning thermodynamicist Ilya Prigogine described [5] as 'a beautiful book' ([7] p. 242), and Schrödinger's exposition not only inspired him but also the mathematician Roger Penrose ([6], p. xx). Prigogine uses the expression 'a flow of free energy carried by matter ... which "feeds" the living system' ([7], p. 239), while Penrose goes right back to the 'Sun's role as a source of low entropy (or 'negative entropy', in Schrödinger's terminology)' ([6], p. xx). This thermodynamic necessity for the existence of living organisms carries forward into current mainstream biology. In the standard US text *Life*, W. Purves *et al.* [8] explain: 'Free energy is what cells require for ... cell growth, cell division and the maintenance of cell health' (([8], p. 97) and context).

A final, more general point, is that systems far-from-equilibrium can undergo evolution to a new level of order. This has application to the behaviour of complex systems.

In this chapter, thermodynamics is applied to examples in physics (magnetism), in engineering (exergy analysis of a nuclear power system) and to biology (glycolysis in cells). While the first two examples are at the level of understanding of the latter years of undergraduate courses, there is sufficient background in [9] and in this chapter to follow these through. The identity of thermal and statistical entropy is then demonstrated and the role of entropy in contemporary studies surveyed. The validity of 'Shannon entropy' is carefully addressed together with the quantitative definitions of information and complexity. The core of the whole biological/thermodynamic synthesis is dealt with, for the cosmos, for the biosphere and then for species.

Overall, we have sought to elucidate our theme from a variety of angles. We conclude with a discussion of some of the consequences of our overall approach, and by stressing that a mainstream reconciliation, if not orthodoxy, of thermodynamics and biology is clearly developing.

1.2 Closed, open and isolated systems

> Thermodynamics is the only science that deals with *inherent* directionality.
>
> Jeffrey Wicken [10], p. 65

> . . . he became fascinated by the apparent contradiction. . . suggested by the second law of thermodynamics and Darwinian evolution. . . to more complex and ordered structures.
>
> Obituary 'Viscomte Ilya Prigogine', *Daily Telegraph*, 5 June 2003

In this chapter, we consider three kinds of systems. A *closed* system is of fixed mass, with possible heat and work transfers across its boundary. An *open* system has, in addition to possible heat and work, mass flows across its boundary. An *isolated* system is a closed system with no heat transfer.

Classical physics and chemistry deal with closed systems in a state of equilibrium. A system is in equilibrium if its parameters are uniform in space. Equilibrium processes are ideal processes which do not occur in nature. When a system changes from one state to another, its equilibrium is lost. In practice though, it is sufficient that the time needed for a state to reach equilibrium is smaller than the time of change from one state to another. In classical terms, the relaxation time is shorter than the duration of the phenomenon. This is a key underlying assumption of 'equilibrium thermodynamics' which, while idealised, is a satisfactory description of a whole range of engineering processes.

An isolated system, initially in a non-equilibrium state, will always tend towards equilibrium. Such a process is called *spontaneous*. Now reversal of a spontaneous process is impossible. This is the most general formulation of the second law of thermodynamics and expresses Wicken's 'inherent directionality' of change in natural processes. It is written in mathematical form using the entropy S, where changes for an isolated system are always non-decreasing:

$$\mathrm{d}S_{\mathrm{isol}} \geq 0. \tag{1}$$

As we have noted, the concept of entropy was formulated by Clausius in 1850. It is defined in the following way:

$$\mathrm{d}S^{\mathrm{def}} = \frac{\mathrm{d}Q^{\mathrm{o}}}{T}, \tag{2}$$

where dQ^o, represents the total heat transfer to or from the system and T, is the absolute temperature of the system.

In Section 1.1 we pointed out that open systems are of increasing interest in science and engineering and are also found in biology, medicine, economics, sociology and history. It is the essential differences between closed/isolated and open systems which removes the 'apparent contradiction' which fascinated Prigogine. Living organisms are open systems.

In an open system, as formulated by Prigogine:

$$dS = d_eS + d_iS, \tag{3}$$

where d_eS denotes the change of entropy due to the influx of mass from the surroundings, d_iS is the entropy rise due to irreversibility of changes taking place in the system; d_eS is always positive whereas d_iS can either be positive or negative. dS, then, can be negative, i.e. the entropy of the system *reduces*, such a process being frequently referred to as 'negentropic'. *In living organisms this 'negentropic' effect is not inconsistent with the second law of thermodynamics.*

1.3 Complex systems

The historical development of thermodynamics focused on the production of mechanical work. This resulted from a change of shape (expansion) of a system, usually comprising a fluid substance such as steam/water or gas. However, thermodynamics is much more generally applicable, and can involve systems such as magnetic substances, surface membranes and elastic (solid) bodies. Correspondingly, thermodynamic work is widely defined to include that caused, for example, by electrical, rotational, chemical and surface tension forces. Systems which can perform a number of different types of work are formally defined in the context of thermodynamics as 'complex systems', involving what are referred to as 'complex substances'. However, these expressions should not be confused with the rather similar expression 'complex structures' as used to describe biological systems. The latter really arises from more general complexity studies, not thermodynamics. Finally, our use of the word *complexity* will refer to the quantifiable Shannon information measure of structured systems in general.

2 Application of classical thermodynamics to physics

2.1 The calculation of mechanical work

One of the first applications of the closed system concept in thermodynamics is its key application to mechanical work. Such work is done when a system increases its volume (see equation (1) in [9]), so that:

$$dW = p dV.$$

Most frequently this involves a compressible substance, such as a gas. For the equation to be integrated, it is essential to know the relationship between p and V. In fact, for a gas, the absolute temperature T is also involved. To calculate the work for a given process, the relationship may be expressed as:

1. model equations, such as 'perfect gas' and 'real gas' equations;
2. thermodynamic graphs, such as the old engineering 'steam charts';
3. thermodynamic tables, such as the engineering 'steam tables' or 'property tables'.

In this treatment, we define the perfect gas and give an example of a real gas, the Van der Waals' gas.

2.1.1 The perfect gas

The thermal state equation for a perfect gas has the form:

$$pv = RT \quad \text{or} \quad pV = mRT, \tag{4}$$

and is sometimes called the Clapeyron equation.

The experiment of Joule concerning expansion of gases to a vacuum proved that the change of the gas volume does not change the energy of the internal gas.

$$\left(\frac{\partial u}{\partial v}\right)_T = 0. \tag{5}$$

That is, for a constant volume there is no work transfer, and the heat transfer dq by the first law of thermodynamics is equal to du.

$$\mathrm{d}q = \mathrm{d}u = \left(\frac{\partial u}{\partial T}\right)_v \mathrm{d}T = f(T)\mathrm{d}T. \tag{6}$$

It follows from the above that

$$\mathrm{d}u = c_v \mathrm{d}T, \tag{7a}$$

which means that the internal energy depends only on the temperature. The above equation is called *the caloric equation for a perfect gas.* For a compressible substance in general u is given by

$$u = u(T, v), \quad \text{and this becomes } u = u(T) \text{ for the perfect gas.} \tag{7b}$$

2.1.2 A real gas

Based on the molecular structure of matter, Van der Waals suggested a modification of the state equation for a perfect gas which takes into account the effect of molecular attraction in the pressure term and effect of molecular volume in the specific volume term:

$$\left(p + \frac{a}{v^2}\right)(v - b) = RT. \tag{8}$$

This model of a real gas is only qualitatively good, and other p, V, T relationships may be specified. As mentioned above, for engineering calculations thermodynamic graphs and tables are used (now, most commonly, the latter) near the phase-change boundaries.

What eqn (4) implies is that the thermodynamic state of a substance may be defined by the choice of two properties. This is an expression of the so-called 'two-property' rule. It is termed a rule, rather than a law, because under certain conditions (e.g. mixed-phase) some of the properties are not independent.

2.2 The simple magnetic substance

Classical equilibrium thermodynamics is relevant to the entire physical world. The above initial example demonstrates how, in order to calculate thermodynamic behaviour, the primary properties of pressure, volume and temperature are related for compressible substances like gases.

A further example, the simple magnetic substance, illustrates how thermodynamics can provide a detailed description of what may be subtle and intricate property behaviour for solids.

For such a substance, it follows from the above two-property rule that by choosing two independent intensive parameters, we may define the state. Using the absolute temperature T and the magnetic field H, and starting from similar premises as for compressible substances (say the perfect gas) the thermal state equation can be written in the form:

$$M = M(T, H), \tag{9}$$

where M is the magnetic moment. This plays the same role for magnetic substances as the volume for compressible substances. The magnetic field H, parallel to pressure, gives the mechanical work in a magnetic field as:

$$\mathrm{d}W = -\mu_0 VH\,\mathrm{d}M = BV\mathrm{d}M, \tag{10}$$

where μ_0 denotes the magnetic permeability and $B = \mu_0$ H is the magnetic induction.

The caloric equation of state that describes the internal energy can have one of the following forms:

$$\begin{aligned} u &= u(T, M), \\ u &= u(T, H), \\ u &= u(M, H), \end{aligned} \tag{11}$$

the first form being analogous to

$$u = u(T, V),$$

for compressible substances in general (equation preceding (7b)).

For magnetic systems we can also define:

- Magnetic enthalpy
$$H_M = U - \mu_0 VHM \tag{12}$$
- Gibbs function
$$G_M = H_M - T\mathrm{d}S = U - \mu_0 VHM - T\mathrm{d}S = \varphi \tag{13}$$
- Specific heat at constant magnetic field H (in analogy to c_p)
$$c_H = \left(\frac{\partial h}{\partial T}\right)_H \tag{14}$$
and the relation between c_H and c_M, where c_M is analogous to c_V (the specific heat of a gas at constant volume):
$$c_H - c_M = \frac{T\mu_0 V\,(\partial M/\partial T)_H^2}{(\partial M/\partial H)_T}. \tag{15}$$

One can also obtain Maxwell relations for magnetic systems.

The magnetic moment M is related to the magnetic field H

$$M = \chi H. \tag{16}$$

If $\chi < 0$, then the substances are called *diamagnetic*. They attenuate the magnetic field. Examples of such substances are mercury, silver, copper, or gold. If $\chi > 0$, the substances are called *paramagnetic*. For these substances $B > H$, however χ does not depend on H. Examples are Curie magnetics for which $\chi = A/T$, where A is a constant. If $\chi < 0$ and χ strongly depends on the magnetic field, then the substances are called *ferromagnetic*. Among ferromagnetic substances are iron, nickel and cobalt.

The magnetic system undergoes changes in structure (phase changes). It appears in two phases: *conductive (paramagnetic)* and *superconductive (ferromagnetic)*. Systems that change from normally conductive to superconductive states are called superconductors. The transition point is called the Curie point. Above the Curie point, the ferromagnetic substance behaves as paramagnetic. For iron, the Curie point is 765°C. For magnetic systems, the Curie magnetic can be considered analogous to a perfect gas, as its state equation is expressed by a simple analytical relation:

$$M = c\frac{H}{T} \quad \text{where } \frac{H}{T} = f(M). \tag{17}$$

Examples of Curie magnetics can be paramagnetic salts at not too low temperatures and not too high magnetic fields. It can be proved that in this case the internal energy is a function of temperature only.

The temperature does not depend on M for $\mathrm{U} = \text{const.}$, which means that $\mathrm{U} = \mathrm{U}(T)$ for substances for which $H/T = f(M)$. For the Curie magnetic the following relation holds:

$$\mathrm{d}u = c_M(T)\mathrm{d}T. \tag{18}$$

For a paramagnetic substance $c_H > c_M$, for diamagnetic $c_H - c_M \approx 0$.

Adiabatic demagnetisation of paramagnetics leads to a decrease in their temperature. This effect is especially strong at low temperatures. Using this approach, temperatures of the order of 0.001°C can be achieved.

2.3 Complex substances

Complex substances are substances which are subject to more than one type of reversible work. Examples are systems exchanging mass within the same substance and at the same time performing volume work. There are two types of such systems: those undergoing phase changes (physical reactions) and those with chemical reactions. Systems within a field of external forces, such as an electric or magnetic field and those which are at the same time subject to mechanical loads, form a class of complex substances. A number of thermal effects, important from the point of view of applications as well as interesting in terms of cognitive values, occur in these systems. *Pyroelectric* or *pyromagnetic* effects appear in electric or magnetic fields. Among these effects are electrocaloric or magnetocaloric effects that accompany adiabatic processes. In systems subject to mechanical loads, thermoelastic effects occur, whereas in electric or magnetic systems subject to mechanical loads additional effects take place resulting from coupling of the phenomena, the so-called piezoelectric or piezomagnetic effects.

Again, we are using the expression ‘complex substances’ as a whole in a manner similar to ‘complex systems’.

2.4 Discussion

In this section we have shown how, using methods of classical thermodynamics, the behaviour of typical gases may be described, and an entire mathematical model constructed for magnetic substances. This model incorporates the properties and behaviour of the different classes diamagnetic, paramagnetic and ferromagnetic. It serves to demonstrate how comprehensively thermodynamics applies in the physical world.

3 Application of laws of thermodynamics in engineering

3.1 Introduction

We turn from physics to engineering and emphasise that the laws of thermodynamics enable a range of engineering problems to be addressed.

Now classical thermodynamics does not make use of a co-ordinate system as it deals with equilibrium systems where intensive parameters (such as temperature field, pressure field and concentration field) are uniform in the whole volume.

An important element of thermodynamic analysis is the choice of the *system*. The classical system is a closed system – connected with a constant mass and same number of molecules. In engineering thermodynamics the concept of the *closed system* is extended to that of the *open system*. In the case of the closed system, the application of mass conservation is redundant. In the case of the open system, the conservation of mass and momentum provides complementary information about the behaviour of the system.

3.2 Energy and exergy analysis: the concept of maximum work

Here the level of understanding is that of the final year of an engineering degree course. We have sought to give sufficient background material, here and in [9], to enable an understanding of *exergy* to be gained, from the following.

An energy conversion chain is accomplished in power stations: from chemical energy to heat to mechanical energy to electrical energy. The efficiency of individual processes is the ratio of the exit energy of the desired type to the energy supplied at the inlet. The efficiency depends on physical and chemical laws governing the processes of conversion. Increasing the efficiency of the process can be achieved by decreasing the amount of energy supplied at the inlet, increasing the energy output at the exit, or a combination of both these methods.

In the companion chapter of the authors in Volume 1 of this Series [9] a description is given of the application of thermodynamics to the generation of power through, especially, fossil fuels. The individual components of a power station (e.g. boiler, steam turbine) are open systems using flow processes. However, for the operation as a whole the H_2O working fluid acts as a closed system, and therefore as a *heat engine*. (The latter is defined as a system operating continuously over a cycle, exchanging heat with thermal reservoirs and producing work.)

The maximum possible theoretical thermal efficiency is provided by the Carnot cycle. However, a simple more practical adaptation results in the Rankine cycle. This, together with superheat and reheat enhancements, provides a theoretical heat engine model for analysis of fossil fuel central electricity generation steam cycles (again see [9] and the typical undergraduate text by Rogers and Mayhew [11]). For the sake of completeness we define our ideal heat engine as follows.

In consistency with the second law of thermodynamics, no heat engine can have an efficiency higher than that of the reversible Carnot cycle. In turn, the first law of thermodynamics implies that the heat input to the cycle, i.e. the difference between the supplied heat and the heat rejected from the cycle, is equal to the cycle work. The efficiency of the Carnot cycle for constant temperatures of the upper source T_2 and surroundings T_1 is equal to:

$$\eta_c = 1 - \frac{T_1}{T_2}. \tag{19}$$

Apart from the Rankine cycle, a number of other heat engine cycles have been invented since the time of Carnot, some of which form a modification of the ideal Carnot cycle but are not easy to implement in practice. Two, the Stirling and Ericsson cycles ([11], pp. 270–272) have the same theoretical efficiency as the Carnot cycle, but even ideal versions of the others have a lesser maximum.

Apart from the overall cycle, the engineer is interested in the individual flow processes making up the cycle, reflecting the real processes taking place in thermal machinery. Determination of the efficiency of the machine where the energy conversion takes place requires application of not only the first but also the second law of thermodynamics, as the conversion of any type of energy always leads to a fraction of it being changed into heat, which as a consequence gives rise to an entropy change of the system.

For an open system, for which heat Q is both supplied and carried away, the first law of thermodynamics for steady states has the form for work output W:

$$W = Q_{in} - Q_{out} + \dot{m}\,(h_{in} - h_{out}), \tag{20}$$

noting that $\dot{m}_{in} = \dot{m}_{out} = \dot{m}$ (steady state) and h is the enthalpy.

The second law of thermodynamics related to the surroundings can be written in the form:

$$\dot{S} = \frac{dS}{dt} = \frac{Q_{out}}{T_{ref}} - \frac{Q_{in}}{T} + \dot{m}\,(s_{out} - s_{in}) \geq 0, \tag{21}$$

$$\dot{S}\,T_{ref} = Q_{in}\left(1 - \frac{T_{ref}}{T}\right) + \dot{m}\,[(h_{in} - h_{out}) - T_{ref}\,(s_{in} - s_{out})] - W \geq 0. \tag{22}$$

The following equation for the maximum work of the system can be derived from eqn (22):

$$W_{max} = Q_{in}\left(1 - \frac{T_{ref}}{T}\right) + \dot{m}\,[(h_{in} - h_{out}) - T_{ref}\,(s_{in} - s_{out})]\,. \tag{23}$$

Now the ability of a system to perform work may be regarded as a sort of 'quality' of its behaviour, where the reference level for this ability is the surroundings. At these conditions the work output is zero.

The above quantitative measure, represented by eqn (23) is termed the *exergy*, and substantial contributions to its recognition and application have been made by Szargut in Polish [12, 13] and by Kotas in English [14].

3.3 Theoretical aspects of exergy

Now the first term in eqn (23) describes the increase of exergy of the heat source, whereas the second term expresses the change of exergy of the working fluid of the system.

For a closed system ($\dot{m} = 0$), W_{max} is equal to:

$$W_{max} = Q_{in}\left(1 - \frac{T_{ref}}{T}\right), \tag{24}$$

which is the same as the work obtained in the Carnot cycle, since $\eta_c = 1 - T_{ref}/T$.

If in a system, heat is converted to work, then the efficiency of this process (defined by using the first law of thermodynamics) has the form:

$$\eta^{I} = \frac{W}{Q_{in} + \dot{m}h_{in}}. \tag{25}$$

For a closed system, i.e. when $\dot{m} = 0$, the efficiency is equal to:

$$\eta^{I} = \frac{W}{Q_{in}}, \tag{26}$$

where

$$W = Q_{in} - Q_{out}.$$

Another definition of the efficiency – exergy efficiency – that takes into account the energy quality can be derived from the second law of thermodynamics. In this approach, entropy changes – i.e. the irreversibility of the process due to which the work W is obtained instead of W_{max} are taken into account. The exergy efficiency found in this way is:

$$\eta^{II} = \frac{W}{W_{max}}, \tag{27}$$

where W is described by eqn (20) and W_{max} by eqn (23). The difference

$$W_{max} - W = \dot{S}\,T_{ref} = I \tag{28}$$

is a loss of work I due to irreversibility of the process. In the case of a closed system $\dot{m} = 0$, eqn (27) takes the form:

$$\eta^{II} = \frac{Q_{in} - Q_{out}}{Q(1 - T_{ref}/T)} = \frac{\eta^{I}}{\eta_c}, \tag{29}$$

where η_c is the efficiency of the Carnot cycle – the maximum efficiency that can be reached. Energy analyses of engineering processes based on the first law of thermodynamics and overall analyses based on the second law of thermodynamics enable the determination of energy losses and the evaluation of the maximum possible work which may be obtained from the system. Such analyses can lead to improvements in engineering processes. In fact, because exergy can be regarded as representing the (exploitable) economic value of an energy source, it can be used for the evaluation of the natural environment itself [12].

3.4 Exergy and Gibbs free energy – an engineering/biology identity

In eqn (23) the second term – the change of exergy of the working fluid – may be re-expressed as:

$$W_{max_f} = \dot{m}\,[\Delta h - T_{ref}\,\Delta s] = G_{ref}, \tag{30}$$

where the Gibbs function G_{ref} is defined by eqn (13).

In [9] Mikielewicz *et al.* pointed out that G, too, represents maximum available work, and the above shows that exergy and Gibbs function are essentially identical, if $T = T_{\text{ref}}$.

Now it is the 'Gibbs free energy' or just 'free energy' that is the standard descriptor for biological energy processes ([8], 97 ff. – compare with the earlier edition [15] (p. 117); [7], p. 239; [10], p. 36; [16], p. 18 or [17], p. 37). So, here too, via thermodynamics, engineering and biology are intimately connected.

3.5 The application of exergy – an example

As part of a series of final-year undergraduate projects supervised by Collins [18] exergy analyses were carried out for two (MAGNOX and AGR) UK nuclear power reactor systems used for electricity generation. Now in using nuclear power as an example, it is important to note that at the time of writing this chapter the first political signs are evident that in the UK at least this energy source may need renewing [19]. Under the subheading 'Timms hints at fresh nuclear builds', the Energy Minister is reported to have said 'In the future we may realize that there is a place for nuclear power and new nuclear builds'.

Based on the work of Kotas [14], and ignoring potential and kinetic energy and chemical reaction effects, the change in exergy between two points is:

$$E_1 - E_2 = (H_1 - H_2) - T_{\text{ref}}(S_2 - S_1) \tag{31a}$$

or, per unit mass,

$$e_1 - e_2 = (h_1 - h_2) - T_{\text{ref}}(s_2 - s_1). \tag{31b}$$

In any real (irreversible process) with W and Q:

$$E_1 - E_2 = W - Q + I, \tag{32}$$

where I is the irreversibility of the process.

What is termed the 'rational efficiency' is:

$$\Psi = 1 - \frac{I}{\sum E_{\text{in}}}, \tag{33}$$

with the efficiency defect δ:

$$\delta = 1 - \Psi \tag{34}$$

and component inefficiencies δ_{i} given by:

$$\delta_{\text{i}} = \frac{I_{\text{i}}}{\sum E_{\text{in}}} \quad \text{where } \Psi + \sum \delta_{\text{i}} = 1. \tag{35}$$

Using these parameters, the exergy balance for a plant may be expressed diagrammatically as Grassman and pie chart figures.

The first phase of the UK nuclear programme was based on the Magnox reactor system, with natural uranium as fuel, non-oxidising magnesium as cladding, carbon dioxide as primary coolant and a steam cycle for electricity generation. The plant diagram was simplified to final-year thermodynamics teaching level, and typical temperature and pressure data defined. The plant diagram is shown in Fig. 1, with calculated values of h, s and e incorporated into Table 1. From this, an exergy pie chart and Grassman diagram were constructed as in Figs 2 and 3.

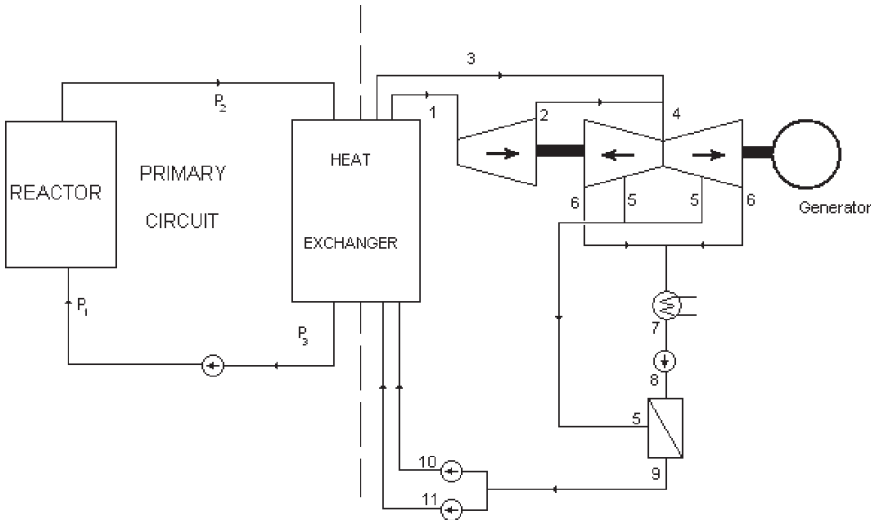


Figure 1: Simplified plant layout for Magnox reactor system.

Table 1: Magnox reactor system thermodynamic calculations.

Position in Fig. 1[a]	m(kg)	T(°C)	p(bar)	s [kJ/(kg K)]	h(kJ/kg)	ε
P_1	4808	180	10	5.95[a]	351[a]	167.0
P_2	4808	390	9.86	6.25[a]	573[a]	299.5
P_3	−4808	174	9.65	5.95[a]	345[a]	160.9
1	258	371	51	6.52	3121	1176.5
2	258		13	6.61	2840	869.2
3	96	371	13	7.25	3197	1035.4
4	354			6.80	2937	909.6
5	36			7.18	2577	431.2
6	318			7.22	2175	22.4
7	318		0.04	0.40	121	2.0
8	318		0.04	0.48	147	3.9
9	354			1.10	370	43.0
10	96			1.10	371	44.0
11	258			1.10	375	48.0

[a]Based on zero of 0 K.

Thermodynamically speaking, the main problem with the Magnox system was the low working temperatures. These were associated with the permitted maximum fuel and cladding temperatures and to avoid excessive creep and other problems in the original steel pressure vessels. As a consequence, the heat exchangers, under the 'burden' of a low temperature difference, had to be of very high surface area. This component inefficiency dominated the whole cycle (see Fig. 2)

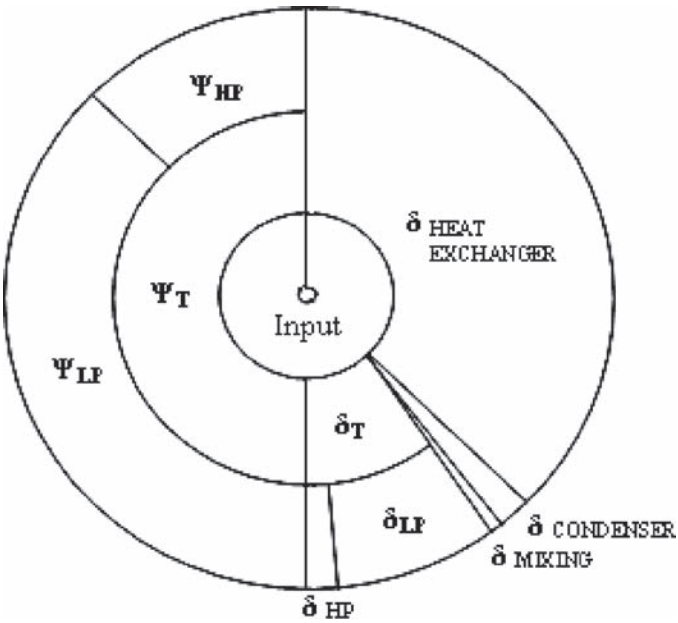


Figure 2: Magnox reactor system – exergy pie chart input.

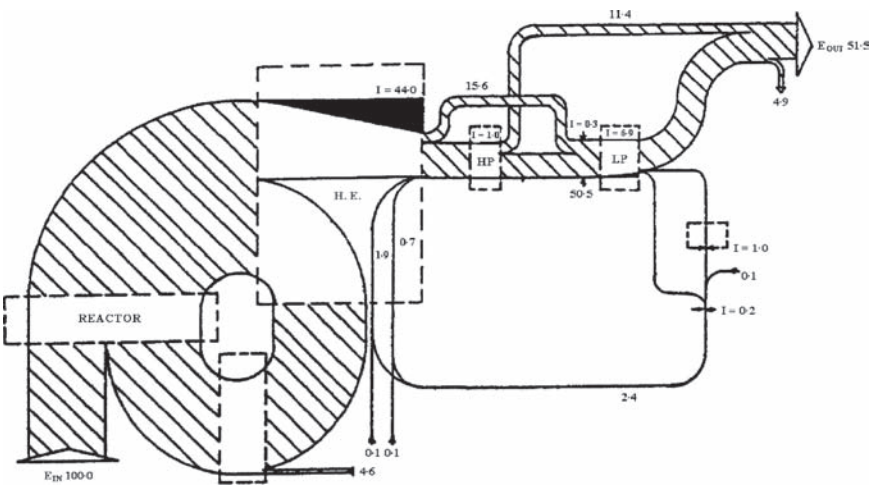


Figure 3: Magnox reactor system – Grassman diagram.

resulting in a low overall thermal efficiency. The subsequent AGR system used enriched uranium fuel and stainless steel cladding, and did not have this drawback.

In general, exergy analyses quantify the various losses in a cycle, giving a focus on where design improvements should most usefully be made.

4 Application of thermodynamics to biology – glycolysis and the tricarboxylic acid (Krebs) cycle

The essential similarity of exergy and (Gibbs) free energy has been noted and is formally analysed by Rogers and Mayhew ([11], pp. 117–125). Our engineering example has not involved chemistry, whereas the free energy analyses for biological cell energy cycles are exclusively chemical in character ([8], chapter 7). By taking glycolysis and the tricarboxylic acid cycle as an example, we are extending the application of thermodynamics to encompass the physical, engineering and biological worlds, and of exergy/Gibbs free energy to both thermal and chemical processes.

Glycolysis is a 10-reaction series of chemical processes in the living cell, whereby the substrate glucose is transformed into pyruvate. It is common to aerobic and anaerobic metabolism, as shown in figure 16 of the companion chapter in Volume 1 of this series [9]. In aerobic metabolism, glycolysis is followed by pyruvate oxidation and the tricarboxylic acid cycle, before entry of reducing potential (arising from oxidative processes in glycolysis and the tricarboxylic acid cycle) into the respiratory chain. Like glycolysis, the tricarboxylic acid cycle is a multi-process series.

Table 2, using data from [8] and [15] shows the progression of free energy through the complex chemistry. It can be seen that the release of free energy is much greater in the tri-carboxylic acid cycle than in glycolysis.

On comparing Tables 1 and 2 it is apparent that the overall multi-stage character of the thermodynamics processes are of a similar order of magnitude. However, while most biological

Table 2: (Gibbs) Free energy changes in the cell (from figures 7.12 and 7.14 [15], 7.7 and 7.9 [8]).

Process	Stages	Free energy change ΔG (kcal)	Process change ΔG (kcal)
Glycolysis	Glucose	0	
	Glucose 6-phosphate	+5.4	
	Fructose 6-phosphate	+7.9	
	Fructose1,6-biphosphate	+11.6	
	Dihydroxyacetone phosphate	+15.9	
	Glyceraldehyde 3-phosphate	+18.0	
	1,3-Biphosphoglycerate	−83.6	
	3-Phosphoglycerate	−110.3	
	2-Phosphoglycerate	−10	
	Phosphoenolpyruvate	−109.1	−137.6
	Pyruvate	−137.6	−112.3
	Acetyl CoA	−249.9	
Pyruvate oxidation/ Tricarboxylic acid cycle	Citrate (citric acid)	−273.4	
	Isocitrate	−259.5	
	α-Ketoglutarate	−370.9	
	Succinyl CoA	−504.1	
	Succinate	−572.0	
	Malate	−581.6	−418.8
	Oxaloacetate	−668.7	

cells have diameters between 1 and 100 microns ($1\,\mu\text{m} = 10^{-6}$ m) ([8], p. 56), the overall site dimension of a Magnox power station is in the kilometre range. Later in the chapter the reasons for this mismatch between engineering and biological 'power systems' will be discussed.

5 Equivalence of thermal and statistical entropy

5.1 The role of thermal entropy – a summary

In the preceding sections, thermodynamics has been applied within the physical, engineering and biological worlds. Whether in the form of mathematical relationships, quantitative calculations or measurement data, thermodynamics facilitates a clear understanding of the underlying science. While the first law focuses on the relatively understandable concept of energy, the second law shows that the absolute temperature is a crucial factor in thermodynamic performance. The combination of 'heat transferred' and the 'temperature at which the heat is transferred' is represented by the (phenomenological or thermal) entropy. So entropy enters, via exergy or Gibbs function, the interpretation of thermodynamic efficiency and maximum available (or 'free') energy, whether in a nuclear power system or in a biological cell. Finally, our engineering example has not involved chemistry, whereas the cell energy processes are exclusively chemical in character ([8], Chapter 7). Thermodynamics copes equally well with both, so this contrast demonstrates still further its all-pervasive applicability within the engineering and natural worlds.

5.2 Statistical entropy

The 'genius' of thermodynamics includes its ability to address large scale structures (e.g. steam turbines) on the one hand and detailed molecular and property behaviour on the other. For engineering thermodynamics in particular, the system concept (whether closed, open or isolated) means that a 'black box' approach is elegantly feasible. By judicious choice of the system boundary, problematic internal components, effects and processes can be virtually ignored, provided, as is usual, equilibrium can be assumed. This is because the first and second laws relate to system boundary transfers such as work and heat, and properties for the system *as a whole*, say internal energy. As a consequence, this can be a powerful approach when applied to living systems such as *Homo sapiens*.

At the same time, thermodynamics can be expressed theoretically in terms of, for example, the partial differential Maxwell relations between properties. This has already been apparent here in the treatment above of magnetic substances. Such properties can be related to molecular behaviour.

This macroscopic/microscopic comparison is directly relevant to entropy. In the section above we surveyed the place of thermal entropy in a variety of situations. Boltzmann, however introduced the concept of *statistical* entropy related to the probability of the microstates of individual molecules comprising a system.

Briefly, this is as follows. Boltzmann treated entropy as a measure of disorder of molecules forming a system:

$$S = -k \sum_i (p_i \ln p_i), \tag{36}$$

where k is the Boltzmann constant equal to 1.38×10^{-23} having dimensions of entropy. The symbol p_i denotes the probability of occurrence of a microstate. The equation corresponds to a

single molecule, or more precisely, to each degree of freedom possessing a few microstates. The contribution of all degrees of freedom is summed up for the entropy of the system. The total energy of the degree of freedom is $E = kT$ and acts as a restriction for p_i. At low temperatures some degrees of freedom are frozen and do not contribute to energy. At zero temperature on the absolute scale, all degrees of freedom have a probability equal to one or zero, and it is then obtained that $S = 0$, in consistency with the third law of thermodynamics. Distinguishing between the molecular microstate and degree of freedom of the system is important from the point of view of information theory because one can speak about the entropy of individual degrees of freedom.

5.3 Equivalence of thermal and statistical entropy

It is possible to prove the equivalence of thermal and statistical entropy using the example of a perfect gas. Let us consider an irreversible process of expansion of a perfect gas in a vacuum. Due to the fact that the surroundings is a vacuum, the system does not yield work (lack of back-pressure). The system as a whole is isolated, therefore its energy and temperature do not change. For an isothermal process

$$\mathrm{d}W = p\mathrm{d}V = \mathrm{d}Q.$$

And therefore the entropy change is equal to:

$$\mathrm{d}S = \frac{\mathrm{d}Q}{T} = \frac{p\mathrm{d}V}{T}.$$

Making use of the state equation for a perfect gas (4), one can obtain

$$\mathrm{d}S = R\frac{\mathrm{d}V}{V}. \tag{37}$$

Let us consider the same process from the microscopic point of view. The probability of finding a molecule in the initial volume V_1 is equal to:

$$p_V = \frac{V_1}{V_2}, \tag{38}$$

where V_2 is the final volume.

The probability of finding N molecules in the initial volume V_1 is still smaller and due to the independence of events is equal to:

$$p_N = \left(\frac{V_1}{V_2}\right)^N. \tag{39}$$

Taking a logarithm from the above equation one can obtain

$$\ln p_N = N\ln p_V = N\ln\frac{V_1}{V_2}. \tag{40}$$

Let us assume that V_2 does not significantly differ from V_1, i.e. $V_2 = V_1 + \mathrm{d}V$. Then

$$\frac{V_1}{V_2} = 1 - \frac{\mathrm{d}V}{V}.$$

Taking a logarithm and expanding into a series, one can get

$$\ln\frac{V_1}{V_2} = \ln\left(1 - \frac{\mathrm{d}V}{V}\right) \approx \frac{-\mathrm{d}V}{V}. \tag{41}$$

Then

$$\ln p_N = -N\frac{\mathrm{d}V}{V}. \tag{42}$$

Using eqn (37)

$$\mathrm{d}S = -\frac{R}{N}\ln p_N. \tag{43}$$

5.4 Consequences

So a relationship between the thermal entropy and the probability of occurrence of a given state of molecules is evident. The consequence is that the entropy statements of the second law of thermodynamics may be reformulated in terms of the probability of occurrence of given states of molecules. Thus an irreversible process now means a process of change from a less to a more probable microscopic state. The (thermal) Planck statement of the second law [9] is: 'it is impossible to construct a system which will operate in a cycle, extract heat from a reservoir, and do an equivalent amount of work on the surroundings'. Such a device is described as a perpetual motion machine of the second kind (or PMM2). However, the *statistical* formulation of the second law states that a PMM2 is not impossible but highly unlikely. There is, similarly, little probability of an expanded gas contracting to its initial volume, and even less probability that such a process could be repeated in a cycle.

To summarise, the statistical aspect of entropy requires the subtle re-interpretation of the second law away from absolutism to that of being very highly probable.

6 Role of entropy in contemporary studies

6.1 The different aspects of entropy

Firstly, we extend our comparison of thermal and statistical entropy to include that due to the information theory of Shannon. In brief the three aspects may be distinguished as [20]:

1. Thermal, or phenomenological, as already applied in Sections 2–4.
2. Statistical, after Boltzmann, as explained in Section 5.2.
3. Informatic, related to the information contained in a message, forming part of communication theory.

A number of researchers believe these three formulations are equivalent to each other. The first to reach this conclusion was Brillouin [21] and more recently, Peters [22]. Others, such as Wicken, strongly dispute this equivalence. Typical of his statements is 'This misapprehension blurs fundamental issues' ([10], p. 18). In Section 7 the pros and cons of 'Shannon entropy' will be reviewed.

6.2 Information theory

This description follows Thoma's comprehensive, but externally unpublished, study [23]; Shannon's source material may be found in [24].

According to Shannon, information is the message content of a signal. Each signal takes a certain position in the message and can assume a number of discrete values, each value with a probability p_i, where i changes from 1 to m and $\sum_1^m p_i = 1$.

The information transported through a given symbol i is $\log_2 p_i$. The logarithm with a base 2 is chosen here, as with a 50% probability, it gives the unit information called a bit. The average contribution of symbol i to information is $p_i \log_2 p_i$. The message is a sum of all possible pieces of information

$$H = -K \sum_i^{\bullet} p_i \log_2 p_i. \tag{44}$$

In binary notation, $m = 2$ and $p = p_i$ or $p = 1-p_i$.

In the above equation, the constant K is usually taken as unity ([10], p. 19) and H is the Shannon or confusion functional ([23], p. 7). In fact, H is customarily taken to be entropy, for example, being so used throughout in a *biological context* by Brooks and Wiley [25].

The transmitted information I is defined as the difference in the values of H before and after the communication of a message:

$$I = -(H_{\text{after}} - H_{\text{before}}). \tag{45}$$

I, being of opposite sign to H, becomes the widely used term 'negentropy'.

As an example, assume that the memory of an element contains 1024 responses with equal probability $p_i = 1/1024$ for each response before the information. Then, according to the above formula, the information before receipt is equal to 10. After receipt of the information only one signal is certain, the other signals having a zero probability. The obtained information contains 10 bits, therefore. Now as Wicken notes ([10], p. 19) constants and logarithmic bases are essentially arbitrary, so the relationships of eqns (36) and (44) 'are identical'. To avoid any ambiguity, H in eqn (44) will be termed 'Shannon entropy'.

6.3 Shannon entropy

Assuming for the moment, the validity of the concept of 'Shannon entropy' we find the applications are wide-ranging indeed.

The following discussion relates to representative contemporary subject areas.

In engineering, the cost of manufacturing an item depends on its complexity, in the sense that a quantity of information is necessary to construct that item. Such information can also form an assessment of the capital and labour costs involved. On that basis Thoma ([23], p. 14) can compare the lifetime information for steam and diesel locomotion. Quantification of information – again identified with entropy – is also a characteristic of the comprehensive application of Brooks and Wiley [25] to the field of biology. Topics, with specific calculations, include DNA (pp. 118/119) and ontogeny (pp. 147 ff.), phylogeny (p. 189), cohesion of populations (pp. 212 ff.), speciation (pp. 232/233) and phylogenetic trees (pp. 234 ff., 273), food webs (p. 304) and ecology (p. 318, p. 340).

In complete contrast to engineering and science is art. Arnheim [26] gives an intensively argued explanation of the meaning of entropy within art in general. He sees (figure 10.2, p. 30) art as 'structural order' brought about by 'two cosmic tendencies'. The *structuring* theme is achieved by an Anabolic Tendency, 'which initiates all articulate existence', with the *ordering* achieved by a Tension Reduction Tendency organizing energy in 'the simplest most balanced' manner. *In addition*, entropy provides a negative catabolic destruction and is the partial (?) cause of tension reduction. Thermodynamicists would wince at all this (to put it mildly!) but the overall argument could possibly be rewritten in a more rigorous manner.

Our final representative study is that of Goonatilake [27] who focuses on information 'flow lines' in DNA, neural–cultural systems and then artefacts (cover blurb). The climax of his treatment is the

presentation of *world history* as a series of bifurcations (pp. 155–162), i.e. as a chaotic system. The chapter in question, rather esoterically entitled 'The deep engines of entropic change', is largely inspired by the work of Shannon, Prigogine, and Brooks and Wiley (pp. 140–151).

Space does not allow discussion of the interpretation of other areas in entropic terms, such as economics (extensively referenced by Goonatilake, p. 152).

6.4 Dissipative structures

This entropic approach, stemming from the second law of thermodynamics, is by no means as revolutionary as might be thought. It should be regarded as a sub-set of the relevance of *thermodynamics* to these subject areas. This is clear from two other recent authoritative publications, which coincide with [28] or overlap [29] Goonatilake's interests. Both have detailed reference to energy (hence first law) considerations ([28], pp. 95–97, 193 ff.; [29], pp. 69 ff.). The point of our chapter is that second law considerations are equally significant, and in particular the developments in non-equilibrium thermodynamics associated with Prigogine. These especially apply to living systems, as noted in Section 1.

Such systems maintain their existence by consuming free energy/(Schrödinger's negative entropy) from their surroundings. They export entropy to their surroundings via chaotically oriented dissipative processes. The supreme example is the closed system of the biosphere, which receives low-entropy energy from the sun, and dissipates high-entropy energy in the form of radiation into space. Simple energy considerations lead to a bottom–up interpretation of organic activity via the food chain, whereas the overarching historical biosphere effect is more of a top–down low entropic driving force.

Once again an excellent summary of the whole material of this chapter, and its inclusion in a cosmic 'programme', is given by Chaisson [16].

7 Pros and cons of Shannon entropy

7.1 Introduction

It is of crucial interest to establish whether the information-based Shannon entropy may be identified, via statistical entropy, with the thermodynamics-based thermal entropy. Three levels of comparison may be distinguished: prima facie, formal thermodynamics and universality of the second law of thermodynamics. The first tends *towards* identification, the second and third *against*.

7.2 Prima facie comparison

This arises from the complete similarity of form of eqns (44) and (36) for Shannon and statistical entropy respectively. Moreover, the underlying concepts of thermal and Shannon entropy are negative in quality: 'disorder' or 'diminution of potential' (constraints; [10], p. 18) in the case of entropy, 'uncertainty' in the case of information. This argument is even finer when Shannon and statistical entropy are compared: 'uncertainty' is an acceptable expression for both. In fact (as explained later) Tribus structures his entire thermodynamic approach on the basis that entropy and uncertainty are coincident ([30], p. 77).

Further, in eqn (45), the positive quality 'information', in the sense of removal of uncertainty, has given rise to negentropy as a defined property.

7.3 Formal thermodynamics

> This practice is not to be recommended.
>
> J.D. Fast [31], p. 330

> … there are in science today two 'entropies'. This is one too many.
>
> Jeffrey Wicken [10], p. 23

A typical selection of recent thermodynamics texts [11, 32, 33] avoids this issue completely. Further, they almost avoid the statistical definition of entropy itself. Out of the over 1500 pages in [11, 32, 33] there are only Winterbone's sentence that '… statistical mechanics and the kinetic theory … do not give a good macroscopic theory of … (irreversible) processes' ([32], p. 316), and Kondepudi and Prigogine's page 'Statistical Interpretation of Entropy' ([33], pp. 91/92). The reader should not interpret our comment as in any way pejorative, but rather note that there is something unusual about thermodynamics to allow such an omission. Further, while the (again recent) specialised discussion of entropy by Dugdale [34] indeed treats statistical entropy, still the issue of Shannon entropy is not addressed. However, we do find 'Entropy and information' discussed in Fast's older study of entropy ([29], pp. 325–332).

Fast's essential conclusion is that Shannon entropy and conventional entropy are not connected at all ('are two entirely different concepts', p. 332) and any similarity starts and ends with the equations we have quoted. His arguments consist of a range of anomalies, including what he terms an 'absurd result' (p. 330) in trying to make a thermal equivalent to information. It is little wonder that 'this practice is not to be recommended' (p. 330) in his view.

Wicken is more sympathetic than Fast, because he uses information theory as an integral part of his overall explanation of evolution. Fast dismisses negentropy as a concept, describing the italicised statement of Brillouin 'Any additional piece of information increases the negentropy of the system' as a 'tautology' (p. 331). Wicken, on the other hand, uses it, but redefined as complexity. 'There is a completely appropriate alternative to "entropy" in information theory. This is "complexity".' ([10], p. 24), leading to his grand biological statement: 'This ordered, information-rich, and *negentropic* complexity lies at the heart of biological organization' ([10], p. 49).

Returning to the question of Shannon entropy, Wicken's conclusion is the same as Fast's – '… two "entropies" … one too many' ([10], p. 23). He wants to 'expunge "entropy" from the lexicon of information theory' ([10], p. 27). In support of this, he adduces a series of inconsistencies in the rationales for the two entropies ([10], pp. 19f). Now these inconsistencies can be more subtle than Fast would allow. In discussing the application of the two entropies to the crucial thermodynamic concept of path-independent changes of state ([10], pp. 22/23), Wicken admits his own confusion in the past: … 'not been exempt … this loose language ….' ([10], p. 23).

So formal thermodynamics refuses to mix with Shannon entropy. Does this mean that the entire enterprise of Brooks and Wiley 'Toward a Unified Theory of Biology' [25], Subtitle, is thereby invalidated? We turn now to the wider question of the universality of the second law of thermodynamics.

7.4 The second law of thermodynamics

> The second law of thermodynamics is not restricted to engineering. It is a fundamental law of nature … biological organisms fully comply with the second law of thermodynamics.
>
> Paul Davies [17], pp. 27/30

> The ubiquitous second law of thermodynamics … is degrading the energy in the food chain at each link.
>
> John Barrow [29], pp. 69/70

> For each structure (sun, earth's climasphere, biosphere, hominid body, human cranial, modern culture) the entropy increase of the surrounding environment can be mathematically shown to exceed the entropy decrease of the system per se, guaranteeing good agreement with the second law of thermodynamics.
>
> E.J. Chaisson [16], p. 18

> Since the Second Law governs all irreversible processes, a materialistically coherent cosmos requires their connection.
>
> Jeffrey Wicken [10], p. 6

The above quotes, which omit any from Brooks and Wiley themselves, almost write this section. The mainstream conclusion is clear – the second law is universal in its applicability. What of the approach of Brooks and Wiley? The second law is indeed relevant (as working systems, organisms are subject to the second law of thermodynamics ([25], p. 33)), but that law is only a *part* of the story. '… this is not enough' (p. 9). Yes, energy considerations are valid for living organisms 'just like steam engines' but '*strictly thermodynamic* considerations are not likely to tell us much …' ([25], p. 33). No, 'information takes precedence over energy' (p. 34), so *their* second law is the second law of (information, energy …) and correspondingly their entropy relates to (information, energy …). 'We are able to see biological evolution and thermodynamic changes as special cases of a more general phenomenon of evolution. … The second law is more than the natural law of energy flows; it is the natural law of history' ([25], p. 355/356). Employing HIT (hierarchical information theory ([25], p. 71)) leads to 'The total information capacity of the system is equivalent to the total entropy ($H_{\max}$) of the system' (p. 72). As a consequence, as neatly summarised by Goonatilake ([27], p. 150), 'they show that biological systems are entropy-increasing with respect to themselves even though they would appear to an outside static observer as entropy-reducing'. This is because Brooks and Wiley distinguish between a static observer and one 'perched on the line that represents entropy maximum' ([25], p. 39).

Allied to this is disagreement with Wicken over 'negentropy'. The latter uses 'the "negentropy" concept … to express the idea of "probabilistic compression" of distance from equilibrium' ([10], p. 36). Moreover 'negentropic complexity lies at the heart of biological organization' ([10], p. 49). For Brooks and Wiley, however, even if the entropy of living systems doesn't increase', this does not mean that (they) are in any sense 'negentropic' ([25], p. 355). But, of course, the meaning of 'negentropic' for Brooks and Wiley must be affected by their understanding of entropy itself.

Since the missions of Wicken and Brooks and Wiley are largely coincident, and ones with which we have great empathy, differences of this kind are unfortunate.

It has to be noted that Brooks and Wiley misunderstand certain thermodynamic issues. For example, on p. 7 they say that the (enthalpy-based) Gibbs free energy G was formulated for use with closed systems. In fact it is the brother function F, the (internal energy based) Helmholz free energy that is tailored to closed systems. The distinction is clear, for example, in non-flow and flow combustion situations which must use internal energy and enthalpy, respectively, as the basis for calorific values of fuels. Of course, once defined as a thermodynamic property G is automatically quantified by the thermodynamic state of the system. Nevertheless, it is the (Gibbs) free energy that 'feeds' the living systems, as all other authors attest.

More significant, however, are their comments on Prigogine's equation (pp. 9 and 57). They indeed recognise it is for open systems, but say it 'neglects the state of the system itself' (p. 9), and 'does not address the very attributes of biological systems that suggest thermodynamic behaviour'. Despite what Brooks and Wiley appear to believe, the energy cost in bringing particles into the system, is *indeed* allowed for in enthalpy-based analyses. While the *state* itself does not explicitly appear in the Prigogine equation, it is changes in state that most thermodynamic equations relate to. Further, they state that 'We do not assert that energy flow is trivial, only that there is no external energetic imperative behind organismic diversification' (p. 34). This is inconsistent with other authors, e.g. 'The biosphere … trapping radiant energy … necessarily provides riches in which AOs can emerge and evolve … the generation of structure: from molecular complexification to microsphere formation' ([10], p. 117). Doesn't this constitute some form of imperative?

Despite all the above, Brooks and Wiley's 'alternative general mechanism for driving systems' (p. 58), which directly relate to expanding phase space cosmological models, has an inherent attractiveness, and later we will return to this point.

Finally, Brooks and Wiley stress the difference between closed and open systems. For a biological (open) system their entropy would increase, but by less than it would for a closed system by virtue of some 'entropy production … dissipated into the surroundings' ([25], p. 9).

Now an engineering approach to an issue such as this postulate would be to identify some limiting case which has least uncertainty. Since the biosphere is a closed system ('the biosphere as a whole is the ultimate unit of cyclic closure' ([10], p. 146)) then Brooks and Wiley's above entropy reduction would be made zero. From that viewpoint, it is advantageous that Wicken studies the biosphere, and, moreover, that the ensemble statistics of thermodynamics can be applied in their theoretical microcanonical form ([10], p. 34). Conversely, it is somewhat disappointing that Brooks and Wiley do not consider biosphere processes.

7.5 The thermodynamics of Tribus

Finally, there is Myron Tribus's complete re-description and exhaustive treatment of thermodynamics in terms of information theory. Tribus took the Shannon information equation (44) as a basis, with S formally defined as both the entropy and uncertainty ([30], p. 77). Applying this to the statistics of a perfect monatomic gas allowed him firstly to explain temperature, thermal equilibrium and the zeroth law of thermodynamics (pp. 117–119), then heat, work and the first law (pp. 124–140), and finally 'classical' entropy and the third and second laws (pp. 140–145). It is presented as an undergraduate course text, with the explicit support in 1961 of L.M.K. Boelter, Dean at UCLA, 'who never doubted that the junior-year students could master this material' (p. ix). The rationale for this radical approach is of considerable interest, as it follows E.T. Jaynes (see p. viii) who '*took the ideas of information theory as primitive and more basic than thermodynamics*'.

The question of the validity of Tribus's approach – and behind it whether Jaynes's assertion is justifiable – is now addressed. We have seen how reluctant thermodynamicists are to welcome Shannon entropy to their high table, and this is reflected in the total lack of mention of Tribus or Jaynes by either Rogers and Mayhew [11], Kondepudi and Prigogine [33] or Fast [31], for example. However, while Tribus majors on background information theory, *in practice* he confines his mathematical application to microstate statistics vis-à-vis Boltzmann. In so doing his rationale is implicitly accepted by Wicken, who at the same time as he refuses 'Shannon entropy' accepts 'Jaynes's approach' ([10], p. 21).

Now this is consistent with the demonstration earlier in the present chapter that thermal and statistical entropies are equivalent. Classical thermodynamics is empirical in character: 'the important fact that thermodynamics is essentially an experimental science. It is not a branch of mathematics' ([11], p. 80). So, by demonstrating, as we have done, the equivalence of the two entropies, the Boltzmann statistical version is at the very least, established. Moreover, if such demonstration is extended to showing *full coincidence* of the two entropies, they then become equally and independently valid. Fast ([31] p. v), supports this: 'the two methods provided by thermodynamics and statistical mechanics'. Tribus and Jaynes, or more correctly, Jaynes and Tribus, take this logic several crucial stages further. Firstly, they adopt the Shannon information theory as a rationale which provides a definition of entropy as 'uncertainty' ([30], p. 77). They do this in a sufficiently carefully applied manner to be acceptable to Wicken ([10], p. 21), and to others. Their 'uncertainty' is microscopically oriented. So we *do* find Tribus's publications referenced by Wicken ([10], p. 232), by Brooks and Wiley ([25], p. 394) and by Winterbone ([32], p. 367). Secondly, Jaynes makes the almost revolutionary assertion of the primacy of information theory. Finally, Tribus restructures the whole thermodynamic enterprise from this direction. Unquestionably, it is mathematically oriented. Tribus, then, gives an alternative approach to teaching thermodynamics, and in fact at City University, London, UK, a colleague of one of the authors, Prof. I.K. Smith, used it as the thermodynamics section of a taught Master's course for some years, in which M.W.C. participated. Furthermore, given the accepted applicability of thermodynamics to biology, it underlines the message of this chapter that the connection of energy with information in a biological context needs full exploration. As Wicken points out, however, what is meant by information requires careful understanding. Brooks and Wiley assert the prior status of information theory – 'information takes precedence over energy when we consider the impact of the second law of thermodynamics on organisms' ([25], p. 34). However, *they* mean what they describe as 'instructional information' rather than the statistically based view of Jaynes. Broadly speaking, it is a macroscopic/microscopic contrast, but even this cannot be used in a trite manner.

7.6 Conclusion

We conclude, therefore, that it is safer not to take 'Shannon entropy' as a valid direct equivalent of thermal or statistical entropy. Despite this, the whole enterprise of Brooks and Wiley is otherwise most satisfying, with careful quantification for all applications, and consistency with conventional biology. In concluding Chapter 4, Populations and Species, they note (p. 255) 'the most startling and most hopeful outcome of this chapter is the recognition that the empirical core of neo-Darwinism, namely population biology, can be accommodated within our theory'. Setting aside the problem of entropy definition, there is a rather fine qualitative description and graph of varying biological timescales and 'production' effects (p. 372) and referring back (pp. 85/86). Their key scientific foundation is mainstream, namely the expansion of the universe and gravitational concentration of matter (p. 58). Finally, various information-related terms such as disorder/order (p. 69) and diversity, stability and evenness (p. 309), give scope for quantitative assessment of biological processes. Later in the chapter, we will give a comparative synthesis of the biological topics covered by them and by Wicken.

The debate has not gone unnoticed. Ji adopts a compromise ([36], p. 123). Like Brooks and Wiley, he uses H as the symbol in the Shannon equation and calls it entropy. However, he interprets it as complexity!

8 Information and complexity

> There is a completely appropriate alternative to 'entropy' in information theory. This is 'complexity'. … The Shannon 'entropy' of a sequence is simply the minimal program of information required for its specification.
>
> Jeffrey Wicken [10], p. 24

> A Shannon measure … is also a measure of complexity. …
>
> D. Brooks and E.O. Wiley [25], p. 41

> The complexity, or number of bits of information. …
>
> Stephen Hawking [37], p. 163

8.1 Introduction

Complexity itself is a field of considerable current interest, but we will focus on the narrow issue of quantitative meaning of the Shannon equation. The treatment is concise as it lacks controversy.

8.2 Information

> The Shannon and Weaver System does not allow for the semantics of information, its context or its meaning, a failing admitted by the founders themselves.
>
> Susantha Goonatilake [27], p. 13

There is some ambiguity about the meaning of information as defined by the Shannon equation. The information specialist W. Gitt gives an incisive – if now possibly needing updating – analysis of the overall problem. He points out ([38], p. 36) that 'Shannon completely ignores whether a text is meaningful, comprehensible, correct, incorrect or meaningless'. He addresses the issue by defining five levels of information – statistics, syntax, semantics, pragmatics and apobetics (or purpose) – supported by 14 theorems. (This paper is interesting also for its comparison of 1989 computer chips with DNA on the basis of information per unit volume. Also, estimating the total world knowledge in libraries as 10^{18} bits, in DNA molecule storage terms 1% of a pinhead volume would suffice ([38], p. 38)). Gitt's reasoned criticism only mirrors the problems of definition. For example, admittedly in the context of complexity, Gell-Mann discusses AIC (algorithmic information content) [39]. He points out its unsuitability 'since the works of Shakespeare have a lower AIC than random gibberish of the same length that would typically be typed by the proverbial roomful of monkeys'. Goonatilake's quote summarises the above.

8.3 Complexity

For our purposes, it is sufficient to note the wide attention being currently given to the concept of complexity and complex systems. Typical of authoritative substantial publications are (also giving subtitles) *Exploring Complexity – An Introduction* by Nicolis and Prigogine [40], *Complexity – Life at the Edge of Chaos* by Lewin [41], and *At home in the Universe – The Search for the Laws*

of Self -Organization and Complexity' by Kauffman [42]. Its engineering and scientific potential is epitomised by the UK EPSRC's (Engineering and Physical Sciences Research Council) 2002 initiative setting up 16 'novel computational clusters which aimed to improve understanding of the organisational process in complex systems' [43]. Most recently, there is the growing awareness of 'the simplicity of the rules … that allows living things to be complicated' (p. 103, *Deep Simplicity – Chaos, Complexity and the Emergence of Life*, [44]).

8.4 Quantification of complexity

As Gell-Mann points out [39] 'a variety of different measures would be required to capture all our intuitive ideas about what is meant by complexity'. His criticism of AIC has been noted, just as Shalizi [45] highlights Kolmogorov's measure 'as useless for any practical application'. To hand is a research paper studying mathematical models for the complex dynamics of fisheries, and the chronic problem of over fishing (*Of fish and fishermen: models of complexity* [46]). These models comprise neoclassical equilibrium approach and system dynamics in the context of the Canadian Atlantic coast.

8.5 Conclusion

In contrast to Shannon entropy, the identification of the Shannon function with a quantitative measure of complexity finds Wicken, Brooks and Wiley, and Ji in accord. Moreover Hawking's terse comment is actually biological: 'the complexity, or number of bits of information, that is coded in DNA is roughly the number of bases in the molecule'. Also, this quite straightforward definition has a parallel in Thoma's application of Shannon's formula to complexity/information for capital/labour/specification in a mechanical engineering context ([23], p. 11).

Finally, since Brooks and Wiley's calculations are Shannon based, a large reconciliation is possible by using the term complexity, rather than entropy, that is to say 'Evolution as complexity' as an alternative title.

9 Evolution – a universal paradigm

… what Darwinism does for plants and animals, cosmic evolution aspires to do for all things.

Eric Chaisson [16], p. 14

… the cosmological evolution, including the evolution of living systems. …

Sungchal Ji [36], p. 156

… evolutionary processes in three discrete domains. These domains are biology, culture and man-made information systems.

Susantha Gonatilake [27], p. 1

Last but not least, the universe as a whole is continuously evolving.

Grégoire Nicolis and Ilya Prigogine [40], p. 37

9.1 Introduction

In certain respects, this section forms the climax of our chapter. In it we discuss the generalisation of the evolutionary rationale to 'all things' (above quote), the underlying driving forces of gravity and the expansion of the universe, and the key roles played by thermodynamics and information/complexity. The issues are well illustrated by alternative graphical presentations, redrawn from the works of a number of authors.

9.2 The expansion of the universe and its gravity

> … a veritable prime mover … is the expansion of the universe itself.
>
> Eric Chaisson [16], p. 13

> The most important fact about the universe today is that it is expanding.
>
> John Gribbin [44], p. 111

> Looking at the universe as a whole … arranged itself into shining proto-galaxies … the expansion of the universe assisted … an entropy gap opened up … all sources of free energy … can be attributed to that gap. … The ultimate source of biological information and order is gravitation.
>
> Paul Davies [17], p. 41

> … the evolution of the cosmos … there has existed an asymmetry between potential and kinetic forms of energy due to cosmic expansion, which makes descents into potential energy wells entropically favourable … constitute *means* for the dissipation of potential energy.
>
> Jeffrey Wicken [10], p. 63/64

It is part of the challenge of our overall study, but also part of its satisfaction, that the entire biological enterprise rests on the secure scientific foundation of the nature and history of the universe. Not only that, but such a foundation is, almost by definition, a thermodynamic one, as our quotes make clear. That history is given in display form by, for example, Rees ([47], p. 119) and Hawking ([37], pp. 168/169) with (p. 78).

The next key event to follow the Big Bang was the decoupling of matter and radiant energy, completed after around 100,000 years. Figures 4 and 5, redrawn from [16], display the processes quantitatively. Following this decoupling the temperatures diverge, and result in the inception of information/complexity as explained by Chaisson.

An extended alternative to Figs 4 and 5 is presented qualitatively in Fig. 6, in three-dimensional form (redrawn from [36], p. 156).

In Fig. 6, E represents qualitatively the energy density of the earth so is somewhat more specific than Fig. 4. Also, the information density I represents only biological information, as opposed to the overall cosmic information of Fig. 5b which increased from zero after decoupling. Ji also postulated a possible non-zero information content of the universe at the time of the Big Bang which (despite 'probably varying with time', Ji, figure explanation, ([36], p. 156)) is represented as a constant thickness in the I direction.

The overall message is clear and presented by both authors in an essentially consistent manner. Chaisson's concise accompanying description, which stresses the thermodynamics ([16], p. 16) is complemented by that of Davies ([17], pp. 37–41), who explains the thermodynamic consequences

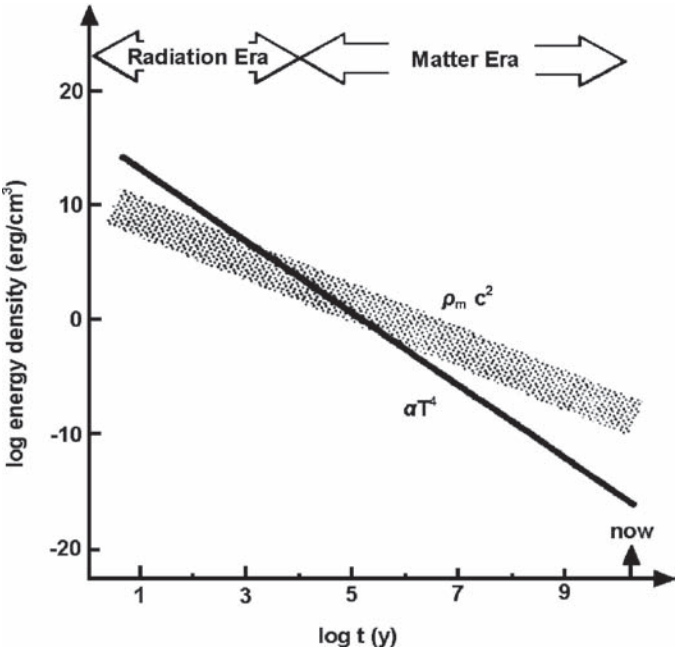


Figure 4: Variation of energy density of matter and radiation. The lines cross at about 10,000 years, one-tenth of the time taken to fully decouple matter and radiation (redrawn from [16]).

of gravity. Figure 5b reflects the 'virtually informationless' character of the originally 'undifferentiated and highly uniform blob of plasma' (Chaisson) – or the 'very little information' of the 'uniform gas' (Davies). Figure 5b also reflects the commencement of cosmic information due to the concentration of matter. 'A smooth gas grows into something clumpy and complex' – 'a star cluster or a galaxy requires a lot of information to describe it' (Davies).

The existence of gravity provides a key to the cosmic thermodynamic enterprise. 'In the 1980's' (Davies explains) 'the puzzle of the source of cosmic energy was solved', 'because its gravitational field has negative energy'. Also 'the universe came stocked with information, or negative entropy, from the word go'. In the context of our post Big Bang plasma/gas 'a huge amount of information evidently lies secreted in the smooth gravitational field of a featureless, uniform gas'.

So as the cosmic matter/radiation system evolves, information emerges, and a cosmic entropy gap opens up – the difference between the maximum possible entropy and its actual entropy.

There is therefore a cosmic entropic driving force with the objective of raising the universe's actual entropy to its maximum – in other words, the second law of thermodynamics. The thermodynamics of our earth is part of this cosmic mission. Although in *energy* terms, 'the earth sends back into space all the energy that it receives from the sun', in *entropy* terms, the energy 'we do receive has a far lower entropy than the (equal) energy that we return', to give Penrose's simple yet telling explanation (Introduction ([6], pp. xi–xii)). Whether concerning our chemical or geothermal energy, or – specifically for this chapter – 'biological information and order', 'the ultimate source is gravitation' ([17], p. 41).

The effect of cosmic expansion gives the 'ever-widening gradient of the universe's 'heat engine' (([16], p. 17); Fig. 5) and so becomes 'the prime mover'. Whereas for Davies the focus is on gravitation, there is no contradiction. In fact, Hawking ([48], p. 165) gives the same sequence as

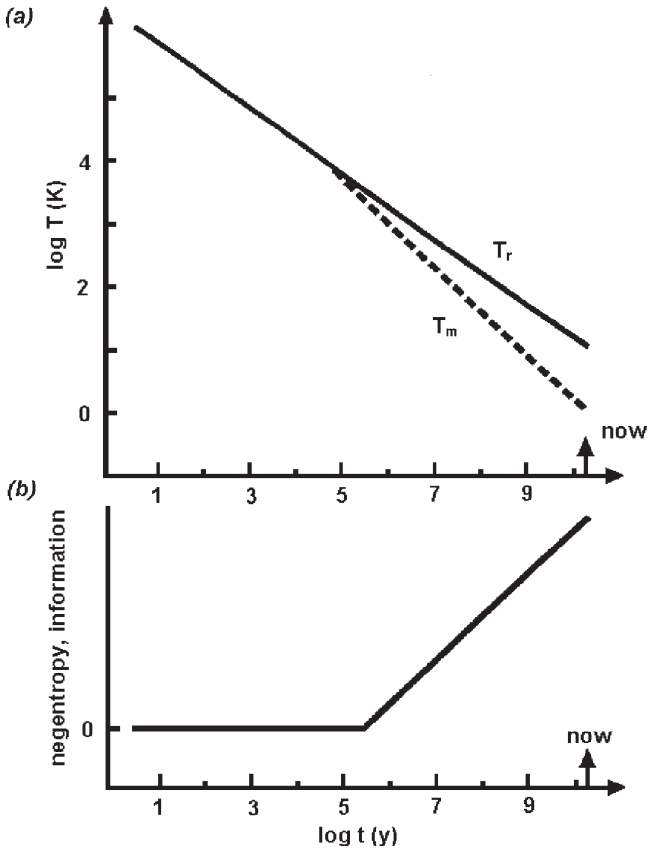


Figure 5: Temperature variation of matter and radiation in the universe, leading to: evolution of information, after Chaisson(Redrawn from [16] and quoting his comments in full.) (a) The temperature of matter and radiation went their separate ways once these quantities became fully decoupled at $t \approx 100{,}000$ years. Since that time, the universe has been in a non-equilibrium state – a kind of cosmic heat engine. (b) The potential for rising negentropy or information content – un-quantified here but conceptual synonyms for 'complexity' – is broadly proportional to the growing thermal gradient in the universe.

Davies, but with the stress laid on the cosmic expansion. So expansion and gravitation jointly form the cosmic fact of life.

Finally, like Davies, Chaisson is particularly attracted by 'life forms' which 'arguably comprise the most fascinating complexities of all' ([16], p. 16). As a consequence his cosmic history has only three eras, the radiation era and matter era (Fig. 4) now succeeded by the very recent life era – 'the emergence of technologically intelligent life' ([16], p. 17).

So the thermodynamics of the universe is identified with our *current* biological perspective.

9.3 The evolution of information/complexity

9.3.1 General

The preceding section focused on two major historical events, the decoupling of matter and radiation, of cosmic significance, and the origin of life on earth of biological significance.

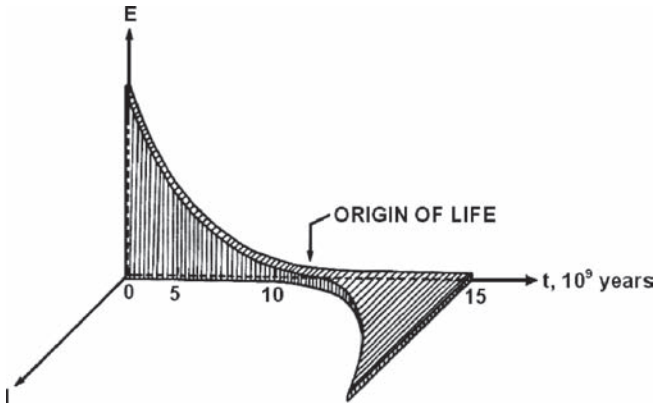


Figure 6: Evolution of information, after Ji (redrawn from Ji). 'The sudden increase in the information density (defined as the amount of biological information divided by the volume of the biosphere) occurred with the emergence of the first self-replicating systems in the biosphere on the earth about 3billion years ago' ([36], p. 156).

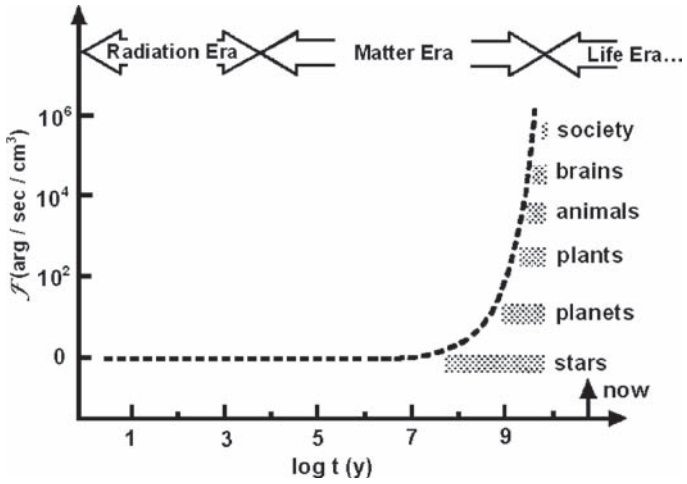


Figure 7: The components of cosmic evolution expressed as a function of their consumption of free energy. (Redrawn from [16] and quoting his comments in full.) The rise in free energy rate density, F, plotted as horizontal histograms for those times at which various open structures have existed in nature, has been dramatic in the last few billion years. The dashed line approximates the rise in negentropy, information, or complexity sketched in the previous figure, but it is energy flow, as graphed here, that best characterises the order, form, and structure in the universe. The three principal eras, discussed in this paper, are bracketed across the top.

In Fig. 7 Chaisson's minimalist history of Fig. 5b is now detailed. Cosmic evolution results in negentropy/information/complexity residing in (a) 'matter scattered throughout the universe' in space, (b) biological structures, and (c) society. Moreover, it is not so much the output which can be quantified, as the rate at which free energy can be processed, as a kind of volumetric effectiveness.

This, we should be reminded, is the direct consequence of the free energy arising, from the entropy gap, and from which 'all life feeds' ([17], p. 41). It is all part of the grand entropic scheme, with the effectiveness of free energy consumption increasing substantially from stage to stage in the evolution.

Goonatilake has the same rationale. It is based on what we might now regard as the new Prigogine-oriented thermodynamic orthodoxy. 'Free energy is continuously fed into the living system to balance the outflow into the environment occurring with the decrease in entropy owing to the increase in information. In the open systems of living organisms, entropy decreases with growing differentiation' ([27], p. 14).

Furthermore, by combining 'the living system and the environment' Goonatilake correctly, and identically with Chaisson ([16], p. 18 and quoted previously) states that consistency with the second law is achieved, as 'the entropy in the total system – of the living (organism) and its environment – increases' ([27], p. 14).

Not only is the basis of Goonatilake's and Chaisson's studies entirely consistent, but so is their attitude to information. Goonatilake sees all information systems, whether they are purely biological, cultural or artificial (called 'exosomatic') as part of the same evolutionary stream. So the evolutionary phylogenetic tree is extended to include the other two information systems, as in Fig. 8.

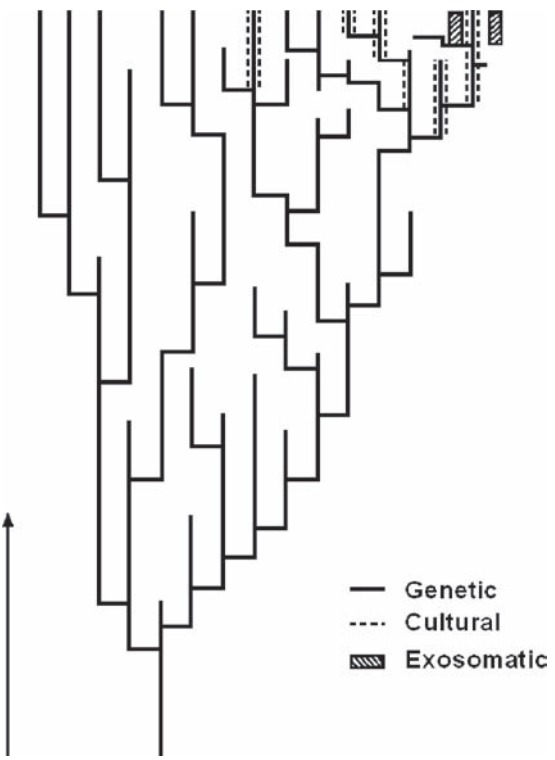


Figure 8: Phylogenetic tree with cultural and exosomatic information systems (redrawn from figure 9.5 [27], p. 139.)

9.3.2 Cultural information

The only eucultural species is man

Susantha Goonatilake [27], p. 31

Using material from [49], Goonatilake presents the various grades of cultural evolution, and the number of species for each grade.

This material enables the (broken lines of) cultural evolution to be superimposed on the phylogenetic tree of Fig. 8, where the tree itself represents the evolution of *genetic* information.

It will be noticed from Table 3 how each new defined grade of culture results in a sharp reduction of number of species, until, at the eucultural level the only species remaining is man. Furthermore, man is also unique in developing Goonatilake's 'exosomatic' or artificial information systems.

9.3.3 Exosomatic information

This is one of the principal themes of Goonatilake's study [27], and is indicated in Fig. 8 as a region of evolution of information for *Homo sapiens* at the upper right of the tree. The region is in parallel, therefore, to the eucultural information. The exosomatic information stream commenced with writing (timescale ∼4000 years ago ([27], p. 130)) through printing (1457), steam printing

Table 3: Explanation of cultural evolution and corresponding species. (Redrawn from figures 4.2 and 4.3 [27], pp. 30–31; Lumsden and Wilson [49]).

	Components				
Grades	Learning	Imitation	Teaching	Reification (including symbolization and abstract thinking)	Species density
Acultural I					
Acultural II	•				All invertebrates and cold-blooded vertebrates, i.e. 1,000,000 species
Protocultural I	•	•			8600 species of birds and 3200 species of mammals
Protocultural II	•	•	•		Seven species of wolves and dogs, single species of African wild dog, one species of dhole, one species of lion, both species of elephant, 11 species of anthropoid apes
Eucultural	•	•	•	•	Man is the only species

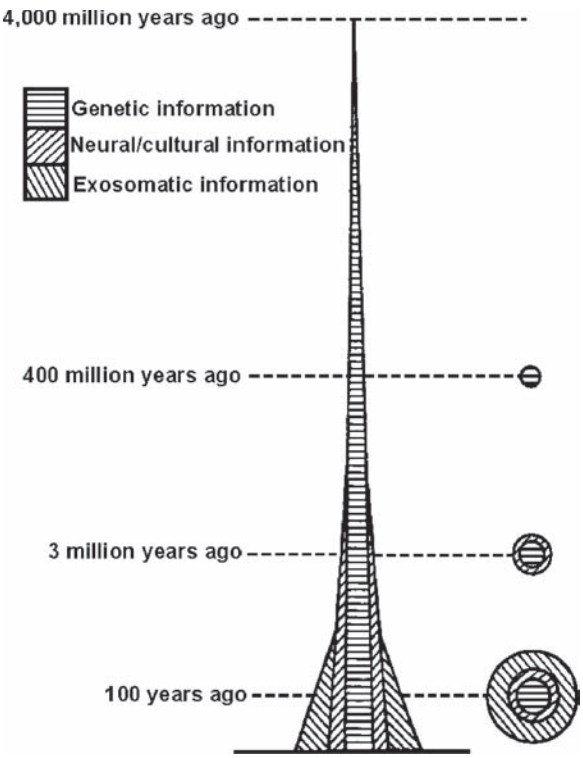


Figure 9: Evolution of information on earth, after Goonatilake (redrawn from figure 9.3 [27], p. 133).

(1814), still photography (1820), phonograph/telephone (1876) as typical landmarks, and is with us today in full flood with a plethora of computer-oriented systems. In parallel with this is an assessment of the information content – cuneiform tablet – 10^2 bits, typewritten page – 10^3 bits, magnetic tape – 10^6 bits, and 'ultrafine silver haloid film' using an electronic microscope to generate microbeam storage – 10^{12} bits ([27], p. 95). A third aspect is the *manner* in which exosomatic information has grown, as plateaus but with each invention causing rapid growth – 'rapid information growth with discontinuities' is Goonatilake's description ([27], p. 128). In fact, Goonatilake points out that it is a 'similar phenomenon to punctuated equilibrium in genetic evolution' (p. 68).

Omitting these discontinuities enables Fig. 9 to be constructed, as a qualitative description of the overall evolution of information since the origin of life on earth. Goonatilake's model may be compared with that of Hawking ([37], p. 163), redrawn as Fig. 10.

Apart from Hawking's omission of cultural information, the interpretation is essentially the same. Moreover, in the 10 years between their publication, the *quantitative* assessment of DNA has become clear, so Hawking can directly compare the 'genetic' and 'exosomatic' aspects of Goonatilake.

9.3.4 Genetic engineering

An extremely important issue for our discussion, is the feedback information loop which enables *Homo sapiens* to adapt its own DNA. 'There has been no significant change in human DNA in the

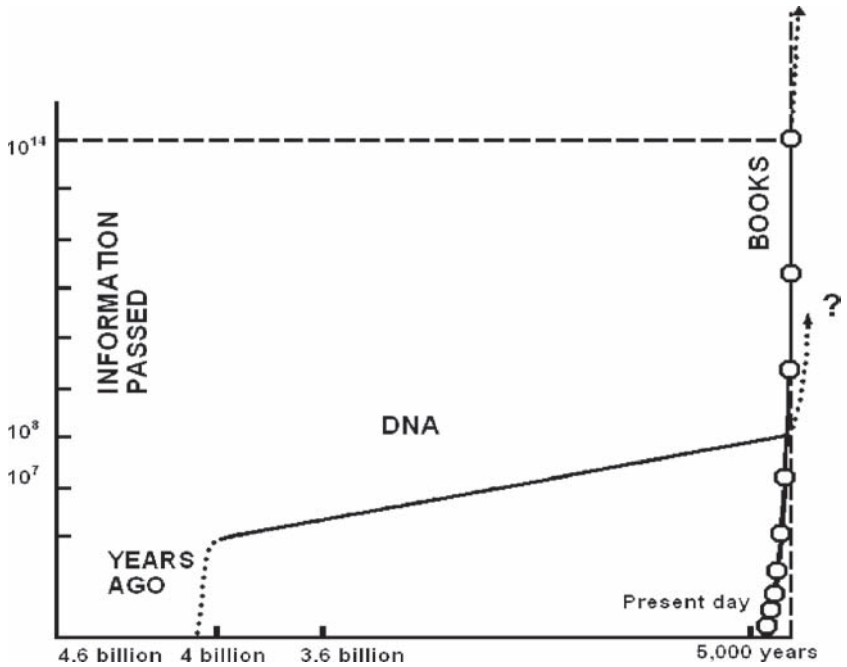


Figure 10: Evolution of information on earth, after Hawking (redrawn from figure 65 [37], p. 163).

last 10,000 years, but it is likely that we will be able to completely redesign it in the next thousand' ([37], p. 165). He points out that this is a rational prediction, irrespective of ethical issues. 'Unless we have a totalitarian world order, someone somewhere will design improved humans'. For Chaisson, this 'feedback loop' (our words) is a key aspect of his cosmic evolutionary scheme. 'Technologically competent life differs fundamentally from lower forms of life … after more than 10 billion years of cosmic evolution, the dominant species on earth – we, the human being – has learnt to tinker not only with matter and energy but also with evolution. … We are, quite literally, forcing a change in the way things change' ([16], pp. 16–17).

9.4 Time's arrow

> The increase of disorder or entropy with time is one example of what is called an arrow of time, something that distinguishes the past from the future.
>
> Stephen Hawking [48], p. 161

> The cosmological arrow generates randomness; the evolutionary arrow generates complexity and organization. We must connect the two.
>
> Jeffrey Wicken [10], p. 77

> When a process is always spontaneously irreversible, it can be said to have an 'arrow of time'. … The arrow of time and the arrow of history for biology are complementary…
>
> Daniel Brooks and E.O. Wiley [25], pp. 6, 63

The interface between thermodynamics and biology also requires an understanding of the concept of time. The second law with its entropy always tending to increase, means that that increase in entropy is also a marker of the passage of time, from past to future. For Brooks and Wiley, their 'why' of the 'path to a more unified theory of biological evolution … must include the contributions of the only natural law with a sense of time' ([25], pp. 50–51). For Goonatilake's study of the evolution of information, it is the same. The second law 'gives direction in time to physical processes, "the arrow of time" in the words of Arthur Eddington', and 'evolution as entropy in biology' is followed by 'historical change and entropy', ([27], pp. 144, 147 and 156).

Here again we detect a common understanding of time as an arrow, and Chaisson uses it as a cosmic 'intellectual road map' ([16], p. 13), as shown in Fig. 11. Its content has already been largely discussed.

It is clear from the introductory quotes above that there is more than one arrow, and this issue is very well explained by Hawking ([48], Chapter 9) 'The arrow of time'. 'There are', he says, 'at least three different arrows of time … thermodynamic …in which disorder or entropy increases … psychological …we remember the past but not the future … cosmological …universe is expanding rather than contracting' ([48], p. 161). The first two 'necessarily point in the same direction', because computational processes (equivalent as far as we know to thought processes) will always obey the second law of thermodynamics. Comparing the first and third is more subtle, but consistent with the sequence described by Davies. 'The universe would have started in a smooth, and ordered state, and … during this expansion … would become lumpy and disordered as time went on. This would explain … the thermodynamic arrow of time' ([48], pp. 165–166).

So, inherent in (the time direction of) the expanding universe is (the time direction of) the disordering process, hence the two arrows point in the same direction. It is the former that gives rise to Wicken's 'evolutionary arrow that generates complexity and organization' ([10], p. 77). Moreover, Hawking's *three* arrows point in the same direction and his argument is that it is only then 'that conditions are suitable for the development of intelligent beings who can ask the question. …' ([48], p. 161). This is connected with his discussion of the anthropic principle – termed 'We see the universe the way it is because we exist' ([48], p. 137 ff.). [The anthropic principle, finally, he connects with possible belief in a Creator – 'One can take this either as evidence of a divine purpose in Creation and the choice of the laws of science or as support for the strong anthropic principle' (p. 139). Although the ethos of this chapter, volume and series is to

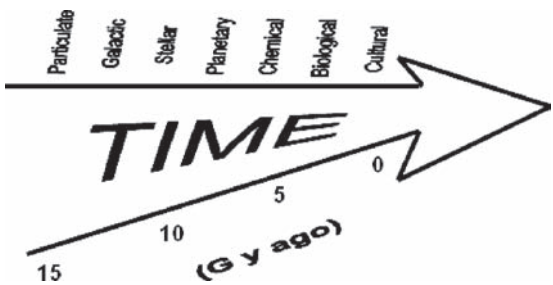


Figure 11: The cosmic arrow of time, after Chaisson (redrawn from [16], figure 1, p. 14). An arrow of time can be used to highlight salient features of cosmic history, from the beginning of the universe to the present. Sketched diagonally along the top of this arrow are the major evolutionary phases that have acted, in turn, to yield increasing amounts of order, form, and structure among all material things.

be neutral on such issues, they tend to appear consistently, and could form an attractive extension of our study. Frequently, the authors of our references either implicitly or explicitly express their individual beliefs. M.W.C. writing as series editor].

A related issue could be mentioned in closing. It is the question of what would happen in a contracting or collapsing universe ([48], p. 166; [10], p. 78). Then, and in common with the behaviour of black holes, the cosmic and thermodynamic arrows would be opposite – 'the thermodynamic and psychological arrows of time would not reverse when the universe begins to recontract or inside black holes' ([48], p. 167). Wicken mentions a 'devolutionary … arrow' ([10], p. 78).

9.5 Conclusion – an evolutionary paradigm

This section, well epitomized by Chaisson, appears to concentrate all science, engineering and humanities into one space – cosmic evolution. We have shown that in fact the ground rules of this concentration are fairly few – the history of the universe and the laws of thermodynamics, especially the second. There is a fairly tight range of individual aspects – such as cosmic expansion – with a common interpretation. Authors have also presented the evolution of information graphically, and again such presentations are consistent.

Overall, this section reflects an emerging mainstream of thought on cosmic and terrestrial history.

10 Evolution of the biosphere

> Such a thermal gradient is the patent signature of heat engine.
>
> Eric Chaisson [16], p. 17

> The source–sink dynamics is responsible for the energetic 'charging' of the prebiosphere prior to life's emergence.
>
> Jeffrey Wicken [10], p. 39

> The biosphere has evolved over the ages … absorbing solar energy, chemically degrading radiant energy, and releasing thermal entropy to space.
>
> Jeffrey Wicken [10], p. 39

> … the evolution of the biosphere is manifestly a physical process in the universe…
>
> Stuart Kauffman [50], p. 245

10.1 Introduction

The history of the prebiosphere/biosphere is the primary key to uniting thermodynamics with biology. Embedded within the cosmic evolutionary programme are two high temperature (low entropy) sources of energy, earth's mantle and the sun. These provide a top–down driving force for the eventual emergence and development of life forms. The processes are especially addressed by Wicken [10].

10.2 The biosphere

As a working definition Wicken uses 'not only the blanket of living things that covers the earth, but also the abiotic matrices in which they live – which include the atmosphere and geochemical stores with which they exchange materials and energy. The biosphere as a whole is a closed thermodynamic system, cycling elements and irreversibly processing energy' ([10], p. 74).

This is quite close to a typical 'popular' encyclopaedic definition: 'that region of the earth's surface (land and water), and the atmosphere above it, that can be occupied by living organisms' [51].

A somewhat weaker definition is provided by the exhaustive undergraduate biological text *Life* ([8], p. 8). 'Biological communities exchange energy with one another, combining to create the biosphere of earth.'

The question of precise definition has a direct relevance, for example, to Lovelock's 'gaia' concept [52]. It is hoped to review this, in the thermodynamic terms of this chapter, in a future volume of the Series.

Further, the energy issues on which Wicken and 'Life' focus above are comprehensively identified by Barrow ([29], pp. 69–70). The issues form the background to a fascinating and quantitatively well informed discussion of the weight, size, complexity, population density and brain size of various species ([29], pp. 68–86). For Barrow, the second law is 'ubiquitous', degrading 'the energy in the food chain at each link' (pp. 69–70).

10.3 The thermodynamic model

10.3.1 The terrestrial heat engine

Under the heading 'Sources, Sinks and Energetic Charging' ([10], p. 70), Wicken describes the charging of the prebiosphere to 'higher levels of free energy' and how, by dissipating that free energy 'molecular complexity' was 'generated'. Assembling the various components, the thermodynamic activity may be visualised as a kind of terrestrial heat engine, but with no work output. Figure 12 gives the two cases of geothermal and solar heating.

There are subtle differences, however. In an engineering heat engine, the thermodynamic state of the system itself varies cyclically, and $Q_1 > Q_2$, with $Q_1 - Q_2$ being the work output. In the case of the prebiosphere, $Q_1 > Q_2$, but the inequality results in a 'charging' of the system.

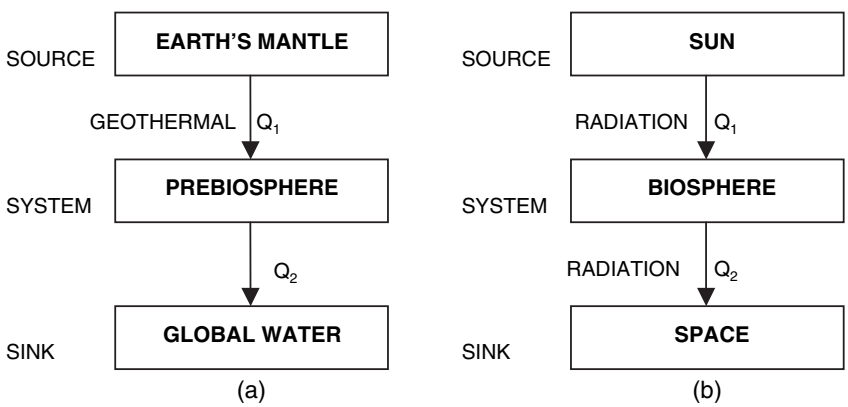


Figure 12: The terrestrial heat engine dominant for (a) prebiosphere and (b) biosphere.

In the case of the biosphere, we have noted [6] that currently $Q_1 = Q_2$. Elsewhere, Wicken ([10], p. 146), points out that 'since irreversible flows of energy through closed systems require cyclic movements of matter' there is a strong cyclic character to the biosphere. For the prebiosphere, there is a drive towards such cycling.

Using the standard approach of Prigogine (eqn (3), but as expressed by Wicken ([10], p. 70)) for either an open or a closed system the change in entropy for the system ΔS is given by:

$$\Delta S = \Delta S_{\mathrm{i}} + \Delta S_{\mathrm{e}}, \tag{46}$$

where ΔS_{i} is the change within the system due to its internal irreversibilities, the second law requiring $\Delta S_{\mathrm{i}} > 0$, and ΔS_{e} is the net exchange with the environment.

From Fig. 11 (or the engineering equivalent ([9], figures 8 and 9, pp. 38–39)) the second law Carnot inequality is:

$$\frac{Q_2}{T_2} - \frac{Q_1}{T_1} > 0, \tag{47}$$

where

$$\Delta S_{\mathrm{e}} = \frac{Q_1}{T_1} - \frac{Q_2}{T_2}, \tag{48}$$

giving

$$\Delta S = \Delta S_{\mathrm{i}} + \left(\frac{Q_1}{T_1} - \frac{Q_2}{T_2}\right), \tag{49}$$

enabling $\Delta S < 0$, i.e. 'a necessary condition for evolutionary self-organization' ([10], p. 71). However, it is not a sufficient condition. The prebiotic evolution leading to emergence involves 'chemical potential and molecular complexity' (p. 71). This sequence also necessitates the '*penetration* of (solar) energy so that at certain points Q_1 exceeds Q_2' (p. 71).

This local penetration is the bottom–up prebiosphere driving force, complementary to the overall heat engine model. Wicken re-expresses Q_1 and Q_2 locally, so a kind of micro heat engine is driven by:

$$\Delta H_s = (Q_1 - Q_2)_s. \tag{50}$$

Correspondingly, a local geothermal driving force is provided by the identity $\Delta H = \Delta G + T\Delta S$, where ΔG is the Gibbs free energy, resulting in:

$$\Delta H_{\mathrm{g}} = (Q_1 - Q_2)_{\mathrm{g}} = (\Delta G + T\Delta S)_{\mathrm{g}}, \tag{51}$$

where subscripts 's' and 'g' refer to solar and geothermal respectively.

10.3.2 The growth of information

> … the randomizing directive of the second law begins to make contact with the integrative movement of evolution.
>
> Jeffrey Wicken [10], p. 76

This is based on information theory in the sense of statistical mechanics, and not on Shannon-type information. So macroscopic information (I_M) refers to a macrostate, the summation of 'microstates contributing to that state' ([10], p. 74). Specifically I_M 'is related to the probability of the macrostate's occurrence' (p. 74). Omitting the underlying statistically related mathematics, and with $I = -S$ as a basic relationship, I_M is given by:

$$I_M = I_c + I_{th} + I_e, \tag{52}$$

where I_e is the energetic information, $I_e = E/T$ (E is the internal energy and T is the temperature); I_c is the configurational information, the overall probability of the 'spatial configuration' of the constituents and I_{th} is the thermal information, the overall probability of 'allocation of kinetic energy among their … quantum states' ([10], p. 75).

Wicken then divides 'the universe' (u) (p. 75) into the 'limited system under consideration' (s), and an infinite-capacity reservoir (r) 'with which s can exchange materials and energy'.

With ΔS representing a change in entropy, the second law can be expressed as:

$$\Delta S_u = \Delta S_s + \Delta S_r > 0. \tag{53}$$

Now

$$\Delta S_s = -(\Delta I_c + \Delta I_{th}),$$

and with no work output from *s*,

$$\Delta S_r = (Q/T)_s = -\Delta E/T = -\Delta I_e.$$

The second law requirement then becomes:

$$\Delta I_c + \Delta I_{th} + \Delta I_e < 0. \tag{54}$$

This is a significant equation, as it re-expresses the second law entirely as changes of information, where 'information' is interpreted in microscopic/statistical terms. (To digress somewhat, Tribus [30] uses the entropy to information conversion in reverse, in his formulation of the laws of thermodynamics we considered earlier).

Equation (54) enables Wicken to explain the core of a thermodynamic understanding, based on statistical considerations, of prebiotic history. Growth in prebiotic complexity requires an increase in $I_{th,}$ i.e. in overall structuring, since structuring requires the 'movement' of thermal energy from practically continuous translational modes to much less densely spaced vibrational modes … reductions in kinetic freedom … hence reductions in thermal quantum states ([10], p. 76). ΔI_{th} is given by:

$$\Delta I_{th} < -\Delta I_c - \Delta I_e. \tag{55}$$

This relationship expresses the thermodynamic constraints which allow 'evolutionary complexification' to occur.

Summarising, Wicken mentions two prebiotic effects:

1. formation of water from hydrogen and oxygen (I_{th} and I_c increase);
2. conditions in the atmosphere and oceans (I_{th} increases, I_c and I_e decrease).

However, he focuses on solar radiation, which provides increasing I_e. Pointing out that I_e/I_{th} changes tend to be reciprocal, this gives:

3. solar radiation, increasing I_e, then reducing I_e, with increase in I_{th}.

10.3.3 The arrow of time

The cosmological arrow generates randomness; the evolutionary arrow generates complexity and organization. We must connect the two.

Jeffrey Wicken [10], p. 77

Wicken points out that the second law does not 'mandate directional changes in complexity with time. All it mandates are expansions in probability space'. However, the route of any process is constrained by the local prebiosphere conditions.

Hence the overall increase in 'randomness of matter – energy in the universe', causing I_e to increase, is followed by the locally constrained and irreversible conversion of $I_e \rightarrow I_{th}$. This is shown in Fig. 13.

Also, he shows that 'matter-randomization promotes …reactions essential to molecular evolution', even in the absence of energetic charging. For these reactions, I_{th} and I_e are approximately constant, but the overall probability, $I_m \approx I_c$ becomes negative – 'a highly creative force' (pp. 79, 80).

Finally, Wicken lists the various stages of prebiotic evolution.

1. formation of simple molecules,
2. formation of biomonomers (amino acids, sugars, etc.),
3. formation of biopolymers (polypeptides, nucleic acids),
4. aggregation of biopolymers into microspheres,
5. emergence of 'protocells'.

While the first four stages 'can be understood straightforwardly' from the preceding, the last 'crucial' step is the 'most difficult to explain' ([10], p. 81) Wicken addresses it in Part III of [10].

Our summary, due to lack of space, must finish at this point. There remain life's emergence and Darwinian evolution to consider, which are addressed jointly in the next section.

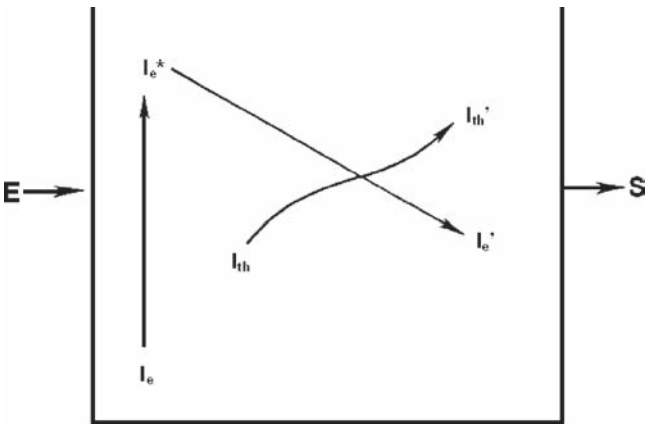


Figure 13: The arrows of time in the prebiosphere – the growth of thermal information under the prebiosphere's source–sink gradient (redrawn from [10]).

11 Thermodynamics, life's emergence and Darwinian evolution

> … a description in which evolution and history play essential roles. For this new description of nature, thermodynamics is basic. This is our message to the reader.
>
> Dilip Kondepudi and Ilya Prigogine [33], p. xvii

> We shall travel together through thermodynamics and biology.
>
> Enzo Tiezzi [35], p. VIII

> This book is written partly as a research program for bringing the mainstream of thermodynamic and statistical–thermodynamic thinking *conservatively* into evolutionary theory.
>
> Jeffrey Wicken [10], p. 7

> In this book we will develop the idea that evolution is an axiomatic consequence of organismic information and cohesion systems obeying the second law of thermodynamics in a manner analogous to, but not identical with, the consequence of the seconds law's usual application in physical and chemical systems.
>
> Daniel Brooks and E.O. Wiley [25], p. xi

11.1 Introduction

It may have been true in the past that 'thermodynamics has been an uninvited guest in evolutionary discourse' ([10], p. 7), but this is no longer the case. The objectives of the above four books make this clear. Like Wicken, our aim is to focus on mainstream thermodynamics, which includes consistency with Prigogine's analyses. It is also apparent from the start that the approach of Brooks and Wiley differs somewhat.

In this section, we cannot hope to do other than provide a skeleton of the material presented as two entire books (by Wicken and by Brooks and Wiley) and other extensive writing (especially by Ji and Kauffmann). On closer analysis, we will find that the approach of Brooks and Wiley is representative of a much wider emerging unity. Three key themes are present in the living systems we now have to consider:

> thermodynamic input, i.e. Gibbs free energy, output, or information/complexity, and internal behaviour of the system.

The fundamental issue is 'What is life?'. This will lead to the question of life's emergence, and thence to Darwinian evolution.

11.2 What is life?

> What is life?
>
> Erwin Schrödinger, Title of book [5]

> What is life?
>
> Jeffrey Wicken, Title, Chapter 2 [10]

… a proper definition of life itself …

Stuart Kauffman, [50], p. 72

… what distinguishes *life* from *nonlife* …

Sungchal Ji, [36], p. 99

… the riddle of life.

Ilya Prigogine, Chapter title [7]

To consider such a weighty question as 'What is Life?' is like climbing a whole mountain of perception. The reader may expect the view from the top should have an intellectual grandeur to reward the effort involved. Such is provided, we feel, by Kauffman ([50], p. 47), '… that mysterious concatenation of matter, energy, information and something more that we call life'. In keeping with our aim of defining the common ground, we have analysed a number of models in some detail. We find two principal features, those of thermodynamics and information.

Table 4 shows a synthesis of five relevant models, which, when further grouped into input – activities – output, show remarkably close agreement. The thermodynamics aspect is virtually coincident, the question of Kauffman's work cycle being held over to the 'fourth law of thermodynamics' discussion in Chapter 6. The information aspect, although more patchy, is hardly less in agreement. *In fact the overall models of Wicken and Kauffman barely differ.* Kauffman stresses the autonomous agent activity, but so does Wicken's 'AO' or autocatalytic organization ([10], p. 17); also both involve reproduction as an additional function.

Are these two aspects of thermodynamics and information competitive? No, for Wicken 'the informational dimension of life is incorporated with its thermodynamic identity, but not *conflated* with it' ([10], p. 31). In fact, an extended third quote from Wicken rather elegantly encompasses Table 4 ([10], p. 25). 'Complexity and entropy have complementary significances in the emergence and evolution of life. The production of one, and its dissipation to the sink of space, provides the driving force for the biosphere's complexification and generation of thermodynamic potential; the creation of the other through these negentropic processes provides the aperiodic, structured substrates from which natural selection can hone molecular information'.

The above leads naturally on to both the origin of life and Darwinian evolution.

11.3 Life's emergence

In their mainstream biological text *Life*, Purves *et al.* ([8], pp. 450–457) explain three conditions that any model of emergence should satisfy: continuity, signature and no-free-lunch. The first means that any stage 'should be derivable from pre-existing states', the second that traces should be apparent in contemporary biochemistry, and thirdly, that sources of energy must be explicit. The extent to which this latter biology is viewed in the same way as our preceding thermodynamics is apparent from their 'two long-term sources of free energy', 'radiation from the sun, and earth's magma', either or both of which 'could have powered the origin of life' ([8], p. 451).

Finally, Purves *et al.* make a telling point, that of *economy*: 'biochemical evolution has been remarkably conservative'. This tends to resonate with Wicken's rationale, which is that of *economy of model*. 'One of the projects of this book', he says ([10], p. 9), 'will be to show that variation and selection emerged as evolutionary principles at the *prebiotic* level, and *led* to Darwin's primordial organization. This extends rather than burdens the evolutionary project', and below '… this

Table 4: What is life? – a synthesis of models.

Model	Input	Activities	Output
Thermodynamics based			
Schrödinger [7]	Negative entropy	Avoids equilibrium	Heat rejection
Prigogine [7]	Free energy	(a) Far-from-equilibrium (b) Dissipative processes	
Wicken [10]	Free energy (p. 36)	(a) Remote from equilibrium (p. 17) (b) Dissipative structures (p. 31)	Entropy (p. 31)
Ji [36]	Free energy (pp. 5, 157)	Dissipative structures (p. 67)	
Kauffman [50]	Free energy (p. 29)		Work cycle (pp. 8, 72)
Information based			
Schrödinger	Aperiodic solids		
Prigogine	Genetic constraints maintaining f–f–e		
Wicken AO autocatalytic organization		Informed autocatalytic system (pp. 17, 31)	Reproducing (p. 32)
Ji	Shannon information (p. 1) Genetic material (p. 99)		
Kauffman AA autonomous agent		Autonomous agent (autocatalytic system) (pp. 8, 72)	Self-reproducing (p. 8)

book … attempts to provide *unifying* principles for evolution …'. Such seems a commendable approach.

We conclude this entrée by noting Ji's rather more speculative postulation based on the anthropic principle. It is that of the inevitability of life's emergence, and the concomitant unity of cosmological and biological information '… the cosmological information encoded in the initial conditions of this planet at the time of the origin of life might have been necessary and sufficient to cause living systems to evolve spontaneously; i.e. these initial conditions might have acted as a "cosmological DNA".' ([36], pp. 154/155). However, Ji's 'inevitability' is mirrored both by Kauffman and by Wicken: 'life is an expected, emergent property … autocatalytic sets of molecules suddenly become almost inevitable' ([50], p. 35) and 'The biosphere … necessarily provides niches in which AOs can emerge and evolve' ([10], p. 117).

A detailed comparison is now made of the models for emergence of Wicken and Kauffman. Tables 5–8 compare respectively their overall strategy, salient features, the AO with AA and

Table 5: Comparison of emergence models of Wicken and Kauffman – (i) overall strategy.

Wicken [10]	Kauffman [50]
Darwinism	
(a) To discuss the structure that *allowed* crossing the bridge of autonomy that has isolated evolutionary theory from the physical sciences (b) … a basis in physical law for the Darwinian principles of variation and selection (c) To show how those laws … prebiotic evolution … conditions for life's emergence (p. 131)	(a) …order can arise without the benefit of natural selection… (b) Self-organization mingles with natural selection… (c) We must, therefore, expand evolutionary theory (pp. 1, 2)
Thermodynamics	
(a) Emergence is a systematic movement away from thermodynamic equilibrium (b) The biosphere … necessarily provides niches in which AOs can emerge and evolve (pp. 116, 117)	The emergence of a metabolism that solves the thermodynamic problem of driving the rapid synthesis of molecular species above their equilibrium concentrations (p. 47)

their triple-cycle autocatalysis concepts. It is important to include Ji's triple-cycle model, the Princetonator, in the last comparison.

Table 5 shows that whereas Wicken has a conservatively Darwinian approach, Kauffman sees self-organization as an independent contributory factor in evolution. Despite this, in the context of their triple-cycle models, both see a *selective* drive. So for Kauffman, 'Darwin's natural selection could, in principle, operate if there were heritable variation in the kinetic constants' ([50], p. 71). Wicken is emphatic. Summarising ([10], p. 131), he says 'Selection was the central topic here' in the whole emergence story. Table 5 also shows there is a common thermodynamic core, despite the latter being a problem for Kauffman, compared with a kind of driving force for Wicken. Table 6 again shows good agreement. Both disagree with the primal replication approach – 'current life is not "nude" replicating DNA or RNA …' ([50], p. 25), 'less reasonable is the assumption … under the conditions of naked RNA competitions' ([10], p. 103).

When comes to hypercycles, Wicken classes them with the above problem, Kauffman accepting them. However, it is not completely clear whether Kauffman's acceptance is for the *prebiotic* circumstances as well as later evolution. Then, it must be admitted that there is no reflection of 'microspheres' in Kauffman's work, that is to say at the *equilibrium* level. Despite this, the end results of autocatalytic models (also see Table 7) and triple-cycle models (also see Table 8) are very similar. In Tables 7 and 8 material from Brooks and Wiley, and Ji are included, respectively.

In fact, the triple-cycle models of Wicken, Kauffman and Ji are strikingly coincident, and the conclusion must be that there is an underlying 'commonwealth' of thermodynamics. (A caveat must be that in constructing Table 8, it was not always clear that we could compare like with like). In fact, the authors' descriptions show their commitment to a thermodynamic/biological synthesis.

Space does not allow further discussion of the still complex issues involved – for instance, Wicken (p. 128) – 'we are still far from crossing the Kantian threshold to living matter. There are no *individuals* yet in this part of the epic …'.

Table 6: Comparison of emergence of Wicken and Kauffman – (ii) salient features.

	Wicken [10]	Kauffman [50]
Primal replication	Primal-replicator scenario paints itself into the … corner (p. 106)	Life need not be based on template replication at all (p. 32)
Hypercycles of Eigen and Schuster [53]	… treating replicators as … primordial objects of selection imposes a need … in hypercyclic couplings – a need that defies the rules of selection (p. 101)	… we have no trouble imagining a hypercycle of autonomous agents (p. 121)
Alternative to primal replication	A more realistic possibility is that life emerged through … coevolution … within catalytic microspheres (p.106)	This radical new view of life that I adhere to … not on template replication per se (p. 32)
Microspheres	… is *already* a dynamic entity, capable of growth, chemical catalysis, and reproduction … (p. 124) … are equilibrium systems (p. 125) … this attempt to combine the best in the microsphere and hypercycle models (p. 110)	–
Autocatalytic systems	AOs (pp. 31–32, 17)	AAs (pp. 8, 72)
Triple cycle models	figures 10-2 to 10-5 (pp. 127–128)	figure 3.4 (p. 65)

Table 7: Comparison of emergence models of Wicken and Kauffman – (iii) comparison of autocatalytic features, with references made by Brooks and Wiley.

	Wicken [10]	Kauffman [50]	Brooks and Wiley [25]
Autocatalytic	√ p. 31	√ p. 72	√ p. 80
Self-reproducing	√ p. 32	√ p. 8	√ p. 77
Non-equilibrium	√ pp. 116, 32	√ p. 8	√ p. 77
Dissipating structure	√ pp. 74, 75	–	√ p. 79
Expanding phase space	? the growth of microscopic information (p. 122)	√ adjacent possible (p. 47)	√ p. 77
Work output/energy storage	–	pp. 8, 72	–
Hydrophobic	√ p. 126	–	√ pp. 77, 79

Table 8: Comparison of emergence models of Wicken and Kauffman – (iv) comparison of triple cycle models, including the Princetonator.

	Wicken ([10], pp. 126–129)	Kauffman ([50], pp. 64–71)	Ji ([36] pp. 224/225) The Princetonator
1. Thermal cycle	Radiation in, photoreceptor in x ground state x*excited state	Photon source, electron in e ground state e*excited state	Solar radiation on day–night cycle
2. Phosphate cycle	Yes	Yes ('chemical engine')	Yes
3. Replication cycle	T nucleic acid, N abiotic protein (see Quote W)	DNA hexamer plus 2 trimers	A, B two kinds of biopolymers

Notes: Authors' descriptions.

Wicken: '… the fates of proteins and nucleic acids were bound together from the beginning – a primordial pair, yin to yang … the emergence of AOs based on their synergistic action was strongly selected for thermodynamically – and with it a translation mechanism. A minimal kinetic cycle based on these considerations is shown in figure 10-5' ([10], p. 128).

Kauffman: 'We measured efficiency thermodynamically as the conversion of available free energy coming into the system from the photon source into the excess hexamer, with respect to the undriven steady-state rate of reaction concentration of the hexamer' ([50], p. 69).

Ji: 'Clearly, the Princetonator provides a theoretically feasible molecular mechanism for biopolymer self-replication in the primordial soup that is driven solely by the solar radiation' ([36], p. 225).

11.4 Thermodynamics and Darwinian evolution

11.4.1 The models of Brooks and Wiley, Ji, and Kauffman

11.4.1.1 Introduction We have pointed out that the model used by Brooks and Wiley appears to be divergent from the developed classical thermodynamics model of Wicken. In that context, Wicken's approach may be regarded as *conservative*, both in his thermodynamics and his biology.

In this section we shall endeavour to show that: Brooks and Wiley's model is redolent of a fundamental aspect of cosmological thermodynamics, reflected in the models developed by Ji and by Kauffman; what Brooks and Wiley term entropy might be replaced by, we suggest, a neutral as yet unquantified term, say 'structure'; a reconciliation of the models of Brooks and Wiley and Wicken may be achieved by postulating the former's 'expanding phase space' as a *consequence* of thermal driving forces, and that thereby their biological analyses are unaffected.

11.4.1.2 The model of Brooks and Wiley Underlying the entire rationale of Brooks and Wiley is the postulation that the second law of thermodynamics is but one manifestation of a more general 'law of history' (p. 356), and 'entropy … a general manifestation of the passage of time indicated to an observer by time – dependent or irreversible, *processes of all kinds* (our italics).

All time-dependent processes, under this view, should exhibit entropic behaviour' (p. 355). So for Brooks and Wiley, the 'second law of thermodynamics' is replaced by the '(second) law of information' and their entropy is no longer 'thermal' but 'information capacity' (pp. 71–73). For those (most?) of us who never found the thermodynamic entropy concept immediately digestible as were, say, heat, work and energy, this seems rather confusing, and Brooks and Wiley, again candidly, admit 'their significant departure from classical thinking in thermodynamics' (p. 52).

The outcome is expressed graphically by the generic model of Fig. 14. H is the information-defined entropy, with H_{max} the information capacity, ever increasing with time in consonance with cosmological expansion – in other words expanding phase space. H_{obs} is the calculable Shannon information relevant to the specific biological application, termed complexity, as in Wicken's suggestion. Brooks and Wiley use versions of this graph about 16 times throughout the book. In one application, that of phylogeny, the graph is quantitatively defined (figures 4.20 and 4.22, pp. 245/248). In this instance, time is defined as number of speciation events, and H in bits of information.

Now, given the concept of increasing phase space (i.e. continuous expansion of the possible states a system can occupy) and ignoring the point that H is claimed as an entropy, this approach gives an integrated and convincing description of the various strategic aspects of biological evolution. Moreover, the 'adjacent possible' concept of Kauffman ([50], p. 47), is completely consistent with this. Most radical is the 'gnergy tetrahedron' of Ji ([36], pp. 160, 231, 234). We have already noted Kauffman's 'concatenation of matter, energy, information, and something more that we call life' ([50], p. 47). Ji combines just those items. 'This line of thinking' (that is of the 'primeval substance of the universe') 'led me to postulate that the universe originated from gnergy, the primeval substance thought to be composed of a complementary … union of four essential entities, namely *energy*, *matter*, *life* and *information*. … I propose to use the term "energy–matter–life–information tetrahedrality of gnergy" to indicate the notion that gnergy is

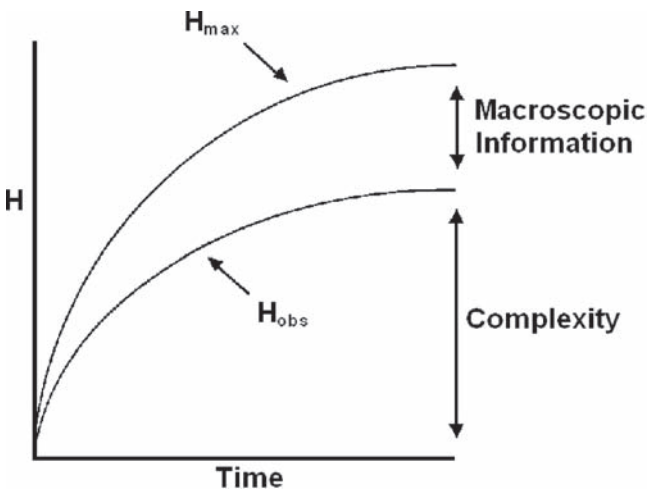


Figure 14: Generic evolutionary model of Brooks and Wiley – the relationship between macroscopic information and complexity of a physical information system under the Brooks and Wiley theory (redrawn from figure 2.3 with title [25], p. 41).

neither energy nor life nor matter nor information but can manifest such properties or entities under the right set of physical conditions' ([36], p. 231). The gnergy tetrahedron is shown in Fig. 15.

Such multi-manifestations have, to us, too high an academic alcohol content. However, a much more dilute version, namely a possible direct energy to information conversion, is of considerable interest, and with this we retrace our steps. It is unquestionable mainstream thinking, as we have reviewed elsewhere, that the second law of thermodynamics is fundamental to the structure and history of the universe. It is also and equally accepted that due to cosmic expansion there is an entropy gap, which must be filled by thermal dissipation. In our case this is provided by solar radiation to us (of lower entropy) and then from us into space (of higher entropy). So the cosmic expansion effects are mediated to the earth *thermally*. Isn't this just the grand 'energetic imperative' asked for by Brooks and Wiley ([25], p. 34)?

Now in consistency with their model Brooks and Wiley postulate that in prebiotic evolution 'monomer space' microstates, then 'polymer space' become available ([25], p. 77). At face value this is correct, and an earlier publication by Kauffman is quoted in support. However, we have already seen that Kauffman himself makes the same point that we do, and in the context of prebiotic evolution '… there is an overall loss of free energy that is ultimately supplied by the incoming photon … plus the 2 substrates. … Thus, we are not cheating the second law' ([50], p. 57).

We can now suggest a holistic reconciliation of models, which only requires setting aside any need for other entropic laws. So Brooks and Wiley's H-type entropy may be described in terms of order ('structure' is an as-yet unused example), and may properly express expanding phase space. This is now a *consequence* of free energy input to biological systems, and the Fig. 14

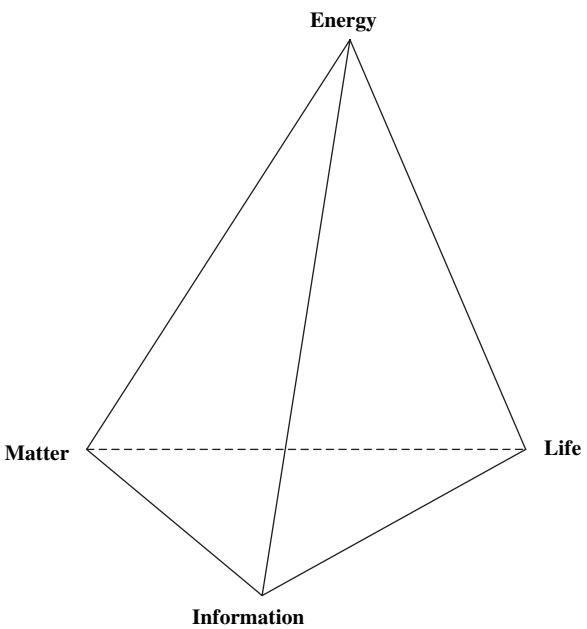


Figure 15: The gnergy tetrahedron of Sungchal Ji (see figure 1.A5 [36], p. 234).

type graphs, both *encompass* the biological applications and are consistent with the second law of thermodynamics.

11.4.2 An integrated survey of the work of Brooks and Wiley and of Wicken

> … the reader may proceed to the summary of this chapter and then on to chapter 3, returning to the more technical parts of this chapter later.
>
> Daniel Brooks and E.O. Wiley [29], p. 50
> (referring to the 50 remaining pages of their 'Core Hypothesis')

> The book in many ways is dense, both scientifically and philosophically. When encountering these densities. …
>
> Jeffrey Wicken, Preface [10], p. vi

It does not help the current authors' case that these two key books, even after reconciliation, are not easy reading. They contrast, for example, with Kauffman's publications in this regard. On the other hand, Brooks and Wiley and Wicken complement one other, the first being biologically oriented ('… a unified theory … in biology' ([29], p. 354)) and the second not ('… little specific biology is discussed…' ([10], p. vi)).

Much more important are their agendas related to natural selection.

For Brooks and Wiley 'Current evolutionary thinking does not reconcile biological evolutionary theory with the four areas of contention discussed in this chapter. Such a reconciliation has not been reached because evolutionists have tended to focus on natural selection as the primary organizing factor in biological evolution. … We will attempt to develop … a theory of biological evolution that unifies the empirical core of neo-Darwinism … with these four major areas of "unfinished business" ', ([35], pp. 29, 30).

For Wicken 'This book synthesizes a decade of my own work in extending the Darwinian program. … I think Darwin would have liked it. …' ([10], p. v).

In both instances then, they constructively interact with natural selection. It is fortunate that both provide intermediate Summary sections, and presuming that readers wish to have direct acquaintance with their work, the following guide is given as Table 9.

Table 9: Identification of summary pages for understanding approaches of Wicken and of Brooks and Wiley.

[A] Wicken [10]	Pages	[B] Brooks and Wiley [35]	Pages
Part I – What is Life?	15/16	Preface	ix–xiv
Part II – Connection	53/54	1. Prelude	29/30
Part III – Emergence	95–97	2. The core hypothesis	102–107
Part IV – Biological evolution	131–134	3. Ontogeny, morphology and evolution	173–175
Philosophic afterword	220–226	4. Populations and species	249–255
		5. Mapping historical change	287/288
		6. Macroecology	346–353
		Reprise and prelude	354–374

12 Conclusions

In this chapter an overview has been attempted of the relevance of thermodynamics to biology. Whereas the companion chapter in Volume 1 of the Design and Nature series focused primarily on energy, here the focus has been on entropy, information and complexity. It is shown that the concepts of free energy and exergy, which are equivalent and involve entropy, are necessary to a full understanding of engineering efficiency and to a thermodynamic model for living systems. It is demonstrated that, including thermodynamic variables, a mainstream interpretation of the evolution of the universe and the biosphere is emerging.

The relationship of entropy to information (Shannon entropy) and of both to complexity, are examined, and again a common interpretation is evident.

A number of authors have presented graphical models of the growth of information in the universe or the earth. These are compared, and again shown to be self-consistent. The concept of an 'arrow of time' of cosmic history is involved, and this stems from the second law of thermodynamics.

The key issues of what is meant by life, life's emergence and of Darwinian evolution are reviewed, focusing on the comprehensive studies of Wicken, Kauffman, Sungchal Ji and of Brooks and Wiley. Comparisons of the various models are made in some detail, and display a convincing underlying unity. With the exception of Brooks and Wiley, the use of thermodynamics by all authors could be described as conservative. However, Brooks and Wiley's model may be reconciled with the others by replacing their use of entropy with complexity, in the sense of consequential output, rather than input to living organisms.

The various authors include mathematicians, physicists, thermodynamicists and biologists. We conclude that their often very extensive studies point to a mainstream synthesis of biology and thermodynamics.

Acknowledgements

This chapter includes material adapted from a previous presentation by one of the authors, J.M.: Role of thermodynamics in science and engineering. *Archives of Thermodynamics*, **25(2)**, 2004.

The efforts of Miss Monika Turska (assistant to Prof. Jan Antoni Stasiek) in preparing this chapter is gratefully acknowledged.

References

[1] Mikielewicz, J., *Principles of Thermodynamics*, Zeszyty Naukowe IMP PAN 182/1098/84 (in Polish).

[2] Mikielewicz, J., *Lectures on Thermodynamics*, 1990–1993 (in Polish).

[3] Wisniewski, S., Staniszewski, B. & Szymanik, R., *Thermodynamics of Non-equilibrium Processes*, PWN Warszawa, 1973 (in Polish).

[4] Guminski, J., *Thermodynamics of Irreversible Processes*, PWN Warszawa, 1983 (in Polish).

[5] Schrödinger, E., *What is Life?*, Canto edn, Cambridge University Press: Cambridge, UK, 1992.

[6] Schrödinger, E., *What is Life*, and Introduction by Rǒger Penrose, The Folio Society: London, UK, 2000.

[7] Prigogine, I., Schrödinger and the riddle of life (Chapter 2). *Molecular Theories of Cell Life and Death*, ed. S. Ji, Rutgers University Press: New Brunswick, NJ, 1991.
[8] Purves, W.K., Sadava, D., Orians, G.H. & Heller, H.C., *Life*, 6th edn, Sinauer Associates/W.H. Freeman and Co. Ltd: New York, NY, 2001.
[9] Mikielewicz, J., Stasiek, J.A. & Collins, M.W., The laws of thermodynamics: cell energy transfer, *Nature and Design*, eds. M.W. Collins., M.A. Atherton & J.A. Bryant, Vol. 1, International Series on Design and Nature, pp. 29–62, WIT Press: Southampton, 2005.
[10] Wicken, J.S., *Evolution, Thermodynamics and Information*, Oxford University Press: UK, 1987.
[11] Rogers, G.F.C. & Mayhew, Y.R., *Engineering Thermodynamics Work and Heat Transfer*, 4th edn, Prentice Hall: Engelwood Cliffs, NJ, 1992.
[12] Szargut, J., *Engineering Thermodynamics*, Silesian Technical University: Gliwice, 1977 (in Polish).
[13] Szargut, J. & Zibik, A., *Foundations of Thermal Power Engineering*, PWN Warszawa, 1998 (in Polish).
[14] Kotas, T.J., *The Exergy Method of Thermal Plant Analysis*, Butterworth: Guildford, UK, 1985.
[15] Purves, W.K., Orians, G.H., & Heller, H.C., *Life*, 4th edn, Sinauer Associates/ W.H. Freeman and Co. Ltd: New York, NY, 1995.
[16] Chaisson, E.J., 'The cosmic environment for the growth of complexity biosystems', **46**, pp. 13–19, 1998.
[17] Davies, P., *The Fifth Miracle*, Penguin Books: London, UK, 1999.
[18] Collins, M.W. & Fleury, R.A., Analysis for nuclear power teaching. Part 2: Conventional and exergetic calculation methods. *Int. J. Mech. Eng. Edn.*, **18(2)**, pp. 131–141, 1990.
[19] Pool, R., *Nuclear whispers*. *IEE Power Engineer*, **18(4)**, p. 9, August/September 2004.
[20] Bilicki, Z., Mikielewicz, J. & Sieniutycz, St., Modern trends in thermodynamics, Institute of Fluid Flow Machinery, Gdańsk, 2001 (in Polish).
[21] Brillouin, L., *Science and Information Theory*, Academic Press: New York, NY, 1962.
[22] Peters, J., *Informations theorie*, Springer Verlag: Berlin, 1967.
[23] Thoma, J.U., 'Energy, Entropy and Information', International Inst. For Applied Systems Analysis, 2361, Laxenburg, Austria, Research Memor. RM–77–32, June 1977.
[24] Shannon, C. & Weaver, W., *The Mathematical Theory of Communication*, University of Illinois Press: Urbana, IL, 1949.
[25] Brooks, D.R. & Wiley, E.O., *Evolution as Entropy*, 2nd edn, University of Chicago Press: Chicago, IL, 1988.
[26] Arnheim, R., 'Entropy and Art. An Essay on disorder and order', University of California Press: Berkley, 1971. Adaptation of web version [http://acnet.pratt.edu/∼arch543p/ readings/Arnheim.html] August 2001.
[27] Goonatilake, S., *The Evolution of Information*, Pinter: London/New York, 1991.
[28] Boyden, S., *Western Civilization in Biological Perspective*, Clarendon Press: Oxford, 1987.
[29] Barrow, J.D., *The Artful Universe*, Clarendon Press: Oxford, UK, 1995.
[30] Tribus, M., *Thermostatics and Thermodynamics*, Van Nostrand: Princeton, NY, 1961.
[31] Fast, J.D., *Entropy*, 2nd edn, 2nd revision, Macmillan: London, UK, 1970.
[32] Winterbone, D.E., *Advanced Thermodynamics for Engineers*, Arnold: London, UK, 1977.
[33] Kondepudi, D. & Prigogine, I., *Modern Thermodynamics*, John Wiley & Sons: Chichester, UK, 1998.
[34] Dugdale, J.S., *Entropy and its Physical Meaning*, Taylor & Francis: London, UK, 1996.

[35] Tiezzi, E., *The End of Time*, WIT Press: Southampton, UK, 2003.
[36] Ji, S., 'Biocybernetics': a machine theory of biology (Chapter 1). Molecular Theories of Cell Life and Death, ed. S. Ji, Rutgers University Press: New Brunswick, NJ, 1991.
[37] Hawking, S., *The Universe in a Nutshell*, Bantam Press: London, UK, 2001.
[38] Gitt, W., 'Information: The Third Fundamental Quantity', *Siemens Review*, **56(6)**, pp. 36–41, November/December 1989.
[39] Gell-Mann, M., What is complexity? Home page www.stantafe.edu, reprinted from *Complexity*, **1(1)**, Wiley, 1995, downloaded 2004.
[40] Nicolis, G. & Prigogine, I., *Exploring Complexity*, Freeman: New York, NY, 1989.
[41] Lewin, R., *Complexity*, JMDent: London, UK, 1993.
[42] Kauffmann, S., *At Home in the Universe*, Oxford University Press: New York, NY 1995.
[43] Coping with complexity, p. 3, *EPSRC News*, UK, Summer 2003.
[44] Gribbin, J., *Deep Simplicity*, Allen Lane/Penguin: London, UK, 2004.
[45] Shalizi, C.R., Complexity Measures, cscs.umich.edu/∼crshalizi, 5 August 2003, downloaded 2004.
[46] Allen, P.M., Of fish and fishermen: models of complexity, *Maths. Today*, pp. 18–23, Feb. 2000.
[47] Rees, M., *Just Six Numbers*, Weidenfeld and Nicolson: London, UK, 1999.
[48] Hawking, S., *A Brief History of Time*, Bantam Books: Toronto, Canada, 1988.
[49] Lumsden, C.J. & Wilson E.O., *Genes, Mind and Culture: The Coevolutionary Process*, Harvard University Press: Cambridge, MA, 1981.
[50] Kauffman, S., *Investigations*, Oxford University Press: Oxford, UK, 2000.
[51] *The Complete Family Encyclopedia, Biosphere or Ecosphere*, Fraser Stewart Book Wholesale Ltd. Helicon: London, UK, 1992.
[52] Lovelock, J., *Gaia*, Reissue with new preface and corrections, Oxford University Press, Oxford, UK, 2000.
[53] Eiger, M. & Schuster P., *The Hypercycle: A Principle of Natural Self-Organization*, Springer: New York, NY, 1979.

This page intentionally left blank

Chapter 6

The laws of thermodynamics and *Homo sapiens* the engineer

M.W. Collins[1], J.A. Stasiek[2] & J. Mikielewicz[3]
[1]*School of Engineering and Design, Brunel University, Uxbridge, Middlesex, UK.*
[2]*Faculty of Mechanical Engineering, Gdansk University of Technology, Narutowicza, Gdansk, Poland.*
[3]*The Szewalski Institute of Fluid – Flow Machinery, Polish Academy of Sciences, Fiszera, Gdansk, Poland.*

Abstract

One of the key developments in the university scene since the publications of *The Origin of Species* has been the emergence of engineering as a discipline in its own right. Over the same timescale the realization has developed that the laws of thermodynamics, universal in their engineering scope, must also apply to biology and living systems. As a matter of historical fact, the relationship between biology and thermodynamics had a bad start, and the reasons for this are discussed. Now one of the foundation concepts for thermodynamics is that of the heat engine, stemming from the radical achievement of the Industrial Revolution of being able to produce mechanical work from heat. In this chapter, we find that by re-defining the heat engine in terms of output, all organisms can be viewed as 'survival engines', certain animal species and man as 'work engines', and man himself as a 'complexity engine'. A number of examples are taken from the process of building the dome of the cathedral at Florence by Brunelleschi, a defining moment in the history of the Renaissance. This interpretation of *Homo sapiens* is consistent with Lumsden and Wilson's assessment of man as being the only eucultural species: various consequences of this are discussed. The chapter completes our trilogy of studies on the laws of thermodynamics, by then focusing on the possibility of further laws for living organisms, especially Kauffman's proposal for a fourth law. Finally, a concluding discussion entitled 'How mathematical is biology?' highlights the question of including mathematics as part of that integration.

1 Introduction

1.1 General

This chapter completes our trilogy of studies addressing the relationship of thermodynamics with biology. In Chapter 5, our direct personal contributions were limited to topics from physics

and engineering. We now seek to apply the practices of *engineering* thermodynamics to living systems in general, focusing on the key concept of the *heat engine*. We will see that it is a quite straightforward approach, verging on simplicity.

Now a perceptive reader might well ask why this has not been attempted before. It possibly stems from two historical reasons: the originally almost non-existent status of engineering in UK universities and the initially bad relationship between biologists and energy specialists.

The ethos of the current volume and series rests on the 'parallel between human design and nature'. The former is self-evidently largely an engineering activity. Today engineering, including even the various branches of engineering, is a highly regarded university discipline in its own right. Because this was not always so, the idea that biology and engineering could have an advantageous dialogue was largely absent.

Much the same could be said of thermodynamics. In Kondepudi and Prigogine's historical introduction to the second law of thermodynamics, pride of place is given to James Watt, who 'obtained a patent for his modifications of Thomas Newcomen's steam engine in the year 1769', ([1], p. 67). Now in a popular encyclopaedia [2] Watt is explicitly described as a 'Scottish engineer'. Little did the Victorians of Darwin's day deem that thermodynamics had anything useful to say about biology.

With the benefit of hindsight, however, the signs were already there, with Clausius' famous summary of 'the two laws of thermodynamics …

> The energy of the universe is constant.
> The entropy of the universe approaches a maximum' ([1], p. 84).

If thermodynamics could cope with the universe, why not with the biological material generated in the universe?

1.2 The heat engine

Now one of the key concepts of engineering thermodynamics is that of the *heat engine*, an 'engine that performs mechanical work through the flow of heat' ([1], p. 69), and this was what Sadi Carnot studied in developing the origins of the second law. This is paralleled by the latest developments in second law interpretation originated by Prigogine and based on non-equilibrium thermodynamics. This allows 'a non-equilibrium system to evolve to an ordered state as a result of fluctuations' ([1], p. 426), arising from the dissipation of free energy. Such are called *dissipative structures.*

The above brief discussion epitomises Kondepudi and Prigogine's approach to the extent that the subtitle of [1] is 'From Heat Engines to Dissipative Structures'. Also, at a conference in 1986 in Piscataway, USA, Prigogine [3] discussed Schrödinger's *What is Life?* study of 1943 [4] – which he termed 'Schrödinger's beautiful book'. One of the focuses is that 'living matter escapes approach to equilibrium', and Prigogine notes that it was one of the sources for his own interest in non-equilibrium processes.

Now in the same conference volume as [3], Sungchal Ji [5] gives an exhaustive exposition of what he terms 'biocybernetics'. He treats a number of living systems, in the broadest sense of the word, as machines. These include the human body (called the Piscatawaytor) and human society (the Newbrunswickator). Such terms are an extrapolation, incidentally, of the original model in dissipative structure studies, the Brusselator (after its place of origin, the Brussels School of Thermodynamics [1], p. 439).

One of Ji's biological machines is the living cell, the model ([5], p. 80) being named the Bhopalator, from another conference at Bhopal, India, in 1983. A crucial component of this machine analogue of the cell is 'Dissipative Structures of Prigogine'.

We end this introduction by noting that Ji concentrates on the *internal* workings of his biological machines, whether, for example, they are living cells, the human body, or even human society. However, if instead we view such living systems as 'Black Boxes', then perhaps they can be interpreted as adapted versions of Heat Engines. In so doing we will be able to combine the earliest and the latest studies of the second law of thermodynamics.

2 Biology and thermodynamics: a bad start to the relationship

… now he lashed the techno – flunkeys from the temple – the 'Engineers … '

Adrian Desmond [6], p. 249

He deserves to be called a mathematician, as well as a physicist and even an engineer.

Denis Weaire [7], p. 57

I.K. BRUNEL ENGINEER 1859 (Monumental lettering on the Royal Albert Bridge)

R.C. Riley [8], p. 8

A bad relationship developed between biology and thermodynamics, following the publication of *The Origin of Species* (henceforth 'Origin'). It was intimately connected with the status of engineering and it focused on two main characters, T.H. Huxley and William Thomson (Lord Kelvin).

Firstly, we consider the status of engineering in the UK in the days of Darwin and T.H. Huxley. As far as universities went, it was poor. Whereas the former was happy to consult the Cambridge *mathematician* 'Professor Miller … this geometer' [9] on the structure of the honeycomb, Huxley saw *engineering* as a contamination of the purity of science. Desmond expresses Huxley's fear that the public veneration of science would switch to its products, 'its *engines* and telegraphs' ([6], p. 250). However, Huxley's attitude was ambivalent. 'Sir Joseph Whitworth put £100,000 into the DSA's science scholarships and joined the steel magnate Sir William Armstrong to fill Huxley's purse' writes Desmond ([6], p. 252). These were the same Sir Joseph Whitworth and Sir William Armstrong who were presidents of the Institution of Mechanical Engineers (IMechE) in 1856/57 and 1861/62, respectively; the IMechE as a learned society having been founded some 10 years earlier in 1847 [10]. Their engineering identities were similarly reflected at the Institution of Civil Engineers (ICE) – already 40 years old at the 'Origin's publication (1818 [10]). There Sir Joseph Whitworth served as Member of Council 1855/64 and 1870/87, with Sir William Armstrong again as president in 1881/82 [10]. Finally, the epitome of Victorian engineering, I.K. Brunel (similarly with the ICE as vice-president 1850–1859) in the very year of the 'Origin', completed his masterpiece railway bridge over the river Tamar. He wanted the world to know *he* was an engineer, so he told them so on his bridge. Huxley, at the same time as he was attacking the engineers 'would enjoy a wing in Whitworth's Matlock mansion', and 'annual holidays at Armstrong's Gothic manor' ([6], p. 252).

Furthermore, university engineering was not actually non-existent. Specifically, it was at Glasgow that 1840 saw the first British chair in 'civil engineering and mechanics' with the appointment of Lewis Gordon ([11], p. 35). We have already mentioned James Watt who became a 'Glasgow-legend' ([11], p. 34). Even as a civil engineer, however, Lewis Gordon's turbine water-wheel studies used a 'criterion for maximum economy' ([11], p. 35), derived from Lazare Carnot, the father of the second law originator, Sadi. A pattern was emerging – at Glasgow the

pioneering of engineering was intimately associated with Watt and thermodynamics. Gordon was followed in the engineering chair by Rankine – the thermodynamics specialist whose cycle is mentioned in [12] (p. 214). In turn, Rankine was succeeded by James Thomson in 1873 ([7], p. 57; [11], p. 104) – the brother of William Thomson, Lord Kelvin, our second principal character. Although Kelvin was appointed to the 'natural philosophy' chair at Glasgow in 1846 he embodied both thermodynamics and engineering. Even in his 'inaugural Latin dissertation' the earth's cooling was a focus ([11], p. 25), and Crosbie Smith describes his coming as 'a quest for engineering credibility' ([11], p. 4).

Kelvin was an immense figure in Victorian science, described by Bill Bryson as a 'kind of Victorian superman' ([13], p. 68) – or by Denis Weaire as 'mathematician … physicist and even an engineer'. He was anchored at Glasgow, declining offers to occupy the Cavendish chair at Cambridge ([7], p. 57). An exhaustive 800-page biography is also provided by Crosbie Smith, with Norton Wise [14]. So the stage was set for a scientific row of the grandest proportions when Darwin published a calculated age for the earth of over 300 million years in his first edition (removed in the 3rd edition due to the controversy ([13], p. 67)). At first, this difference was not too bad – 100 million years from Kelvin [15] (p. 370). Unfortunately, Kelvin kept revising his figures *downwards* ultimately, in 1897, 'to a mere 24 million years' ([13], p. 69). The argument was crucial for Darwin and Huxley, and the latter 'in his role of defence counsel' ([15], p. 370) was right on this issue. The problem was the lack of the effect of radioactive heating (then unknown) in Kelvin's model.

It is difficult for us to overestimate the effect that Kelvin's lack of success must have had on long-term biological/thermodynamics attitudes and relations. After all, in a crucial thermal subject area, the 'amateurs' proved to be right and the 'superman' wrong.

Looking at the broader issue of the assessment of Kelvin as an *engineer*, is a rather disappointing process. J.G. Crowther, while noting Kelvin's 'seventy engineering patents' and his 'theoretical and practical engineering ability' insists this was at the expense of his not achieving highest scientific success. Had he 'concentrated … on a few fundamental problems he might have become … the second Newton' ([16], p. 236). With the notable exception of [11], these days Kelvin is claimed essentially to be a physicist [17, 18].

Time does not permit us to discuss the wider issues of Crosbie Smith's convincing core thesis of the underlying battle between the scientific naturalism of the X-club centred in London, and the 'North British scientists of energy' ([11], p. 7 and 171). While 'Darwin's evolutionary theory' was the chief weapon, John Tyndall, a fellow-member with Huxley of the X-club, also introduced 'energy conservation'. '… he unleashed a massive extension of hostilities' ([11], p. 171). It focused on Tyndall's promulgation of the merits of Mayer 'the German physician' as opposed to those of Joule 'the English natural philosopher' ([11], p. 9) in the field of energy. In this equally strategic, but much less well-known dispute, it was the 'scientists of energy' who succeeded and the X-club who lost out.

It was no wonder, then, that biology and thermodynamics could not engage constructively at first.

3 The heat engine and the work engine

3.1 The heat engine re-visited

The idea of stressing the engineering focuses of thermodynamics is almost a tautology. However, we now repeat this, together with the point made in [8], that 'a careful progression of precise

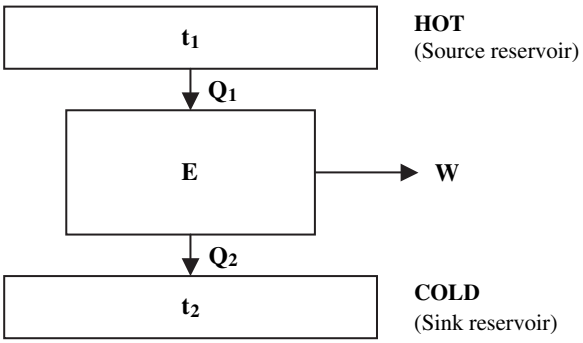


Figure 1: Heat engine diagram.

definitions' is crucial. We have already given Prigogine's definition of a heat engine. This could be extended to: 'a heat engine is a system operating continuously over a cycle, exchanging heat with thermal reservoirs and producing work'. This is shown in Fig. 1 (figure 8 of [12]), where the number of reservoirs is the conventional two. (Also, a direct heat engine is shown, but reversed heat engines for refrigerators and heat pumps are also possible – figure 9 of [12]).

We now propose a unified treatment of biological organisms as thermodynamic engines.

We will re-define the 'heat engine' in terms of *output* and identify three alternative modes:

(i) survival engine, relating to all species, including man (engineering equivalent, an idling heat engine);
(ii) work engine, relating to draught animals and man (engineering equivalent, a heat engine proper);
(iii) complexity (or information) engine, relating to man only (engineering equivalent, designing, computing, etc.).

We will examine to what extent the biological and engineering aspects are self-consistent, and whether the complexity engine concept can shed light on the entropy/information equivalence.

Finally, the focus on the *output* of such engines, allows the inclusion of pre-Industrial Revolution engineering in its scope. In fact, Brunelleschi's building of the dome of the cathedral at Florence will provide the majority of the examples used in this study. Information is derived from the recent publications of Ross King [19] and Paolo Galluzzi [20].

3.2 Internal combustion and the work engine

Although the heat engine is a core thermodynamic concept, it is not readily applicable to many real situations. In particular, internal combustion is an 'open system' or 'control volume' into which flow fuel and air, and out of which flow combustion products. As shown in Fig. 2 it may still be regarded as a black box. In practice, in thermodynamic analysis, it is conceptually replaced by an *external combustion* engine, with air as a working fluid. When, in addition the processes within the cylinder are re-defined by approximate reversible non-flow processes, a complete cycle is defined, and may be completely analysed thermodynamically. Such cycles are termed air standard cycles, and their efficiencies, air standard efficiencies.

In the case of spark-ignition engines, locomotion from which we will compare with that from the horse, the air standard cycle is called an Otto cycle.

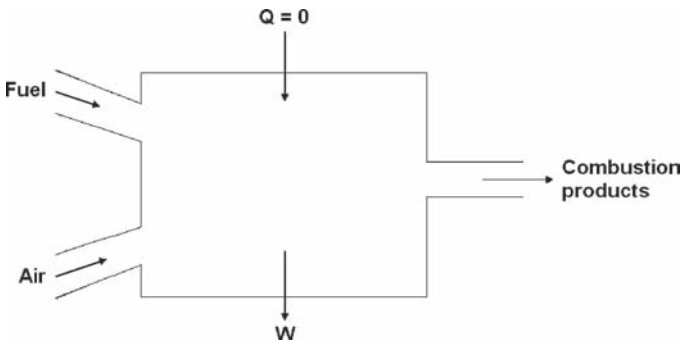
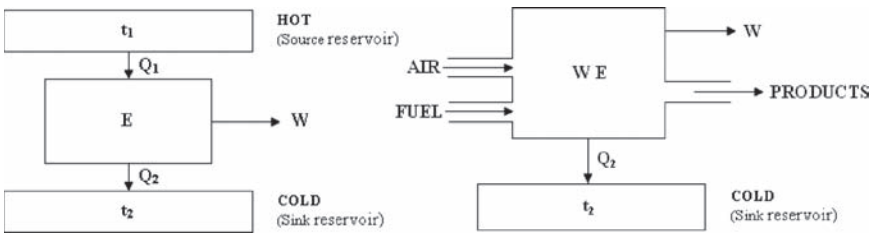
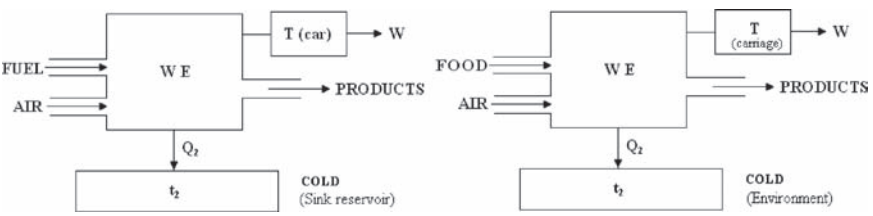


Figure 2: Diagram of internal combustion engine (after [9]).



(a) Heat Engine diagram (b) Work Engine diagram

Figure 3: The heat engine and the work engine.



(a) By Internal Combustion (b) By Horse

Figure 4: Locomotion by internal combustion and the horse.

Figure 3 shows how the heat engine may be replaced in terms of *output* from an *open system* by what we term a *work engine*. The latter is defined with as much thermodynamic flexibility as possible.

In the case of the spark-ignition internal combustion engine, the black box E represents an air standard Otto cycle device. For the work engine, the new black box WE represents an authentic internal combustion cylinder–valve–piston arrangement. There is still a heat rejection from WE to the environment, as it operates at a temperature well above ambient.

3.3 Locomotion by car and horse

We may now compare locomotion by mechanical and biological means, as in Fig. 4.

The equivalence is extremely close, the fuel and products of the internal combustion engine being replaced by food and waste in the case of the horse. For multiple passengers, in both cases

an additional tool T is required – either the car proper or a carriage. Also, for each arrangement, a form of intelligent control by man is needed.

In the history of thermodynamics, this logic was used in reverse, to such an extent that the horse power was used as an unit of performance. James Watts is credited with its introduction ([11], p. 198), 1 horse power being 550 ft · lb (force) of work per second in the old British Imperial system of units, or 746 watts.

3.4 Other draught animals: the ox

While the above comparison is rather precise in the case of horse-drawn locomotion, the work engine concept also applies to any draught animal including elephant, camel or even man himself. In particular, in agriculture there was an ongoing competitive use of horse or ox in Europe over the large time span of 1100–1800 AD. While in Britain the number of horses increased in agricultural and draught work over this period, in southern Europe the donkey and mule were preferred to the horse – the ox being common to both ([21], pp. 154/155).

4 The survival engine: e.g. the lizard

Thermodynamically speaking (and not considering the food chain), most species do not produce work: they just survive. Hence they, as survival engines, approximate to the idling heat engine which consumes fuel, which is ultimately entirely lost as a heat rejection. Such an engine, of course, has zero work output.

We take as an example an *ectotherm*. This animal has, as an integral part of its metabolism, an external source of heat, which, in the case of the lizard means solar radiation. Figure 5 demonstrates the thermal regulation that the lizard has to undertake over the daytime period.

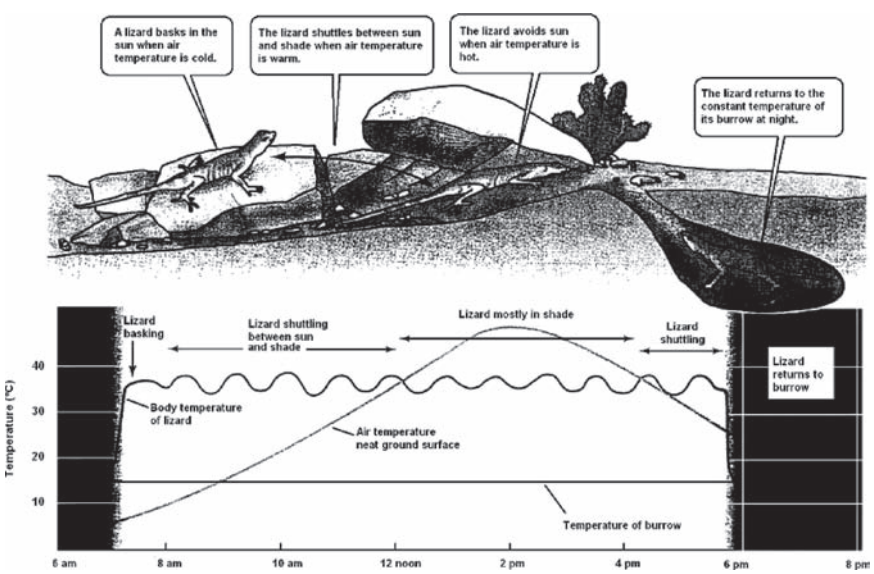
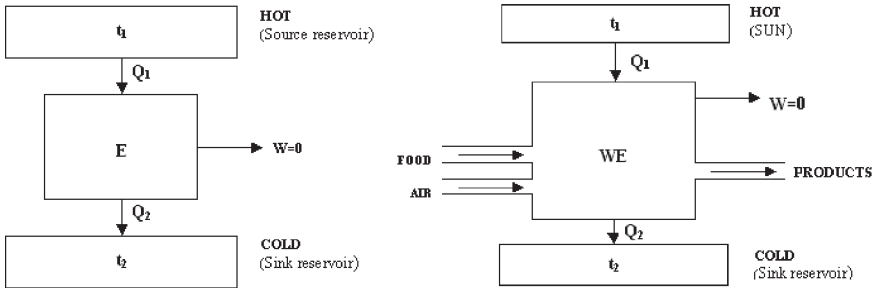


Figure 5: Thermal regulation of an *ectotherm* (after [22]).



(a) Idling Heat Engine (b) Survival Engine *(Ectothrem)*

Figure 6: An engine with no work output: mechanical and biological.

As an *ectotherm*, the lizard exemplifies an even closer comparison with the idling heat engine, as solar heating (i.e. using a hot source reservoir) is an integral part of its metabolism. Figure 6 compares the two.

5 Work engines and the dome of the Florence Cathedral

5.1 The dome

The building of the dome of the cathedral at Florence in Italy was a massive undertaking, and a defining moment in the history of the Renaissance. It took 16 years to build, from 1420–1436 ([10], p. 141). The name of the architect and builder, Filippo Brunelleschi, is synonymous with that of the dome, and it seems incredible that previously he was 'a goldsmith and clockmaker' (p. 11). The achievement was indeed impressive. The base and top of the dome were around 54.5 and 87.5 m above ground level, the lantern adding a further 22 m. Its weight has been estimated at 37,000 metric tons, with probably more than 4,000,000 bricks used. The key design feature is that it is the largest dome ever built without having to use wooden centring.

5.2 The rota magna or treadmill

From Roman times, the device used for major work output of the kind envisaged in lifting the materials for the dome was the rota magna or treadmill, described in detail in [23] (pp. 11–13). Presumably, the Romans would have used slave power. In fact, the rota magna had been used for the construction of the cathedral proper.

Our work engine definition accommodates the rota magna, as shown in Fig. 7. The work engine itself comprises *Homo sapiens*, the tool being the treadmill/rota magna mechanism.

5.3 Brunelleschi's ox-hoist

As part of the competition for the building of the dome, the Opera del Duomo (the office of works in charge of the cathedral [19], p. 1) had included a call for plans for a lifting device that could replace the rota magna. None were forthcoming (p. 59). Having been appointed as one of the three *capomaestri* (architects-in-chief) (pp. 6/59), Brunelleschi almost immediately started designing what was to prove a completely successful ox-powered hoist. His sheer genius was displayed, not just in the size and power output, but especially in the hoist's reversible three-speed gearing

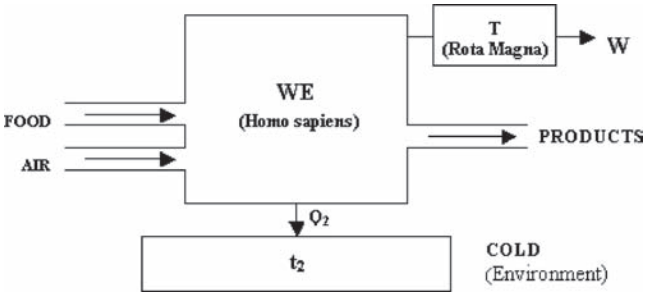


Figure 7: The rota magna.

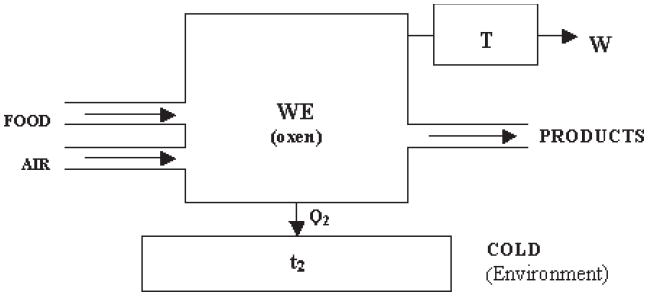


Figure 8: The ox-hoist.

'for which there is no known precedent in the history of engineering' (p. 62). Others followed his example. Taccola (Mariana di Iacopo) designed a single-speed horse-powered version ([19], p. 61; [20], figure I.2B.1b). Also, Leonardo da Vinci analysed the whole 'motor' in detail, from which a working model has been constructed ([20], figure I.2B.1a).

It seems unbelievable that such a revolutionary engineering design was conceived (using Brunelleschi's words) as a relaxation – 'relief … at night … in my mind … from bitter worries and sad thoughts' ([19], p. 58).

Figure 8 shows the ox-hoist as a diagram, the work engine being the ox or horse, and the tool the ox-hoist mechanism. From Ross King's quoted figures ([19], p. 63), one ox achieved a work output – including all the mechanical losses in the hoist itself – of about 15,500 ft · lb force/minute. This compared with the notional Horse Power unit introduced by Watt, of 33,000 ft · lb-force/minute.

5.4 Brunelleschi's revolving crane or castello

Brunelleschi also designed a revolving crane, probably used to position accurately the final stone blocks and to build the lantern. This was an impressive mechanism, at least 20 m high, and it must have dominated the Florence 'skyline' of the dome when it was erected at the apex.

This was powered by *Homo sapiens*. Four teams were needed. The first rotated the vertical shaft, which had a complete angular flexibility of 360°. Two more teams worked the screws for radial positioning of both the load and the counterweight which was needed. The fourth team operated the vertical screw. Again Leonardo da Vinci sketched the mechanism in detail, and again a working model has been constructed ([20], figure I.2B.2a).

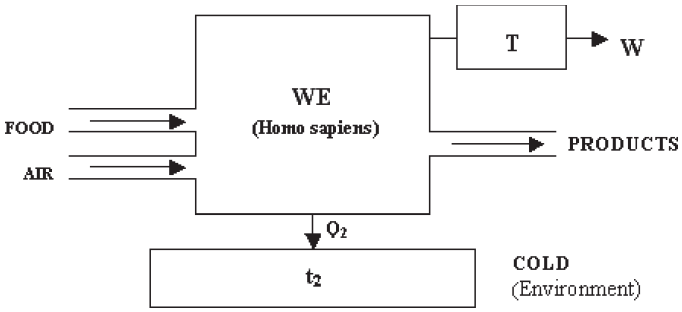


Figure 9: The revolving crane.

Figure 9 gives the thermodynamic diagram, with the four teams as the work engine and the castello mechanism as the tool.

6 Brunelleschi, the complexity engine

6.1 The ox-hoist

The new hoist was conceptually designed and built over a comparatively short period. The information content of an engineering design has been postulated as the entire specification necessary for manufacture. Considering complexity, it should also include, arguably, the performance of the designed object, such as degrees of freedom. For our purposes, it is not essential to quantify complexity, except to note it *must* include Shannon information.

Over this period of time, then, Brunelleschi had – as a biological/*Homo sapiens* 'black box' – a thermodynamic input of food and oxygen. His brain, as a sort of computer hardware analogue, we could assume conservatively was in a similar state at the end – when the hoist was made – as at the beginning (although his skill had gone up). As information (complexity) input, there would have been the rota magna design and the hoist requirements – what we would term a 'design specification'. As information (complexity) output, there was the entire design and manufacture of the hoist.

So, in this case, Brunelleschi – the *Homo sapiens* – acted neither as a work engine nor a survival engine, but rather as a complexity engine. He converted the (Gibbs) free energy in the food into the increase in complexity represented by the ox-hoist design. Figure 10 gives the appropriate postulated engine diagram.

Exactly the same logic may be used for the castello design.

6.2 The dome

The design and construction of the dome is universally recognised as a masterpiece of engineering and architectural genius. It was largely due to having two shells, a photograph of the intervening air gap, staircase and special brick arrangement being shown as a colour plate in [19]. To use Ross King's words ([19], p. 167), 'Today, as for the past five centuries, the mountainous form of the cupola dominates Florence … . The fact that it was built by men – and built amid war and intrigue, with only a limited understanding of the forces of nature – only makes it more of a wonder.'

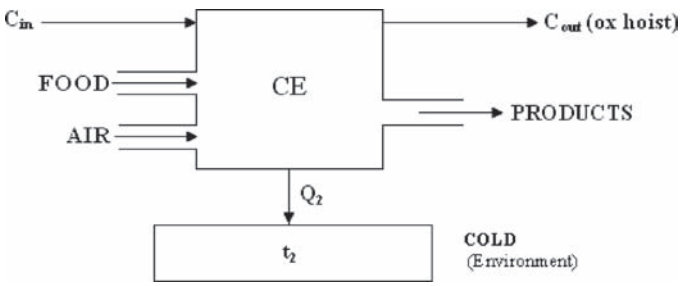


Figure 10: Complexity engine diagram (few years timescale) for Brunelleschi – designer of ox-hoist.

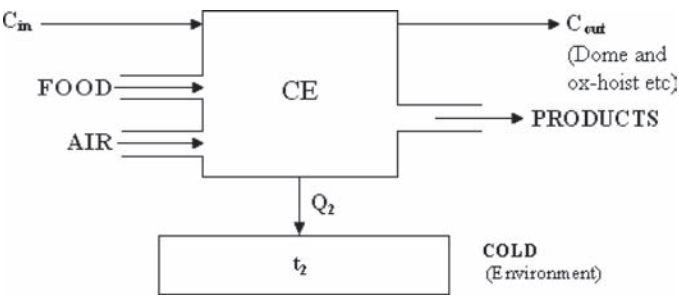


Figure 11: Complexity engine diagram (lifetime) for Brunelleschi – designer and builder of the dome.

6.3 The dome was Brunelleschi's overall achievement

Together with the mechanical devices, it has a huge information (complexity) content, which in principle could be quantified. For a complexity balance on Brunelleschi's lifetime (and ignoring input from his assistants and workforce) we must include both:

The genetic complexity content of his brain
The complexity content of his training/education

Apart from that, Brunelleschi, like all other *Homo sapiens*, had essentially food and air as input. As with the hoist and castello, Brunelleschi the engineer, as a far-from-equilibrium organism, converted the (Gibbs) free energy into complexity. He was a supreme complexity engine. Finally, financial resources are implied – Brunelleschi's food was found for him. Figure 11 shows the corresponding lifetime engine diagram, appropriate to the entire enterprise.

7 Some consequences for *Homo sapiens*

7.1 Man the engineer

In the preceding section, we have shown, essentially at the undergraduate level, how a re-defined thermodynamic heat engine can be applied to living organisms. The free energy/heat rejection

model is based on the original Schrödinger *What is Life?* study as re-interpreted by Prigogine [3]. *Homo sapiens* emerges as a complexity engine par excellence. This is not to deny a complexity engine interpretation of the outputs, say of spiders, of beavers and of the tool-making activities of various species. However, in engineering terms there seems to be no contest, reflecting the one eucultural group of Lumsden and Wilson [24], illustrated by Goonatilake [25] (figure 4.3).

The above, of course, is a reflection of the complexity of *Homo sapiens*. Despite the scientific language of the following quotes, they read almost like poetry.

> By far the most complex systems that we have are our own bodies.
>
> Stephen Hawking [26], p. 161

> The human body is the most complex material system known to us. It consists of $10^{12}-10^{14}$ cells that are organized in space and time (i.e., different parts of the body have different cells and cells at a given site in the body change their properties with time).
>
> Sungchal Ji [5], p. 141

> More than one hundred trillion (10^{14}) cells make up an adult human. As a fertilized human egg – a single cell – develops into a university student, its nucleus gives rise to over one hundred trillion nuclei, each containing basically, the same genetic information as did the fertilized egg.
>
> Purves *et al.* [22], p. 191

> To build the most basic yeast cell … you would have to miniaturize about the same number of components as are found in a Boeing 777 jetliner and fit them into a sphere just 5 microns across … but yeast cells are as nothing compared with human cells.
>
> Bill Bryson [13], pp. 329/330

Also, this interpretation is consistent with Chaisson's work [27], already discussed in Chapter 5 in some detail. His table 1, we are reminded, gives computed free energy rate densities (in ergs/cm^3) for, in particular, brains (human crania) and society (modern culture), together with the developmental timescale. '*Homo sapiens* the engineer' is an expression of the progress from brains to society, with Brunelleschi a notable instance over a very short timescale.

This consideration of complexity output, then, consistent with the complex structure of *Homo sapiens*, tends to stress the *distinctive nature* of our species. Even given this distinctiveness Brunelleschi demonstrated himself to be an outstanding example of 'Man the engineer'. His epitaph uses the medieval Latin expression 'INGENII' related to the building of machines. 'It refers to him, therefore, not directly as an architect but as a man of mechanical genius, alluding to the machines he invented in order to raise the dome' ([19], p. 156).

The above distinction of *Homo sapiens* is possibly related to Sungchal Ji's studies [5]. Whereas our focus is on the processes *external* to the human body as an open system, in his 'biocybernetics' Ji studies the *internal* aspects. His Piscatawaytor, redrawn as Fig. 12, demonstrates how complementary the two approaches are.

Ji concludes that '*voluntary bodily motions*, including thought processes' are 'at the centre of our biological being'. He then poses two questions – Is it possible that there is some deep philosophical significance to this conclusion? Have we underestimated the fundamental biological and evolutionary significance of our voluntary bodily motions?

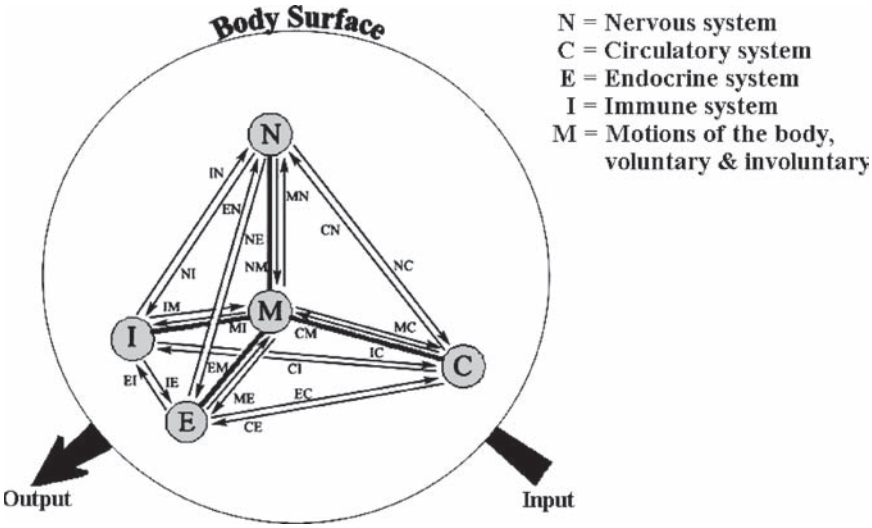


Figure 12: Ji's biocybernetic model of the human body – the Piscatawaytor.

The trillions of cells constituting our body can be divided roughly into 5 groups: the nervous (N), circulatory (C), endocrine (E), immune (I) and voluntary muscular (M) systems. The N, C, E and I systems are thought to act as control systems that cooperate to enable the human body to carry out voluntary motions, M, both macroscopic (i.e. bodily movements) and microscopic (i.e. thought processes).

Sungchal Ji [5], p. 144

7.2 Should there be a biological/engineering synthesis? The case of locomotion

The complexity outputs of species other than man have already been noted. Studies of such are regarded as biology. Consistently, since man is also a biological being, his engineering achievements could also, then, be regarded as biological achievement. This is especially telling in the case of *locomotion*. Bill Bryson, in his usual graphic manner, points out that man is inherently poor at adaptation – 'pretty amazingly useless' ([13], p. 217). Focusing on flight, the best we do is jump some few metres above the ground surface! And yet, with the invention of powered flight (let alone space exploration) hundreds of men and women can at one go outfly all other species.

So in the context of flight man is simultaneously very inadaptable and highly adaptable. The reasons for this mismatch must have a thermodynamic background, for the calorific value of typical foods and typical fuels are of the same order of magnitude [28]. As may be inferred from Table 2 of Chapter 5 on page 140, the cell energy processes are complex. There are a number of reasons for this ([22], pp. 125/126). Firstly, molecules are produced 'that have other roles in the cell', and secondly the total energy release is split by 'a *series* of reactions, each releasing a small manageable amount of energy'. As a consequence, the organism temperature is not too high above ambient, which means the 'heat rejection' is correspondingly reduced. Thermal regulation is very precise.

However, second law considerations require a heat engine to operate at a temperature as high as possible for maximum efficiency, and this means the thermodynamic efficiency of *Homo sapiens* is very low. In contrast, powered flight requires as high an energy *density* as possible – kilowatts per cubic metre (or size considerations) and kilowatts per kilogram (or weight considerations)

([12], p. 249) – and uses internal combustion at high temperatures. The levels of calorific value (either fuels or food) both permit powered flight on the one hand, and minimise the bulk of food intake (and the fraction of *Homo sapiens*' time necessary for eating) on the other.

Strictly, it could be argued that there is not a mismatch. The adaptability of *Homo sapiens*, such as it is, is good, almost instantaneous, on what could be described as the microscale. Noting in passing that they were preceded by the 'purely' biological, equine-based locomotion, invented forms of land-, water- and air-based locomotion give man a very good adaptability on, say, the macroscale. In our overall interpretation of *Homo sapiens*, the point of this discussion is to suggest that the two could be *synthesised*. It is allied to the questions posed above by Ji.

7.3 Is *Homo sapiens* just a machine?

This chapter essentially addresses the thermodynamic aspects of living systems in general and *Homo sapiens* in particular. Immediately above we have studied the concept of man as a complexity engine and noted that our concept is complementary to Ji's rationale of various internal biocybernetics models. The reader might well ask whether on these grounds we should regard *Homo sapiens* as *just* a machine, whether thermodynamic or biocybernetic.

The mainstream thermodynamic sequence may be summarised as follows. Kondepudi and Prigogine present two key concepts ([1], p. 427), 'order through fluctuations' and 'dissipative structures'. The first is a general chaotic phenomenon, where random fluctuations cause the system to evolve to one of a number of possible states, which in general is not predictable. In such non-equilibrium systems, chemical reactions/diffusion/concentration patterns are due to dissipative processes. Both are an expression of the consumption of free energy as input to the system. (We have used in the above a number of the specific expressions in [1] (p. 427).) In the model of the cell known as the Bhopalator ([5], p. 80), Intracellular Dissipative Structures (IDSs) as proposed by Prigogine figure prominently. Ji ([5], pp. 71/72)discusses space-dependent and time-dependent IDSs, leading to a 'bifurcation tree for cell metabolism', consistent with the above chaos description. Further, the application of this to variations in calcium concentration in the cytoskeleton is completely consistent with the model Webster and Goodwin use [29] (p. 200) in explaining the morphogenesis of the cap of *Acetabularia*.

However, it is the thermodynamic light that this throws on the sociobiological debate that is of interest here. Ji focuses on the implications of these 'deterministically chaotic' attributes of the Piscatawaytor, agreeing both with E.O. Wilson's stress on the crucial genetic role in social behaviour [30] and with Wilson's *opponents* who introduce the necessity of cultural influences ([31], p. 148). Ji optimistically believes that the model can 'begin to unite such … opinions into a harmonious and synthetic view of humankind'. He points out that the great sensitivity of chaotic processes to the input boundary condition decouples the predictability so that development can be controlled mainly by cultural influences, '*largely unaffected by individual genetic differences*'. Further, the *control* aspects of the Piscatawaytor (figure 21) suggest to Ji, that 'individuals can exhibit … egotism, altruism, sex drive, hate, love etc. irrespective of educational or cultural background'.

Wicken concludes *his* thermodynamic study ([32], p. 244) under the subheading 'Chance, Necessity and Freedom'. Like Ji, he has a focus on control. After confessing his ignorance of a number of issues such as 'sense, perception and subjectivity', he asserts 'I do know that *decisions* are operative ingredients in biological nature'. As with Ji, E.O. Wilson appears in the discussion. Wicken contrasts Wilson's confidence in a complete scientific explanation (with reference to a quote from Job in the Old Testament) [31], with his own attitude. 'Wilson feels that science has affirmatively answered this challenge. I do not.' Again reflecting Ji 'The genes – first motif misrepresents the rationale of evolution for many *scientific* reasons discussed over this book'.

Goonatilake [25] also addresses the gene–culture issue in his study of thermodynamics/information. So, on p. 42 and following: 'there are two types of information … in cultural animals … the genetic line and … the cultural line which is non-genetic … there are mutual influences between the two … '. Goonatilake does not see conflict with E.O. Wilson, whose work is extensively discussed in pp. 25–31.

Finally, there are the rather more discursive publications of Ridley [33] and Tiezzi [34]. Ridley's is a gene–focused book considering only the nature/nurture issue. Thermodynamics and entropy do not figure. Culture, however, is addressed and Ridley confirms our conclusion about the distinctiveness of *Homo sapiens*. 'The cultural gap between a human being and even the brightest ape or dolphin is a gulf' (p. 209). He concludes that gene adaptation – 'genes that allow culture' (p. 228) – occurred for culture to be initiated. E.O. Wilson appears again (pp. 236, 241–246) and is assessed fairly comprehensively. Ridley is a reconciler – 'Nature versus nurture is dead. Long live nature via nurture' is his overall conclusion (p. 280).

Tiezzi on the other hand, is wedded to thermodynamics, embracing Prigoginism. Entropy is essential – 'a modern scientific culture cannot do without entropy and Darwin's theory of evolution. Besides entropy and evolution have much in common' (pp. 36/37). Again we have a discussion of Wilson – Wilson seems to be *de rigeur* – with a stress on his and other sociobiologists' reductionism (p. 53). Tiezzi is explicitly a non-reductionist – 'man, unlike animals, has a socio-cultural history', the future needing 'scientific humanism and ecological culture … the only hypothesis suitable for the species *Homo sapiens*' (p. 163).

So, by introducing thermodynamics, we seem to be consistently led to a conclusion that man is *not* just a machine. In addition, we feel the method is even more important than the conclusion – thermodynamics and entropy must always be on the guest list for any scientific banquet.

8 Is there a fourth law of thermodynamics?

8.1 Kauffman's statement

> … a First Law of ecosystem dynamics, analogous to the First Law of Thermodynamics …. The competition among alternative patterns of energy flow for resources (natural selection) is a Second Law of ecosystem dynamics.
>
> Jeffrey Wicken [32], p. 141

> … my hoped-for fourth law of thermodynamics for open self-constructing systems.
>
> Stuart Kauffman [35], p. 84

In this chapter, we have applied formal engineering thermodynamics to living systems. Particularly in the case of locomotion, the definition of 'work engine' allows animal-based and mechanical machines to be treated in the same way. Others have taken this logic further to affect the laws of thermodynamics themselves. At one end of the debate is the generalisation of the second law of thermodynamics into 'the natural law of history', Brooks and Wiley [36] (p. 356). At the other end is the idea of *additional* laws. So we have Wicken's two laws of ecosystem dynamics, notably for *natural selection* itself. Also, we have Kauffman's four 'candidate laws for the co-construction of a biosphere' ([35], p. 161). These latter laws, for the evolution of 'co-evolutionarily constructible communities of molecular autonomous agents' (p. 162) overlap the area of the laws postulated by Wicken above. Wicken's statement may be compared with, say, Kauffman's figure 8.12 ([35], p. 191).

However, it is Candidate Law 4 that Kauffman generalizes into a fourth law of thermodynamics. This former is stated as ([35], p. 160):

> Autonomous agents will evolve such that causally local communities are on a generalized 'subcritical–supracritical boundary' exhibiting a generalized self-organized critical average for the sustained expansion of the adjacent possible of the effective phase space of the community.

This statement becomes a postulated fourth law for the biosphere, which Kauffman summarises as (p. 207):

> We enter the adjacent possible, hence expand the workspace of our biosphere, on average, as fast as we can.

The context of Kauffman's statement is his other candidate laws. These address the biological hierarchy of organisms–to–species–to–community, and in so doing are comparable to the areas of interest of Wicken and of Brooks and Wiley. The respective quotes below make this comparison explicit.

> All autocatalytic organizations – from organisms to ecosystems to socio-economic systems – have two referents: the survival and propagation of certain organizational types, and the provision of the biosphere with patterns of entropy production.
>
> Jeffrey Wicken [32], p. 144

> Ecological associations and communities represent higher levels of biological organization above that of species.
>
> Daniel Brooks and E.O. Wiley [36], p. 312

> A coassembling community of agents … will assemble to a self-organized critical state with some maximum number of species per community … .
>
> Stuart Kauffman [35], Law 2, p. 160

In brief then, and in bridging the gap from organisms to biosphere, we see a desire for new biologically – oriented statements arising from the thermodynamic foundations – four candidate laws of Kauffman, two analogous laws of Wicken, and a single generalised version of the second law of thermodynamics, of Brooks and Wiley.

8.2 Discussion

We now consider a number of issues raised by Kauffman. Firstly, coincident with his overall concept is the importance of the direction of time. 'Some fourth law of thermodynamics? An arrow of time?' ([35], p. 151). Although Kauffman claims this is in 'sharp contrast to the familiar idea' (p. 151 and p. 48), his ideas are quite consistent with those of the authors reviewed in the relevant section in Chapter 5. Secondly, his emphasis on the 'adjacent possible of the effective phase space', is parallel to the expanding phase space rationale of Brooks and Wiley. Thirdly, and more subtly, Kauffman views the biosphere as an open system (e.g. pp. 2, 3). This is contra-Wicken, who specifically defines it to *include* all material transport. From choice, it is much easier to use a closed system, but there is a still more significant issue.This is Wicken's degree of closure concept. 'As open systems, organisms fulfil their thermodynamic destinies by

fitting into higher-order systems that express greater degrees of closure' ([32], p. 146). In other words, in parallel with the ascent up the biological hierarchy is a matching increase in the thermodynamic closure. As far as we know, the possible biological significance of this has not yet been addressed.

Finally, Kauffman expresses what may be described as a thermodynamic imperative. It is not just that the 'workspace of the biosphere expands' it does it 'on average, *as fast as it can*' (p. 209), or 'we … expand the workspace … on the average, *as fast as we can*' (p. 207), or subject to selection constraints 'a biosphere expands … about *as fast as it can get away with*' (p. 155). This is caused thermodynamically. '… sunlight shining on our globe, plus some fussing around by lots of critters, has *persistently exploded the molecular diversity of the biosphere into its chemically adjacent possible*' (p. 47). 'Persistence' is a theme word of Kauffman's – we find it on pp. 48, 82, 93, 143, 151, for example. All this adds up to what could be described as a *driven* biosphere – driven by the low entropy free energy from the sun.

While Kauffman's real interest is in the *consequential* effect of complexity or order – 'Order for Free' [35] – he properly connects it with the thermodynamic causes. His 'small print' at the end of Chapter 4 makes this clear: '… while I have called it order for free … it is not "for free" thermodynamically' (p. 92). This leads us on to an intriguing interpretation of (biological) evolution away from the idea of accident or marginality, and towards the concept of a synthesized biothermodynamic enterprise.

We suggest that this could be described as a 'thermodynamic imperative', inspired by a comment made by Brooks and Wiley. 'We do not assert that energy flow is trivial', they admit (p. 34), 'only that there is no external energetic imperative behind organismic diversification'. However, it is our belief that this decoupling of biology from 'energetic' thermodynamics is not only unnecessary but also difficult to justify, particularly as, like Kauffman, they make the connection with free energy. In this regard we read (their p. 48), 'if free energy is used to maintain, transmit and express biological information, we assume that information must be subject to thermodynamic constraints and *expectations*' (our italics). In so saying, Brooks and Wiley point towards the very imperative they deny. We have previously suggested that by re-interpreting their 'entropy' as 'complexity', their whole approach can be broadly made consistent both with their fellow biologist Kauffman and the thermodynamicist Wicken. Further, we conclude, that a thermodynamic imperative may indeed be postulated.

8.3 *Homo sapiens* the engineer

We have considerable empathy with Kauffman's approach. Our rationale is somewhat different and possible complementary. We have endeavoured to be as conservative as possible in applying engineering thermodynamics to biological organisms, by using the whole logic of the industrial revolution in reverse. So the 'closed system heat engine' is given an 'open cycle thermodynamic engine' partner applicable to both engineering and biological systems. This is seen at its clearest in comparing engineering locomotion with draught – animal locomotion. Both use free energy as input, both provide work as output, so being termed 'work engines'. Most species do no *net useful* work and are thus termed 'survival engines'. Kauffman's bacterium ([35], pp. 7, 8), while generating thermodynamic work cycles with its flagellar motor, is doing it exclusively within the food chain. Our distinction between this (technically correct) work and net useful work is also implied by Kauffman in the statement 'sense of "useful" outside the context of autonomous agents' ([35], p. 91). Like Kauffman, Wicken, and Brooks and Wiley we have focused on information, but in our case this is exosomatic (to use Goonatilake's descriptor) or Shannon

information/complexity (after Wicken). Hence *Homo sapiens* becomes a 'complexity engine' par excellence or by further using eucultural in addition (after Lumsden and Wilson/Goonatilake ([25], p. 31)) the *only* 'complexity engine'.

Also a final key point is, that because *Homo sapiens* converts free energy to Shannon information, the latter can become a true thermodynamic variable. This logic points to some resolution of the arguments about 'Shannon entropy'. In principle, it would be possible to make a conversion quantification between net entropy gain and increase in complexity.

8.4 Concluding comments

Several, more points need to be made. Firstly, the timescale of our biological engines is an engineering rather than evolutionary one, with a dominant cycle period of 24 h. A part of the free energy input must go into Kauffman's 'expanding phase space' on an evolutionary/natural selection timescale.

Secondly, because the biological engine open system is still a black box, it can also accommodate a cumulative hierarchy of communities, not just individual organisms. As the hierarchy ascends, such communities develop into the biosphere itself. In this way, Kauffman's 'fourth law' focus on the biosphere could be reconciled with that of Wicken, either by Kauffman re-expressing it for a closed system, or by Wicken extending his thermodynamic treatment to open system organisms. The former is probably thermodynamically preferable because simpler.

Finally, this gives an intriguing parallel path situation for evolution. Genetic development may occur in two ways – either via natural selection, or through *Homo sapiens* 'tinkering' with the biology (after [27]). Hawking puts it more dramatically. 'Now we are at the beginning of a new era, in which we will be able to increase the complexity of our internal record, the DNA, without having to wait for the slow process of biological evolution' ([26], p. 165). *Homo sapiens* has become a complexity engine indeed!

9 How mathematical is biology? How chaotic is evolution?

9.1 Introduction

> I wrote to Professor Miller, of Cambridge, and this geometer … tells me that it is strictly correct. If a number of equal spheres be described … there will result a double layer of hexagonal prisms … the same with the best measurements which have been made of the cells of the hive – bee.
>
> Charles Darwin [9], pp. 173, 174

> Physics seems to be mostly sums, biology mostly essays. I like biology better than physics, but I'm better at sums than I am at essays. Anyway, my best friend is going to do biology, so I can keep asking him … . That does it – I'll do physics.
>
> Eric Laithwaite [37], pp. xi, xii

> I will not need to use in this book any more mathematics than can be expressed in words than in abstruse equations.
>
> Philip Ball [38], p. 14

'Mathematical Biology'

James Murray [39], Title of book

The rationale for the three chapters we have written is that of a biological/thermodynamic synthesis which has been largely accepted into mainstream thought. Here we wish to extend our consideration to the idea that the synthesis should also include *mathematics*. Of necessity our discussion must be brief and we feel that the material, although convincing to us, is best expressed as questions and not postulations.

The general relevance of mathematics to nature was discussed in the main Series Introduction in Volume 1. While biology's public face is that of an experimental science (Darwin consults with Miller rather than jointly researching with him) there are now clear indications of a possible holistic mathematical basis. This is especially represented by the massively comprehensive studies of Murray [39] stemming from reaction–diffusion systems. This really harks back to the celebrated analysis of pattern formation by Turing dating from 1952 [40]. This was recently nicely assessed by Maini [41]. Maini, in fact, closes our thermodynamics circle, pointing out the connection between reaction–diffusion (or activator–inhibitor) systems and dissipative structures of Prigogine.

Mathematically (see Maini for a concise explanation) such equations exhibit diffusion-driven instabilities which result in stable pattern formations – a kind of 'order out of chaos' or emergent property, where complexity results from an 'integration of fundamental units'.

Overall, the above points to the possible synthesis of thermodynamics, mathematics and biology into a kind of extended Prigoginian model. We could describe this by:

> Living systems maintain their far-from-equilibrium state by extracting free energy from, and rejecting heat to, the environment. The energy so gained dissipates itself in a chaotic manner (sometimes termed deterministic chaos) and results in an eventual increase in complexity.

This much can, we believe, be described as a mainstream thermodynamics/biology interpretation which points towards the involvement of chaos-oriented mathematics. For example, there is increasing evidence that the internal system behaviour of *Homo sapiens* is of a chaotic (non-linear dynamic) manner. In this context see [42, 43] relating to the cardiovascular system, in a journal special issue on Chaos published in March 2006. Also, Alzheimer's and Parkinson's diseases are being investigated along similar lines [44]. Such behaviour does not so much prove the above model, but rather that the model is consistent with current medical evidence. In the same way it is consistent with Ji's previously-discussed implications of a chaotic descriptor of overall human behaviour. While the core message of this chapter is that living organisms may be fairly regarded as *operating thermodynamic systems* there is increasing evidence that this has a mathematical character.

9.2 Self-organization: a new keyword

We now wish to extend the applicability of chaos theory to general biology. We remind ourselves that the same power source (thermodynamics free energy) that permits living systems to operate also 'drives' morphogenesis and evolution itself. This fact represents a prima facie case that the same principles applying to our day-to-day existence as living organisms, could also apply to the biological processes (whether developmental or evolutionary) that brought such existence into being.

This brief review is introduced with the well-accepted work of Turing [40]. This concerned pattern formation and morphogenesis. Combined with Murray's subsequent studies, Turing's

approach firstly provides a mathematical basis for *surface pattern formation* for organisms. Certain shell patterning, as pointed out by Ball ([38], p. 89), has no survival implications, and so is free of natural selection as a driving mechanism. Most patterning, however, whether of cats and zebras (Futuyama ([45], p. 43); Ball ([38], pp. 84–88)) of angel fish (Wolpert *et al.* ([46], p. 317 or Ball (pp. 93, 94)) or of butterflies (Ball (pp. 94–99)) does have survival effects, and this means that *both* natural selection *and* 'mathematics' simultaneously contribute to the end result. In the case of the angel fish, development is involved in the patterning too. This is obviously potentially a most important conclusion, which Futuyama describes ([45], p. 431) as 'an example of possible constraints on the phenotype arising from physical chemistry'.

Now the expression 'mathematics' has been used above as a cause complementary to natural selection. Turing postulated that in an initially uniform chemical mixture, the processes of diffusion and auto-catalytic chemical reaction could result under certain conditions, in forms of patterning. This reaction diffusion system could be solved via a partial differential equation. Ball, who gives on excellent non-mathematical explanation in [38] (p. 79, 80) describes [40] as 'undoubtedly one of the most influential in the whole of theoretical biology'. A rather fine description with attractive diagrams is also given by Wolpert *et al.* ([46] p. 317). Wolpert *et al.* give the important caveat of 'as yet no direct evidence'. Harold, on the other hand, in his extensive review of morphogenesis in micro-organisms [47] points out that some of the resultant patterns 'bear an uncanny resemblance to biological ones' (p. 414).

At the start of his treatise, Murray ([39], p. 1) gives the generic PDE as:

$$\frac{\partial u}{\partial t} = \mathrm{f}(u) + \mathrm{D}\nabla^2 u,$$

where u is the reactants vector, $f(u)$ is the non-linear reaction kinetics, and D is the diffusivity matrix.

It is important to realize that the above equation is only 'obeyed' in the sense that it aims to be a high fidelity conceptual representation in quantitative terms of what is actually a fundamental mode of physical behaviour (Futuyama's 'physical chemistry'). In fact, in biological terms 'activator–inhibitor system' is used (see, e.g. [38, 46]), and the process is termed 'self-organization'. As is apparent from Table 1 the expression has become a new keyword.

To conclude this section, we need to make the statement '*both* natural selection *and* self-organization simultaneously contribute to the end result'.

9.3 Self-organization (mathematics, chaos theory): how powerful an effect?

> … I don't believe I am saying anything that will disturb molecular biologists …
>
> Philip Ball [38], p. 9

> This undoubted instance of the inheritance of acquired characteristics leaves both the theory of evolution by natural selection and the central dogma quite unscathed, but has important implications for the inheritance of form.
>
> Franklin Harold [47], p. 387

> If Kauffman is correct, the dynamic states of interacting genes also generate domains of attraction, or cell types.
>
> Rudolf Raff [49], p. 330

Table 1: Self-organization – a new keyword.

Author	Title	Detail
Harold, F.M., 1990 [47]	To shape a cell: an inquiry into the causes of morphogenesis of microorganisms	[Self-Assembly] 'a single principle of morphogenesis universally acknowledged by biochemists and cell biologists' (p. 393) [Self-Organization] Section Heading (p. 413)
Kauffman, S.A., 1995 [48]	*At Home in the Universe*	The Search for the Laws of [Self-Organization] and Complexity (sub-title)
Webster, G. & Goodwin, B. 1996 [29]	*Form and Transformation*	*Phyllotaxis* as a [Self-Organising] Growth Process. Section Heading (p. 215) 'an … unresolved tension between principle of [self-organization] and natural selection' (p. 23)
Raff, R.A., 1996 [49]	*The Shape of Life*	'results of developmental genetics show that [self-assembly] in itself is an incomplete model for ontogeny'
Wolpert, L. *et al.*, 1998 [46]	Principles of development	'[Self-organization] may be involved in pattern formation in the limb bud', Section 10-9 (p. 315)
Ball, P. 1999 [38]	*The [self-made] tapestry – Pattern Formation in Nature* (Book Title)	

If, then, there appears to be general acceptance that self-organization à la Turing (and, in its wake, mathematics and chaos theory) has a real part to play in morphogenesis, and if in the reverse direction, enthusiastic supporters like Ball and Harold maintain a neo-Darwinian orthodoxy, why is there a strong element of controversy?

Table 2 attempts to summarise the debate by way of cross-referencing research workers with a number of recent and definitive publications. The workers comprise D'Arcy Thompson (the originator of the idea of form being defined not just by evolution but also by forces [52]) Turing himself, and the proponents of self-organization/mathematics/chaos theory – Murray, Kauffman and Goodwin. Prigogine is also included, in consistency with our overall thermodynamic context. Finally, plant phyllotaxis appears, as that has the same rationale of self-organization (Webster and Goodwin [29], figure 1).

The controversy exists for the following reasons. On the one side, the general acceptance implied by Table 1 is too strong an expression in the light of Table 2. So Futuyama writes 'possible constraints' ([45], p. 431). So Wolpert writes 'may be involved', 'there may therefore be', 'one possibility is' ([46], pp. 315–317). On the other side, Goodwin in particular is convinced, that 'there are some fundamental aspects of an organism's form that persist *in spite of* natural selection', and of 'features that evolution is powerless to erode away' ([38], p. 9). This is referred to by Raff as a 'strong version' of the 'structuralist hypothesis' which he finds 'somewhat startling' ([49], pp. 312, 313). Further there is Goodwin's (to Raff) 'most extraordinary statement of this view' to do with the common supposed generation of 'the mouth of ciliate protozoa such as

Table 2: How mathematical is biology? References to some significant research workers.

Publication	Research workers referenced						Plant phyllotaxis
	D'Arcy T.	Turing	Prigogine	Murray	Kauffman	Goodwin	
Futuyama, 1986 [45]	p. 422	p. 429	–	p. 430/431	–	–	–
Harold, 1990 [47]	p. 393	p. 414	p. 414	–	–	p. 423	na
Goodwin, 1994 [50]	p. xiii	p. 97	–	p. 135, 150, 226	p. 171 f.	na	p. 105f.
Webster & Goodwin, 1996 [29]	p. 116	p. 199	–	p. 199	p. xiv	na	p. 215
Raff, 1996 [49]	–	–	–	(p. 471)	p. 179, 298, 329	p. 298, 312	na
Wolpert *et al.*, 1998 [46]	–	p. 317	–	–	–	–	–
Ball, 1999 [38]	p. 6f	p. 78 etc	p. 81 f., 255	p. 96, 102	–	p. 9	p. 104
Purves, *et al.*, 2001 [22]	–	–	–	–	–	–	–

Notes:

1. Discussions (verbal/written) with Wolpert referenced by: Murray ([39], p. 75), Raff ([49], pp. 312, 313), Goodwin ([50], p. vii), Webster and Goodwin ([29], p. 136f).
2. Goodwin, Kauffman and Murray have collaborated – see Goodwin, Kauffman and Murray [51], referenced by Webster and Goodwin [29] (p. 233).
3. Goodwin, in particular, has encountered the sharp disagreement of Raff (see above), and others (Ball [38], p. 9).

Tetrahymena, and the dorsal lip of the blastopore in amphibia'. Raff finds 'no support in the data for a strong version of structuralism' (p. 313). This sounds damning, but it is telling to compare Raff's reaction to that of Harold's. In the same subject area of ciliate protozoa, Harold registers 'Goodwin's remarkable proposition that the diversity of ciliate shapes can be encompassed by a simple mathematical formula, Laplace's general field equation' [47], p. 423).

Harold's consistent rationale is that generic specifications and morphogenesis are only 'linked quite indirectly' and that 'there appear to be few true morphogenes (genes dedicated to specifying shape) at least in microorganisms' (p. 386). Possibly consequently, he finds Goodwin's ideas stimulating rather than upsetting. Ball, too, is cautiously supportive ([38], p. 9) – 'it seems to me that (Goodwin's) arguments become weakened only when extended from specific instances to the status of a new developmental principle in biological growth'.

At this stage, we stress our diffidence in monitoring this biological debate. However, we do have an interest. One of us (M.W.C) is co-supervising a joint biological/engineering research project with Dr Dave Roberts of Natural History Museum, London in just this subject area. Specifically, it is the investigation of a possible generic body plan for especially, the ciliate protozoon *Tetrahymena*.

9.4 Self-organization and thermodynamics

Table 2 shows that the only texts referring to Prigogine are those of Harold and of Ball. In fact, the underlying intellectual material of the current chapter can also be gained in a biological context by studying the 50 pages of Harold's review, combined with the 16 pages (pp. 252–267) of Ball's last chapter. Ball includes much stimulating comment on the role of attractors, and of noise and robustness, topics which we have omitted for brevity's sake. The stance of Ball and Harold reflect, we feel, the way in which we have raised the issue of the significance of mathematics in biology. Moreover, it is telling that once the thermodynamic contributions are accommodated the possible scientific consequences appear to be the same. Overall it tends to justify our suggestion of a thermodynamic/mathematical/biological synthesis.

9.5 Self-organization, chaos and evolution

In a way, this is an easy subject to review very briefly, as Kauffman's entire research enterprise focuses on this. We refer, for example, to his recent *At Home in the Universe. The Search for the Laws of Self-Organization and Complexity* [48]. His book 'describes my own search for laws of complexity that govern how life arose naturally from a soup of molecules, evolving into the biosphere we see today' (p. viii). To quote the blurb, '… the forces for order that lie at the edge of chaos … ', 'Kauffman extends this new paradigm to economic and cultural systems, showing that all may evolve according to similar general laws'. In the book we find consistent reference to the components of chaos theory – attractors, in particular.

It is certainly not just Kauffman. The very recent *Deep Simplicity. Chaos, Complexity and the Emergence of Life* by Gribbin [53] covers much the same field with a virtually identical approach. Also, the above mention of economic and cultural systems is expanded by, again, an entire book, that of Goonatilake *The Evolution of Information* [25]. In his case the chaos theory is highlighted by the presentation of biological and economic history as a series of bifurcations (p. 162). One or two brief quotes demonstrate the commonwealth of understanding Goonatilake shares with the other authors: '… dissipative structures operate, giving rise to evolutionary processes' (p. 167), and 'the deeper – level dynamics of the thermodynamic engine of open systems' (p. 168).

All in all, the proposed triple synthesis of Section 9.4 is extended by Kauffman and Goonatilake to include evolutionary processes wider than in biology.

10 Conclusion

This chapter completes our survey of the relevance of thermodynamics to biology, in the form of a trilogy of chapters. The first was in Volume 1 of the Design and Nature Series, and the others are in this volume. We have endeavoured to make them understandable for those 'without previous knowledge'. In this chapter we have studied the key concept of the heat engine, which stems from the roots of thermodynamics in the Industrial Revolution. Moreover, by widening the definition to focus on *output* and on *open systems* it has proved possible to accommodate pre-Industrial Revolution machines and living organisms. The 'black box' character of the definition of the thermodynamic system makes this a relatively straightforward procedure.

By considering locomotion, it is possible to view draught animals (and *Homo sapiens*) as work engines and all species as survival engines. Also on this basis, *Homo sapiens* is the only 'eucultural' complexity engine, able to convert free energy input to exosomatic information output to a very high degree. The significance of this interpretation is investigated and related to the possibility of further laws of thermodynamics for living systems.

The dissipative structures of Prigogine's thermodynamics necessitate the involvement of non-linear dynamics/chaos theory. An initial consideration is given to the question of how mathematical biology actually is. The assessment includes Turing's and subsequent studies on pattern formation and morphogenesis.

We conclude that there is a convincing synthesis of thermodynamics and biology. This is the main message of our entire study. Also, we believe that the future may well see that the synthesis includes mathematics as well.

Acknowledgements

This chapter includes material adapted from a previous presentation by one of the authors, M.W.C.: Man the engineer – an interpretation of *Homo sapiens* based on engineering thermodynamics. Introductory address, 2nd International Conference on Design & Nature, Rhodes, Greece, June 2004. *Design and Nature II*, eds. M.W. Collins & C.A. Brebbia pp. xv–xx WIT 2004.

We are most grateful to Miss Monika Turska (assistant to Prof. Jan Antoni Stasiek) who prepared this chapter for publication.

References

[1] Kondepudi, D. & Prigogine, I., *Modern Thermodynamics*, Wiley: Chichester, 1998.
[2] Fraser Stewart Book Wholesale Ltd, *The Complete Family Encyclopaedia*, Helicon Publishing: London, 1992.
[3] Prigogine, I., Schrödinger and the riddle of life (Chapter 2) *Molecular Theories of Cell Life and Death*, ed. S. Ji, Rutgers University Press: New Brunswick, NJ, 1991.
[4] Schrödinger, E., *What is life?,* Canto edn, Cambridge University Press: Cambridge, 1992.
[5] Ji, S., 'Biocybernetics': a machine theory of biology (Chapter 1). *Molecular Theories of Cell Life and Death,* ed. S. Ji, Rutgers University Press: New Brunswick, NJ, 1991.
[6] Desmond, A., *Huxley: Evolution's High Priest*, Michael Joseph: London, 1997.

[7] Weaire, D., William Thomson (Lord Kelvin) 1824–1907 (Chapter 8). Creators of mathematics: The Irish Connection, ed. K. Houston, University College Dublin Press: Dublin, 2000.
[8] Riley, R.C., *The West Country,* Railway History in Pictures Series, David & Charles: Newton Abbot, UK, 1972.
[9] Darwin, C., *On the Origin of Species by Means of Natural Selection* or *the Preservation of Favoured Races in the Struggle for Life*, Murray: London, 1859.
[10] Information from Library, IMechE, and Archives, ICE, to M.W.C., 24 November 2004.
[11] Smith, C., *The Science of Energy*, Athlone Press: London, 1998.
[12] Rogers, G.F.C. & Mayhew, Y.R., *Engineering Thermodynamics Work and Heat Transfer*, 4th edn, Prentice Hall: Engelwood Cliffs, NJ, 1992.
[13] Bryson, B., *A Short History of Nearly Everything*, Doubleday: Canada, 2003.
[14] Smith, C. & Wise, M.N., *Energy and Empire: A Biographic Study of Lord Kelvin*, Cambridge University Press: Cambridge, UK, 1989.
[15] Desmond, A., *Huxley: The Devil's Disciple*, Michael Joseph: London, 1994.
[16] Crowther, J.G., *British Scientists of the Nineteenth Century,* Vol. II, Pelican Books A, Penguin: Harmondsworth, Middlesex, 1941.
[17] McCartney, M., William Thomson: king of Victorian physics, *Physics World*, December 2002.
[18] Ruddock, I., Lord Kelvin's science and religion, *Physics World*, February 2004.
[19] King, R., *Brunelleschi's Dome*, Pimlico: London, 2001.
[20] Galluzzi, P., *Mechanical Marvels*, Giunti: Florence, 1996.
[21] Clutton-Brock, J., *Horsepower*, Natural History Museum Publications: London, 1992.
[22] Purves, W.K., Sadawa, D., Orians, G.H. & Heller, H.C., *Life*, 6th edn, Sinauer Associates/W.H. Freeman and Co. Ltd: New York, NY, 2001.
[23] Landels, J.G., *Engineering in the Ancient World* Constable: London, 1997.
[24] Lumsden, C.J. & Wilson, E.O., *Genes, Mind and Culture: The Co-evolutionary Process*, Harvard University Press: Cambridge, MA, 1981.
[25] Goonatilake, S., *The Evolution of Information*, Pinter: London/New York, 1991.
[26] Hawking, S., *The Universe in a Nutshell*, Bantam Press: London, 2001.
[27] Chaisson, E.J., The cosmic environment for the growth of complexity. *Biosystems*, **46**, pp. 13–19, 1998.
[28] Mikielewicz, J., Stasiek, J.A. & Collins, M.W., The laws of thermodynamics: cell energy transfer, *Nature and Design*, eds. M.W. Collins, M.A. Atherton & J.A. Bryant, Vol. 1, International Series on Design and Nature, WIT Press: Southampton, UK, 2005.
[29] Webster, D. & Goodwin, B., *Form and Transformation*, Cambridge University Press: Cambridge, UK, 1996.
[30] Wilson, E.O., *Sociobiology: The New Synthesis*, Harvard University Press: Cambridge, MA, 1982.
[31] Wilson, E.O., *On Human Nature*, Harvard University Press: Cambridge, MA, 1978.
[32] Wicken, J.S., *Evolution, Thermodynamics and Information*, Oxford University Press: Oxford, UK, 1987.
[33] Ridley, M., *Nature via Nurture*, Fourth Estate: London, UK, 2003.
[34] Tiezzi, E., *The End of Time*, WIT Press: Southampton, UK, 2003.
[35] Kauffman, S., *Investigations*, Oxford University Press: Oxford, UK, 2000.
[36] Brooks, D.R., & Wiley, E.O., *Evolution as Entropy,* 2nd edn, University of Chicago Press: Chicago, IL, 1988.

[37] Latithwaite, E., *An Inventor in the Garden of Eden*, Cambridge University Press: Cambridge, UK, 1994.
[38] Ball, P., *The Self-Made Tapestry*, Oxford University Press: Oxford, UK, 1999.
[39] Murray, J.D., *Mathematical biology II: spatial models and biomedical applications*, 3rd edn, Springer: Germany 2000.
[40] Turing, A.M., 'The chemical basis of morphogenesis', *Phil. Trans. Roy. Soc., London*, **B 237**, pp. 37–73, 1952.
[41] Maini, P.K., The impact of Turing's work on pattern formation in biology, *Mathematics Today*, pp. 140–141, August 2004.
[42] Griffith, T.M., Parthinos, D. & Edwards, D.H., Nonlinear analysis and modeling of the cellular mechanisms that regulate arterial vasomotion. Special issue on Chaos in Science and Engineering. *Proc. of the Institution of Mechanical Engineers* Part C Journal of Mechanical Engineering Science, 220(C3), pp. 367–383, 2006.
[43] Xu, S.-X., Wang, Q.-W., Chen, D.-D. & Collins, M.W., Nonlinear dynamic simulation of mechanical periodicity of end diastolic volume of left ventricle under the influence of Baroreflex. *Special issue on Chaos in Science and Engineering*. *Proc. of the Institution of Mechanical Engineers* Part C Journal of Mechanical Engineering Science, 220(C3), pp. 357–366, 2006.
[44] Elnashaie, S. & Mahecha-Botero, A., Towards understanding Alzheimer's and Parkinsons'. *The Chemical Engineer*, pp. 29–31, August 2005.
[45] Futuyama, D.J., *Evolutionary Biology*, 2nd edn, Sinauer Associates, MA, 1986.
[46] Wolpert, L., Beddington, R., Brockes, J., Jessel, T., Lawrence, P. & Mayerowitz, E., Principles of development. *Current Biology*, Oxford University Press: London, UK, 1997.
[47] Harold, F.M., To shape a cell: an inquiry into the causes of morphogenesis of micro-organisms. *Microbiological Reviews*, pp. 381–431, December 1990.
[48] Kauffman, S.A., *At Home in the Universe*, Oxford University Press: New York, 1995.
[49] Raff, R.A., *The Shape of Life*, The University of Chicago Press: Chicago, II,1996.
[50] Goodwin, B., *How the Leopard Changed its Spots*, Phoenix: London, UK, 1994.
[51] Goodwin, B.C., Kauffman, S.A. & Murray, J.D., Is morphogenesis an intrinsically robust process? *J. Theoret. Biol*., **163**, pp. 35–144, 1993.
[52] Thompson, D'Arcy W., *On Growth and Form*, Canto, Cambridge University Press: Cambridge, UK (reprint from 1917), 1992.
[53] Gribbin, J., *Deep Simplicity*, Allen Lane', Penguin: London, 2004.

Chapter 7

Information theory and sensory perception

M.D. Plumbley & S.A. Abdallah
Department of Electronic Engineering, Queen Mary University of London, London, UK.

Abstract

In this chapter, we explore Shannon's information theory and what it may tell us about the biological processes of sensory perception. We begin by discussing some historical theories of sensory perception, including concepts of objects and invariances as well as principles from Gestalt psychology. We consider the ecological principle that perception should be useful for an organism, and introduce information theory as a possible way to measure this usefulness. After introducing some of the key concepts from information theory, such as entropy, information, redundancy and factorial coding, we draw parallels between communication in engineering systems and sensory perception. We discuss Attneave's early proposal that perception should create an economical description of sensory inputs, as measured by information theory, and Barlow's suggestion of lateral inhibition as a possible mechanism to achieve this. We discuss the important role played by noise in an information processing system, and its importance when determining how to optimise the communication of information. This leads to the concept of information-optimal filters which change their characteristics at different signal-to-noise levels, a feature exhibited by fly and human visual systems. For information processed across topographic maps, optimal information processing leads to a principle of uniform information density, whereby a larger area in the cortex allows higher information density, and hence higher accuracy or sensitivity, at the corresponding sensors. We also discuss some recent investigations of information theory applied to spiking neural systems, and the concept of economy of impulses. Finally, we discuss some related concepts such as structure in a perceptual stimulus and implications for Gibson's ideas about perceptual systems, and we speculate about the possible importance of attention and active perception for efficient information processing in a sensory system.

1 Introduction

The problem of sensory perception has been of interest to philosophers and psychologists for hundreds of years. During the last 50 years or so, since the introduction of Shannon's information theory [1], the parallel between perception and communication has been explored by many

researchers, such as Attneave [2] and Barlow [3]. More recently, this approach has led to the development of adaptive neural network models that optimise information [4–6], and to considerations of energy-efficient sensory coding [7].

In the first part of this chapter, we will give a brief overview of some historical theories of sensory perception. This will lead us to the consideration of perception as something which is *useful* to an organism, and therefore to the idea that perception should be performed *efficiently*. From a design viewpoint, we are therefore looking to assess how we measure the usefulness of a particular sensory system, and its efficiency in performing the task of perception. In the second part of the chapter, we will consider how concepts from *information theory*, such as entropy, mutual information, and redundancy, can be used to measure this usefulness. We will see how these help to explain certain features of sensory systems, and what features are desirable in such a system.

2 Theories of perception

2.1 What is perception?

According to Fodor ([8], p. 40]), 'what perception must do is to so represent the world as to make it accessible to thought'. Thus the central problem of perception is to construct such a representation from the collection of signals emanating from sensory transducers. Fodor goes on to say, in more precise terms, that although these signals are best thought of as representing conditions at the *surface* of an organism, the representations produced by perceptual systems are 'most naturally interpreted as characterising the arrangements of *things in the world*'. The process of going from one to the other is essentially one of *inference*, where 'proximal stimulus' provides the evidence or 'premises', and the conclusions are 'representations of the character and distribution of distal objects'. Similarly, Barlow [9] describes perception as 'the computation of a representation that enables us to make reliable and versatile inferences about associations occurring in the world around us'.

As a conceptual framework for thinking about perception, this is by no means a recent development. For example, in the 18th century, Berkeley wrote ([10], § 18)

> It remains therefore that if we have any knowledge at all of external things, it must be by reason, inferring their existence from what is perceiv'd by sense.

Notwithstanding his use of the word 'perceived', the gist is the same. Helmholtz's theories about vision and audition, from early in the 20th century, also hold up remarkably well today. With regard to vision, he suggested that ([11], § 26)

> such objects are always imagined as being present in the field of vision as would have to be there in order to produce the same impression on the nervous mechanism, the eyes being used under normal ordinary conditions.

This is an early expression of the idea that perception may be concerned with inferring the worldly *causes* of sensations, the sensations themselves being incidental. He added that 'we are wont to disregard all those parts of the sensations that are of no importance so far as external objects are concerned'.

Addressing the commonplace observation that perception seems to be a transparent, effortless activity, Helmholtz [11] argued that this occurs through a process of 'unconscious induction'. However, Gibson [12] claimed that 'the senses can obtain information about objects in the world

without the intervention of an intellectual process'. By *intellectual process*, he almost certainly did not mean what we might now call a computational process. What he probably objected to is the need for perceptual *inference*, because, he maintained, under normal conditions there is enough information available to leave no room for uncertainty.

Gibson also rejected the idea that passive analysis of transduced sensory signals forms the sole basis for perception ([13], Ch. 4): 'The inputs of the nerves are supposed to be the data on which perceptual processes in the brain operate. But I make a quite different assumption'. This leads to the concept of *active perception*, where a perceptual system is engaged not just in the passive analysis of whatever stimulation happens to be playing over an organism's receptors, but in the active exploration of an 'ambient stimulus field'. Thus the visual system is responsible not just for the analysis of retinal data, but also for the control of systematic eye and head movements designed to extract more information from the available stimulus. It is this information which removes the ambiguity that would otherwise necessitate an inferential system. For example, even slight movements of the head provide strong, unambiguous cues for localisation in 3D for both vision and audition. Biosonar in bats and dolphins is another good example of an intrinsically active perceptual system.

Incidentally, Helmholtz [11] touched upon similar ideas while discussing his thesis that we perceive *causes*, arguing that 'it is only by voluntarily bringing our organs of sense in various relations to the objects that we learn to be sure as to our judgments of the causes of our sensations'. This is a view of active perception as a kind of *experimentation*, by which we distinguish causal relationships from mere coincidence.

2.2 The objects of perception

Just what is it that we actually perceive? In the context of auditory perception, Bregman ([14], p. 9) put it like this: 'The goal of scene analysis is the recovery of separate descriptions of each separate thing in the environment. What are these things?' A fairly uncontroversial answer would be that they are *objects*. Bregman goes on to discuss how objects serve as foci for all the perceptual qualities one experiences in response to a scene. This is implicit in the idea of a *property*: it must be a property of *something*. The object (as a mental construct) plays a syntactic role, binding together properties pertaining to the same physical object.

Syntactic structure aside, what is the phenomenal manifestation of *object-ness*? What aspect of the stimulus triggers the mental organisation? A common observation is that a reliable *association* or *correlation* between a number of features signals the presence of an object, as noted by Berkeley ([10], Part 1, § 1):

> As several of these [sensations] are observ'd to accompany each other, they come to be marked by one name, and so to be reputed as one thing. Thus, for example, a certain colour, taste, smell, figure and consistence having been observ'd to go together, are accounted one distinct thing, signified by the name *apple*.

Mach [15], in a similar vein, thought that bodies, as we perceive them, are made up of 'complexes of sensations' with some relative permanence, and, according to Helmholtz [11], 'experience shows us how to recognise a compound aggregate of sensations as being the sign of a simple object'.

The prevailing view is that we perceive the world in terms of objects, their qualities, and their adventures. An object has coherence and continuity, but also variability. The object, as a mental construct, also serves as a scaffold on which to affix properties that seem to be 'hanging around

together' in a suspicious way. These mental objects tend to agree with external objects because the latter do indeed tend to leave a trail of physical evidence, like an animal in the woods, that sometimes leaves tracks, spoor, broken twigs, remnants of its last meal and so on. Someone who knows what to look for can infer the existence of the animal, and learn a lot about it, without directly seeing it.

Gibson [12, 13] took a different position, suggesting *invariance* as a more fundamental construct than objects. Invariants can only be discovered in the context of change, whether in a single thing evolving over time, or conceptually across a set of distinct things ([12], p. 275): 'In the case of the persisting thing, I suggest, the perceptual system simply extracts the invariant from the flowing [stimulus] array; it *resonates* to the invariant structure or is *attuned* to it. In the case of the substantially distinct things, I venture, the perceptual system must *abstract* the invariants.' In short, invariants represent things that remain constant despite change, like the shape, size, and indeed the identity of an object as it moves about.

The Gestalt psychologists were concerned with organising parts into wholes. These *gestalten*, or emergent shapes, can have properties, or *gestaltqualitäten*, that have meaning only for the whole and not its parts. Köhler [16] gives the example of a musical interval: *fifthness* is not a quality possessed by notes, but by the gestalt, the interval. *Major-ness* or *minor-ness* of chords is another good example. Gestalt theory sees perceptual organisation as a grouping process, in which parts of the visual image (and they were defined primarily in the context of vision) are allocated to one or other object. These have subsequently been adapted for use in audition [14, 17], largely via the adoption of auditory 'images' which take the form of time–frequency distributions such as spectrograms. Energy in different regions of the time–frequency plane is grouped according to such criteria as onset synchrony, proximity in time or frequency, harmonicity, and the principle of *common fate*, which recognises that sounds that undergo, for example, common frequency or amplitude modulations, are likely to have come from the same source.

2.3 Dealing with uncertainty

Sensory evidence is sometimes uncertain, inconclusive or ambiguous. Usually, we do not notice this uncertainty, filling in the gaps or making assumptions based on prior experience of what is likely, with all of this happening below the level of conscious awareness. In speech, for example, phonemes in the middle of words or phrases which have been completely masked by noise are often subjectively heard without the listener even noticing that they were missing; this is known as *phonemic restoration* [18], a specific type of *perceptual restoration*. A similar effect has been reported for music [17].

Gestalt theory deals with this type of uncertainty using the principle of good closure, or *prägnanz*. When there are competing alternative organisations suggested by the other rules, or when there is missing data, this principle says that *good forms* should be preferred. Here, the term 'good form' refers to qualities such as simplicity, regularity, symmetry, or continuity. However, as Bregman ([14], p. 26) observes, we *do* sometimes see forms with gaps in them. The principle is really for completing *evidence* which has gaps in it, and its job is to fill in missing or doubtful data.

One may then ask, what does the brain choose to fill-in the gaps with? What is 'good' form? Hochberg [19] discusses how the Gestalt principle of good *prägnanz* can be interpreted as assuming either the *simplest* (in some sense), or the *most likely* resolution of an ambiguity. This suggests

that there is an interpretation of this principle in terms of likelihood and probability. Mach ([15], § 10.8, p. 213) observed:

> If the visual sense acts in conformity with the habits which it has acquired under the conditions of life of the species and the individual, we may, in the first place, assume that it proceeds according to the principle of probability; that is, those functions which have most frequently excited together before, will afterwards tend to make their appearance together when only one is excited.

Brunswick (according to Hochberg [19]) thought that the 'Gestalt laws were merely aspects of stimulus organisation, reflecting the probability that any parts of the visual field belonged to the same object' and Bregman ([14], p. 24) wrote 'It seems likely that the auditory system, evolving as it has in such a world, has developed principles for "betting" on which parts of a sequence of sensory inputs have arisen from the same source.' We will see later how these ideas can be interpreted in terms of *redundancy* of sensory information.

2.4 Representation and cognition

The acquisition of perceptual knowledge can be identified with the formation of *representations* that fulfil certain goals, or that make plain relevant or interesting aspects of that which is being represented. Fodor ([8], p. 29) states that 'contemporary cognitive theory takes it for granted that the paradigmatic psychological process is a sequence of transformations of mental representations and that the paradigmatic cognitive system is one which effects such transformations' and he goes on to give an informal definition of *representation* ([8], p. 38): '*Any* mechanism whose states covary with environmental ones can be thought of as registering information about the world; … the output of such systems can reasonably be thought of as *representations* of the environmental states with which they covary.'

Fodor's 'paradigmatic psychological process', a sequence of representations each of which is a transformation of the previous one, can be recast as a sequence of *black box* processing units connected by communications channels. Indeed, Marr ([20], p. 3) notes a duality between processing and representation. In doing this, our focus is shifted to the characteristics of the potentially imperfect or noisy channels. If our channels are imperfect, then the concepts of *distance* and *similarity* become important. If a representation is slightly corrupted, what will it become? If we cannot be sure that we recognise an object, what others will it 'look like?'

Richardson [21] proposed that subjective judgements of similarity between stimuli be represented as distances between points in the Euclidean plane. This has since developed into the notion that psychological dissimilarity can be equated with the idea of a distance in some metric space. As Davidson [22] points out, this is not automatically a valid identification, as geometric distances are required to satisfy a number of constraints, such as symmetry, whereas psychological dissimilarities need not. With this caveat in mind, the geometric visualisation of similarity is still a useful and intuitive aid to understanding.

Distances and representations tend to be tied together quite closely. Distance measures operate on representations of pairs of objects: if we change the representation of our objects, the distances will change too. So, if we wanted to build an artificial cognitive system that would report similarities between objects that correspond to human subjects, one way would be to use representations such that a simple distance measure would produce corresponding distances between objects.

Shepard's principle of *psychophysical complementarity* [23] addresses the issue of perceived distance. It says that mental structures should reflect certain transformational symmetries of the real world objects they represent. If an object can be subjected to a group of transformations

(say, rotations) without changing its essential identity, this should be mirrored by a corresponding group of transformations of the internal representation. These would not necessarily be implemented as physical rotations or in a physically circular structure, but as a set of states and operators that relate in the same way. For example, they might replicate the composition of large and small rotations, and the cyclic nature of rotation.

2.5 Mental structure vs stimulus structure

Hochberg [19] discusses the debate between those who believe that perception is driven by mental structure and those who believe that it is driven by stimulus structure. Hochberg defines stimulus structure as *intra-stimulus constraints*, and mental structure as *intra-response* constraints. Shepard [23] put it more directly: 'Does the world appear the way it does because the world is the way it is, or because we are the way we are?'

Gestalt theory holds that perceptual organisation is an achievement of the nervous system, and that *gestalten* do not exist outside the organism. However, Köhler ([16], Ch. 5) did concede that this interpretation may have an objective value: it may 'tell us more about the world around us', have 'biological value', and 'tends to have results which agree with the entities of the physical world'.

Mach ([15], § 1.13, p. 29) also had mentalist tendencies: 'Bodies do not produce sensations, but complexes of elements make up bodies.' However, he did not seem to imply that sensations are amalgamated into bodies in some ad-hoc manner, but rather that there are statistical regularities which are exploited. Gibson, with his theory of the *direct perception* of invariants, and the pick-up of information without an intervening 'intellectual process', believed that perception was driven by stimulus structure, not mental structure, even if this stimulus structure is only fully available to an active perceptual system.

Perception dominated by stimulus structure corresponds roughly with bottom–up or *data-driven* processing, which looks for structures in data without referring to any stored knowledge. Mental structure corresponds with top–down processes, variously known as *knowledge-driven* or, in psychology, *schema-driven* processes. The use of high-level knowledge in perception is not in doubt [24]. A more pertinent question is: How does high-level knowledge get to the top in the first place? In a *knowledge engineering* methodology, it is placed there by a human designer. In unsupervised learning, on the other hand, the acquisition of high-level knowledge is itself a data-driven process: it collects over time, percolating from the bottom up. For example, Leman and Carreras [25] contrast '*long-term data-driven* perceptual learning with *short-term schema-driven* recognition'. This is a sensible distinction to make when there is not enough structure in short spans of data to enable a fully data-driven approach, but there is enough structure in data accumulated over longer periods; the schema represents this accumulation of experience.

In conclusion, to quote Shepard's answer to his own question [23],

> (1) The world appears the way it does because we are the way we are; and (2) we are the way we are because we have evolved in a world that is the way it is.

Each of us has been immersed in this world since birth, and therefore we have had ample opportunity to build and adapt our mental structures to suit our environment.

2.6 An ecological perspective

Let us ask ourselves the question: 'What *use* is perception?' This immediately brings us to the viewpoint that perception must be *useful* for an organism. From a biological and evolutionary point

of view, the machinery of perception is an expensive burden–in humans, the brain is responsible for a large fraction of the body's energy consumption–and must therefore confer considerable advantages to the creature that owns it.

This view is more or less taken for granted. For example, Shepard [26] writes: 'The brain has been shaped by natural selection; only those organisms that were able to interpret correctly what goes on in the external world and to behave accordingly have survived to reproduce.' This amounts to the *ecological* approach to perception: one that recognises the mutual relationship between the organism and its environment.

The basic idea, like perceptual theorising in general, has a long history. Nearly 300 years ago Locke ([27], BII, Ch. IX, § 12) wrote:

> Perception, I believe, is, in some degree, *in all sorts of animals*; though in some possibly the avenues provided by nature for the reception of sensations are so few, and the perceptions they are received with so obscure and dull, that it comes extremely short of the quickness and variety of sensation which is in other animals; but yet it is sufficient for, and wisely adapted to, the state and conditions of that sort of animals who are thus made …

By the 1950s, evolutionary ideas were current. Mach ([15], § 13.3) speculated about the biological relevance of the 'sensations of tone' which 'constitute the means by which we distinguish large and small bodies when sounding, between the tread of large and small animals', and also that 'the highest tones presumably are of extreme importance for the determination of the direction from which a sound proceeds'. Of the perception of colour, he thought it ([15], § 6.2) 'essentially a sensation of favourable or unfavourable chemical conditions of life. In the process of adaptation to these conditions, colour-sensation has probably been developed and modified'.

Gibson [12, 13] mapped out a more complete ecological theory, stressing that an animal and its environment cannot be considered without one another. One of the elements of the theory is the perception of *affordances* ([13], p. 127): 'The affordances of the environment are what it *offers* to the animal, what it *provides* or *furnishes*, either for good or ill.' He argued that these affordances are apprehended in a quite direct fashion, that an intrinsic part of the perception of an object is an appreciation of what can be done with it. An apple says 'eat me', a predator says 'run away from me', a chair says 'sit on me'. As Gibson suggests in his definition, the distinction between *positive* and *negative* affordances ([13], p. 137) is an important part of the perceptual experience and of great relevance to the animal concerned. Shaw *et al.* [28] coined the term *attensity* to rate the relevance or usefulness of available information to a particular organism. Things that offer positive or negative affordances would have a high attensity and should be *attended to*, but things that offer neutral affordances have a low attensity, and presumably can be ignored.

Interestingly, this ecological approach has also been adopted in relation to music by Leman and Carreras [25]. While it may be objected that music is not a 'natural' phenomenon, this raises the interesting point that exposure to a musical tradition might constitute an *environment* of sorts, though a *cultural* rather than a natural one. To the extent that perceptual systems are shaped by experience rather than by evolution, there should be no difference between natural and cultural influences. The ontogenesis of the auditory system is likely to be as responsive to one as to the other [29].

2.7 Summary

To conclude the first part of this chapter, we have seen that there is a long and varied history of theories and viewpoints about perception. The ecological perspective proposes that perception performs a useful task for the organism, and therefore we are likely to find that organisms chosen

by natural selection are those with more *efficient* perceptual systems. Given that there is a limit to the amount of information that can be processed–or equivalently, that there are costs attached to information processing–it is reasonable to suppose that available resources are directed towards the detection and interpretation of events that have some *biological relevance*, such as the presence or absence of food, predators, potential mates and so on; in short, *useful* information.

In the next section we will see how Shannon's information theory [1] has been used to measure the *information* that might be useful to an organism, and how the question 'What is perception for?' leads to proposed answers to 'What should perception do?'

3 Information and redundancy

Information theory, first introduced in Shannon's 'Mathematical theory of communication' [1], has been used and developed extensively by communications engineers ever since its publication. Communications engineering is concerned with transmitting information from one place to another as efficiently as possible, given certain costs and constraints which are imposed on the communications system which we wish to use. For example, we may have a maximum number of *bits* (binary digits) per second that we can send down a certain binary transmission link, or we may have a radio transmitter with a limit on the maximum power level which we can use. Either of these define constraints, within which we must work.

Information theory gave communications engineering a precise meaning to the idea of *rate of information transmission*. This helped to explain how the properties of a communication channel can limit how fast information can be transmitted through it, and how to *code* signals to make most efficient use of such a channel. One of these results showed that a channel has an innate limit on its information rate, its *capacity*. It is impossible to send information through a channel faster than that channel's capacity, no matter how the information is represented or coded.

Not long after information theory first appeared, psychologists began to see potential applications in their own field [30, 31]. In particular, Attneave [2] and Barlow [32] drew at least a partial analogy between communication and sensory perception, the idea being that the initial stages of a sensory system serve to communicate information about the outside world to higher centres in the brain. This view was not universal, however. Green and Courtis [33], for example, objected to the use of information theory, on the grounds that sensory perception had no objective alphabet of symbols and no objective set of transition probabilities, as required by Shannon's formulation of *information*. They suggested that this would mean that different people might get different amounts of 'information' from the same picture or visual scene. Nevertheless, the last few years have seen an increasing interest in the use of information theory in the exploration of sensory systems, giving us some insight into why perceptual systems are organised as they are.

Before we consider these approaches in more detail, let us briefly introduce the key ideas from information theory itself.

3.1 Entropy and information

The two central concepts of information theory are those of *entropy* and *information* [1]. Generally speaking, the entropy of a set of outcomes is the uncertainty in our knowledge about which outcome will actually happen: the less sure we are about the outcome, the higher the entropy. If we know for sure what the outcome will be, the entropy will be zero. Information is gained by reducing entropy, for example, by making an observation of an outcome. Before the observation, our knowledge of the outcome is limited, so we have some uncertainty about it. However, after the

observation the entropy (uncertainty) is reduced to zero: the difference is the information gained by the observation.

Consider an experiment with N possible outcomes i, $1 \leq i \leq N$ with respective probabilities p_i. The entropy H of this system is defined by the formula

$$H = -\sum_{i=1}^{N} p_i \log p_i, \tag{1}$$

with $p \log p$ equal to zero in the limit $p = 0$.

For example, for a fair coin toss, with $N = 2$ and $p_1 = p_2 = 1/2$, we have

$$\begin{aligned} H &= -\left(\frac{1}{2}\log\frac{1}{2} + \frac{1}{2}\log\frac{1}{2}\right) \\ &= \log 2. \end{aligned}$$

If the logarithm is taken to base 2, this quantity is expressed in *bits*, so a fair coin toss has an entropy of 1 bit.

For any number of outcomes N, the entropy is maximised when all the probabilities are equal to $1/N$. In this case, the entropy is $H = \log N$. On the other hand, if one of the outcomes has probability 1 with all others having probability 0, then the entropy H in (1) is zero: otherwise, H is always non-zero and positive.

As we mentioned before, the information gained by an observation is the entropy before it, less the entropy after it. As an example, consider our coin toss again, and assume that we observe the outcome to be a 'head'. We denote the state of the coin by the random variable Ω, and write the entropy of the coin toss before any observation by $H(\Omega)$ (Fig. 1(a)). If we denote the observed face by X, we write the *conditional entropy* of the coin after the observation as $H(\Omega|X =$ 'head'), meaning 'the entropy in Ω given we know that $X =$ "head" ' (Fig. 1(b)).

The situation if the outcome is a 'tail' is exactly the same. The information in the observation X about the coin state Ω is then written

$$I(\Omega, X) = H(\Omega) - H(\Omega|X) = \log 2,$$

i.e. 1 bit of information was gained by the observation.

We can also measure the entropy and information content of several variables. If we have a vector $X = (X_1, \ldots, X_n)$ consisting of several variables X_i, then it is possible to show that

$$H(X) \leq H(X_1) + \cdots + H(X_n), \tag{2}$$

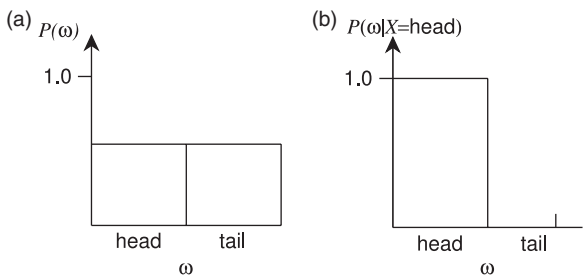


Figure 1: Probabilities of coin state $\Omega = \omega$, (a) before and (b) after observation of a 'head'.

or in other words, the entropy of the entire vector is no larger than the sum of the entropies of its elements. Equality holds when the set of elements are statistically independent. If the individual variables X_i have the same entropy $H(X_1) = H(X_2) = \cdots = H(X_n)$, then we have that $H(X) \leq nH(X_1)$.

In theory, any transformation, or *encoding*, of X, will require at least $H(X)$ bits on average to transmit to a remote observer. If we know the probability distribution of X accurately, and can use this information to construct an appropriate code, we can get as close to this theoretical limit as we like, but we can never compress X below this limit without some loss of information.

For an example, imagine a passage of English text. In a typical computer system this is encoded in ASCII symbols, taking 8 bits per letter. Should we be able to encode the information more compactly than this? What is the entropy of English text?

As a first step, consider that (ignoring spaces, punctuation and capitalisation) we have 26 letters. If each of these had equal probability, we would have $H = \log 26 = 4.7$ bits per letter. However, this is the maximum value that H can take: for any non-uniform distribution H will be less than this. For a closer estimate of the entropy H, we could measure the probabilities p_i, $i = 1, \ldots, 26$, for each different letter of the alphabet. This would take into account the statistics of individual letters, giving an estimate of about 3 bits per letter, but it would not take into account the dependencies *between* letters. For example, if we know a sentence begins with 'T', then there is quite a high probability that the next letter is 'H', and the following letter is 'E' (take a look at some of the sentences beginning with 'T' in this chapter). To get a better estimate of the entropy, Shannon [34] introduced a trick: a 'guessing game' that (in theory) performs a lossless encoding of a text passage. Measuring the code length (entropy) of this encoding gives an alternative estimate of the entropy of text, which turns out to be very low.

The idea of the guessing game is to take advantage of the language knowledge of a human subject to generate an efficient encoding for text. The guessing game works like this. Suppose that you have two identical twins, who always behave deterministically and identically. An experimenter has a passage of text, and asks the first twin to try to guess each letter, continuing to guess until the correct letter is guessed. This experimenter, who we call the *encoder*, records the number of guesses taken by the first twin to correctly guess each letter. The sequence of numbers, representing the number of guesses taken for each original letter, is passed to a second experimenter, called the *decoder*. The decoder asks the second twin to guess each letter of the passage of text, pretending to have the real text passage. However, the decoder simply asks the second twin to keep guessing until the required number of guesses has been made, writing down the appropriate letter reached each time. Since the twins behave identically, they will make the same sequence of guesses, and the decoder will then recover the original text. Therefore, the sequence of numbers is a *lossless encoding* of the original text passage, and the entropy of the entire sequence of numbers is equal to the entropy of the entire original passage [34].

Of course, identical twins such as these are hard to find in practice. Nevertheless, by playing the *encoding* part of the game only, it is possible to estimate the entropy of a text passage, since we know that we should, in theory, be able to decode the message. (If the reader is unhappy with the idea of identical twins, imagine instead a robot or computer system doing the guessing, which has linguistic knowledge similar to a human.)

The point of all this is that the sequence of numbers has a probability distribution with most of the probability mass concentrated around a small number of guesses. After the first few letters of a word or sentence, the guesser will mostly guess the correct letter first time, or at least within the first few letters, so the probability of '1 guess' is much higher than '10 guesses'. This probability distribution is much more concentrated than the probability distribution for individual letters, so the entropy per symbol is likely to be smaller. In fact, If the probabilities p_i are calculated for the

sequence of guess numbers, this gives an entropy of about 1 bit per letter position. Therefore the entropy of the original text cannot be more than about 1 bit per letter position [34].

This shows that there is a significant amount of *redundancy* in a typical passage of text. Redundancy is the difference between the information capacity used to represent a signal and the information actually present in the signal, and gives us an idea of how much we can *compress* a signal by transforming it into a smaller representation, while still retaining all of the original information. Sometimes redundancy is expressed as a ratio, rather than a difference, so that it varies between zero, when the information capacity is used to the full, and one, when the information capacity is completely wasted with no information actually communicated.

3.2 Redundancy reduction in perception

Attneave [2] suggests that perception may be the process that creates an economical description of sensory inputs, by reducing the redundancy in the stimulus. To explore this idea, he adapts Shannon's guessing game to a visual scene instead of a text passage, to see what redundancy might be present. As a concrete example, he considers a picture of an inkbottle on a desk, similar to Fig. 2. This picture, as for the original text-encoding problem, is composed of discrete symbols. A pixel can take any of three values: 'black', 'white' or 'brown' ('brown', which is the colour of a wooden desk, is shown as grey in this figure). The information content of the picture is equal to its entropy, since that is the amount of information we would gain if we reduced the entropy to zero.

Now, if a subject is asked to guess the colours in the picture, scanning along the rows from left to right, taking successive rows down the picture, we would find that most mistakes are made at the edges and corners in the picture. Applying this 'Shannon game' on the scene suggests that most information is concentrated around the edges and corners of a scene. This fitted well with psychological suggestions that these features are important, and with Hubel and Wiesel's

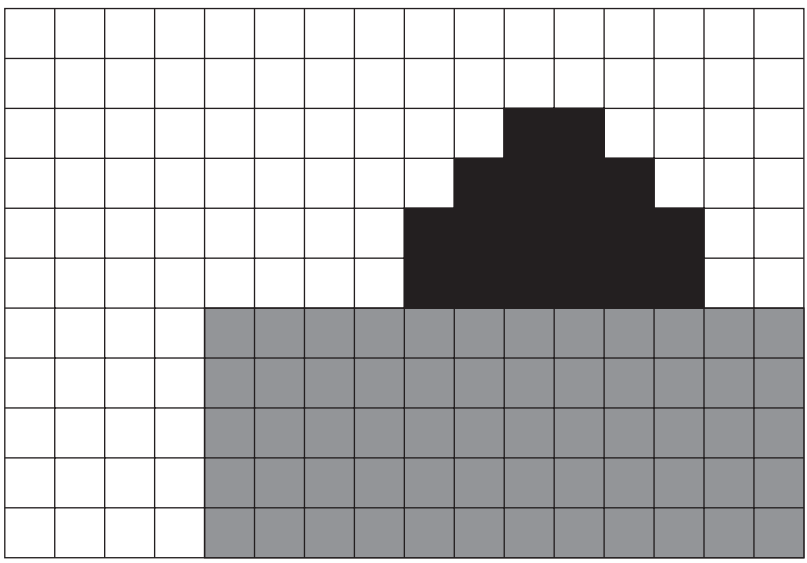


Figure 2: A redundant visual stimulus (after [2]).

discovery of simple cells and complex cells in the visual cortex, behaving as line detectors and edge detectors [35, 36].

3.3 Redundancy reduction and decorrelation

Soon after Attneave's [2] suggestion that the visual system should create an economical description of a scene, Barlow [3] suggested that *lateral inhibition* may be a mechanism by which this is achieved. Lateral inhibition involves inhibitory connections between cells in a parallel sensory pathway, and is a well-known feature of sensory systems [36]. Any signal which is common to many neurons in this pathway will be suppressed by the action of the lateral inhibition, while differences between inputs to neighbouring neurons, such as edges in a visual scene, would produce a significant output. In this way, lateral inhibition would produce a more economical representation of the scene, and could achieve the desired *redundancy reduction*. Thus, the information present in the retina, which was estimated to have a capacity of about 10^9 bits per second [37], would be recoded in a less redundant form as it travelled through the visual system. (Interestingly, Oldfield [38] also proposed a similar mechanism of adaptive redundancy reduction by successive stages of recoding, without explicitly referring to Shannon's information theory.)

Barlow [3] further argued that these lateral inhibitory connections should be *adaptive*, allowing an efficient code to be constructed based on the statistics of the actual sensory input data. There would be some genetically specified mechanisms of redundancy reduction, to deal with ever-present environmental regularities, but there would also be a learning mechanism to deal with those conditions and contingencies peculiar to the individual. While this would require a sensory system to *learn* its final redundancy-reducing encoding, it would have the advantage that this would not have to be specified in detail in the genetic material of that organism, and would allow an organism to adapt to the particular environment in which it finds itself.

Barlow and Földiák [6] suggested a neural network learning algorithm to perform this adaptive lateral inhibition. In their network, the inhibition v_{ij} between neurons i and j in the same layer is increased in proportion to the product of their activities. Mathematically, we can express this as $\Delta v_{ij} \propto x_i x_j$ where x_i and x_j are the activities of two neurons $i \neq j$, and Δv_{ij} is a small update to the inhibition v_{ij}. When this learning algorithm has converged, the activities of the neurons have been *decorrelated* from each other: that is, the correlation $E(x_i x_j)$ of activities of neurons i and j is zero if $i \neq j$, so they are now *uncorrelated*. In contrast to a *Hebbian* learning algorithm [39], whereby the strength of *excitation* between neurons is increased when they are active together, the Barlow and Földiák algorithm causes the *inhibition* to increase, so is called an *anti-Hebbian* learning algorithm.

3.4 Factorial coding

The logical conclusion of a process of successive stages of redundancy reduction would be a completely *non-redundant* code Y. The elements Y_i of Y would be statistically independent: this is called a *factorial* code, because the joint probability distribution for independent variables factorises into

$$P(Y = y) = \prod_{i=1}^{m} P(Y_i = y_i). \tag{3}$$

Harpur [40] called factorial coding the 'holy grail' of unsupervised neural network learning: this is also the goal of algorithms that perform independent component analysis [41]. It means that

all structure due to dependencies between code elements Y_i has successfully been identified, accounted for, and removed.

Barlow [9] also made an argument for factorial coding based on requirements for versatile and reliable associative learning. In order to identify an association between some stimulus y and another stimulus z, an organism must notice that y accompanies z more often than would be expected by chance if y and z occurred independently. To do this efficiently it must compare the joint probability distribution $P(y, z)$ with the product $P(y)P(z)$ of marginal probability distributions: these will be equal if y and z are independent. Thus for an association to be identified in this way, the organism must be able to estimate the marginal probabilities $P(y)$ and $P(z)$. Barlow described this in the context of a single neuron, with a number of different synapses representing the elements of Y. In this case, one can imagine that each synapse is responsible for building up a picture of $P(y_i)$ using locally available information only. If the Y_i are statistically independent, eqn (3) applies, and these marginal probabilities can then be multiplied in the cell body to yield the probability of the entire stimulus, $P(y)$. He also observed that in a dynamic environment, factorial coding means that new structure can be identified as 'new regularities' or 'suspicious coincidences' that 'betray the presence of new causal factors in the environment' [42].

This argument explicitly recognises the desirability of modelling the probability distribution of the stimulus. Factorial coding is useful precisely in that it facilitates a particularly simple computation, but other probabilistic models might do equally well. Building and using these models might form a large part of the operation of a perceptual system.

4 Information and noise in continuous signals

So far we have used a simple discrete representation for sensory information. This is consistent with early models of neurons, which represented information using binary states, such as *firing* or *not firing* [43]. In this discrete approach, entropy and information are identical, a sensory scene is assumed to have finite entropy, and hence it contains a finite amount of information. Redundancy reduction is then the process of encoding this finite amount of information as efficiently as possible, without losing any of the information originally present (an idea that is also closely related to the concept of *minimum message length* [44] or *minimum description length* [45]).

However, on closer inspection, the real world does not appear to be composed of such discrete values. For example, a typical table on which an ink bottle might rest is not a single uniform brown colour, but consists of a seemingly continuous range of subtle shades of brown. If we used the original discrete-variable formula (1), a continuous-valued source would have an infinite number of possible values it could take. Therefore we would end up with an infinite entropy, meaning that, in theory, a visual scene might contain an *infinite* amount of information! In fact, what limits the amount of information available to our sensory systems is *noise*.

For continuous variables, we use an alternative form of entropy

$$H = -\int_{-\infty}^{\infty} p(x) \log p(x)\, \mathrm{d}x. \tag{4}$$

Equation (4) is normally finite, but no longer guaranteed to be positive, and is also dependent on the scaling of variables: scaling x by a factor of n will add $\log n$ to the entropy.

The information $I(\Omega, X) = H(\Omega) - H(\Omega|X) = H(X) - H(X|\Omega)$ derived from this continuous case *is* scale independent, however, since any scaling will add the same value to both 'before' entropy $H(\Omega)$, and the 'after' entropy $H(\Omega|X)$. The entropy $H(\Omega|X)$ is the uncertainty left in Ω once we know X, and so represents the *noise* present in the observation X (Fig. 3). For example,

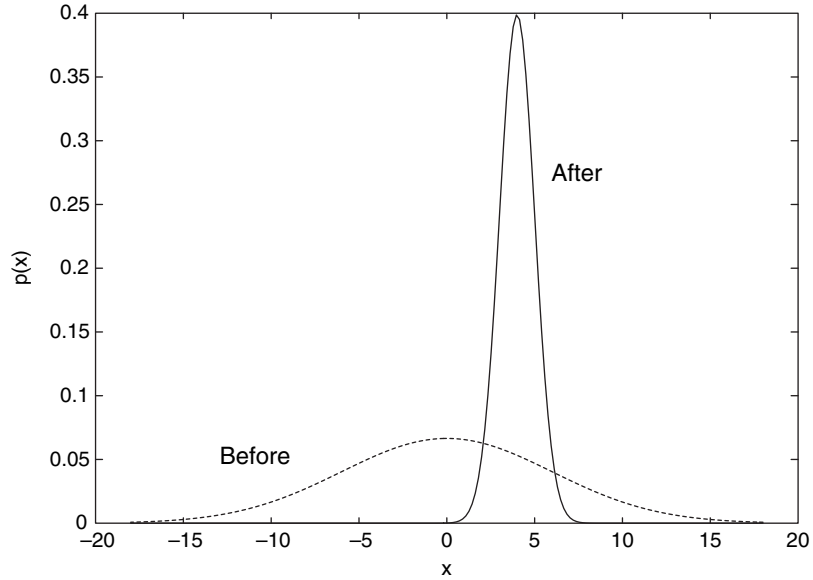


Figure 3: Probabilities of a Gaussian distribution before and after a noisy observation. The 'before' distribution has the signal entropy $H(\Omega)$, while the 'after' distribution has the noise entropy $H(\Omega|X)$ for an example observation $X = 4$.

for a Gaussian signal of variance $\sigma_S^2 = S$ and noise of variance $\sigma_N^2 = N$, we can calculate that the mean information gained from an observation is

$$I = 0.5 \log(1 + S/N),$$

where S/N is the signal-to-noise power (variance) ratio. As the noise power N goes to zero, we can see that the information gained will become infinite: therefore if we could measure a continuous quantity with complete accuracy, we would gain an infinite amount of information. Consideration of noise is therefore very important when determining the information available from a continuous value.

Strictly speaking, the 'noise' in an observation is the uncertainty due to all factors which we are not interested in. We cannot simply measure $H(X)$ any more: we either need some information about the source Ω or the noise which causes Ω and X to differ. Often noise is considered to be random effects such as thermal noise in a sensor, but in a more general case, it may simply be an aspect of a signal that we are not interested in. For example, in a cocktail party, all except one conversation going on in the room will count as 'noise' (something in X not due to Ω), while the one we are interested in will count as 'signal' Ω. Furthermore, the 'signal' may not be the obvious one, perhaps if the conversation taking place to our right is more interesting than the person talking in front of us! Nevertheless, in applying information theory to sensory perception, we normally assume that we are interested equally in all information available in a sensory input, apart from those random variations that would not tell us anything about in the outside world.

There are many sources of this noise that might corrupt our sensory signal, and so reduce the amount of information available. Some is in the outside world, such as the *shot noise* due to the individual photons that are carrying the information about brightness of a particular surface to a retina. Other noise is present in the sensory system itself, due, for example, to variability in

behaviour of neurons, or varying amounts of transmitter substance emitted when a nerve impulse arrives at a synapse. This then leads us to the conclusion that it will be difficult for a sensory system to transmit without loss *all* the information that arrives at its receptors. This leads to the proposal that a perceptual system should adapt itself to preserve the maximum amount of information: this is Linsker's *infomax* principle [4].

4.1 Infomax and information loss

If we consider the early parts of a perceptual system such as vision to be a system for transmitting information about the environment on to higher centres in the brain, it seems reasonable to suggest that the more of the available information which is transmitted, the more effective the system will be. Some visual systems may be optimised to extract information about very specific stimuli early on: an example might be the apparent 'bug detectors' observed in the frog retina [36]. For higher animals, however, it is more likely that early parts of the visual system should process all input information equally. Linsker's *Infomax* principle therefore suggests that a perceptual system should attempt to organise itself to maximise the rate of information transmitted through it [4].

As a concrete example, suppose we have real-valued signals with independent Gaussian noise on the inputs. (For biological sensory systems, which largely use spiking neurons to communicate information [46], we might consider these real-valued signals to be encoded in the firing rate of neurons.) In the case of a sensory system consisting of a linear neural network, Linsker [4] showed that Infomax leads to the well-known linear data reduction approach of *principal components analysis* (PCA), whereby variables are transformed to retain the components in the directions of the eigenvectors with largest corresponding eigenvalues, i.e. the directions with largest variance.

An alternative view, closely related to Infomax, is that we should try to *minimise* the *loss* in information about some original signal Ω as the sensory input is processed by the perceptual system or neural network. Although this approach is in many ways equivalent to Linsker's Infomax principle, it is useful in certain technical cases, and has some interesting properties [5]. If we denote the information loss about Ω across the system (Fig. 4) which transforms X to Y by

$$\Delta I_{\Omega}(X, Y) = I(X, \Omega) - I(Y, \Omega) \tag{5}$$

then it has the following properties [5]:

1. ΔI is positive across any function f where $Y = f(X)$;
2. ΔI is positive across any additive noise Φ where $Y = X + \Phi$ (Fig. 5(a));
3. ΔI is additive across a chain of transforms (Fig. 5(b)).

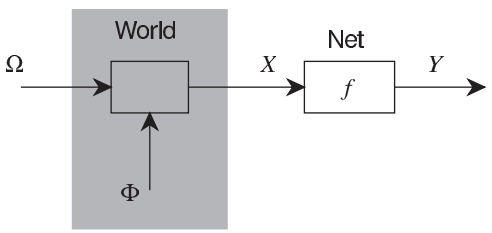


Figure 4: The original signal Ω is corrupted by irrelevant noise Φ to give the stimulus X. This is then transformed by the neural network f to give the output Y.

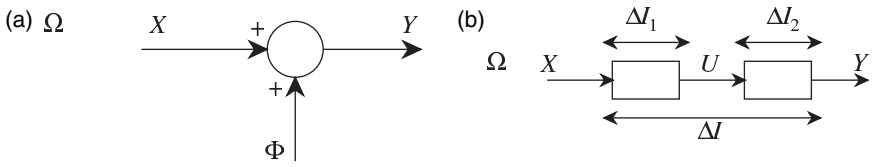


Figure 5: Information loss is (a) positive across additive noise, and (b) additive in series.

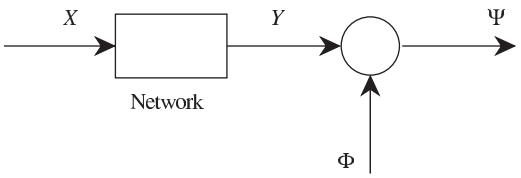


Figure 6: Neural network output signal Y is corrupted by noise Φ.

So, to minimise the information loss across a series of transforms, the information loss across each network should be minimised. Once information about any signal Ω has been lost, it is lost for good.

4.2 Information optimisation and whitening filters

Infomax tells us that PCA is useful when noise is present on the input of a linear sensory system. However, this is not the case when the noise is on the *output* of the transform, i.e. in the sensory system itself. To investigate this situation, we can instead use an approach formulated by Shannon for economical transmission of communication signals through a noisy channel [47].

Consider a sensory system which transmits its real-valued output signal Y through a channel where it will be corrupted by additive noise Φ (Fig. [6]).

If there were no restrictions on Y, we could simply amplify the signal until we had overcome as much of the noise as we like. However, suppose that there is a power cost

$$S_T = \int_0^B S(f)^2 \, \mathrm{d}f, \tag{6}$$

associated with transmitting the signal Y through the channel, where $S(f)$ is the power spectral density of the signal Y at frequency f, and B is a bandwidth limit. Then, assuming both signal and noise are Gaussian, we wish to maximise the transmitted information

$$I(\Psi, X) = \int_0^B \log \frac{S(f) + N(f)}{N(f)} \, \mathrm{d}f \tag{7}$$

for a given power cost S_T, where $N(f)$ is the power spectral density of the noise Φ.

Using the Lagrange multiplier technique, we attempt to maximise

$$J = I(\Psi, X) - \lambda S_T = \int_0^B \left(\log \frac{S(f) + N(f)}{N(f)} - \lambda S(f) \right) \mathrm{d}f \tag{8}$$

as a function of $S(f)$ for every $0 \leq f \leq B$. This is the case when

$$S(f) + N(f) = \text{constant}, \tag{9}$$

so if $N(f)$ is *white* noise, i.e. the power spectral density is uniform, or *flat*, the power spectral density $S(f)$ should also be flat [47]. A filter which performs this flattening is called a *whitening* filter. It is well known that a signal with flat power spectral density has an autocorrelation function $R_y y(\tau) = E(Y(t), Y(t+\tau))$ which is proportional to a delta function $\delta(\tau)$. In other words, the time-varying output signal $Y(t_1)$ at any time t_1 should be uncorrelated with the signal $Y(t_2)$ at any other time $t_2 \neq t_1$.

This approach leads to an equivalent condition where the signal is represented over a regular grid of units in *space* instead of *time*, such as a signal from a grid of visual receptors. In this case the signal should be transformed so that the outputs of the transform are decorrelated from each other. This decorrelation is precisely the property achieved by the Barlow and Földiák [6] algorithm that we discussed in Section 3.3.

Another way to achieve this decorrelation of outputs is to use the technique of *linear predictive coding* [48]. Consider a time-sequence of input values x_i, x_{i-1}, ... where x_i is the most recent value. We can form the least mean square (LMS) linear prediction $\hat{x}_i$ of x_i from the previous values as follows:

$$\hat{x}_i = a_1 x_{i-1} + a_2 x_{i-2} + \cdots , \tag{10}$$

where the coefficients a_j are chosen to minimise the expected squared error

$$\epsilon = E\left(x_i - \hat{x}_i\right)^2 . \tag{11}$$

Taking the derivative of this with respect to each coefficient a_j, the condition for this minimum is

$$E\left[(x_i - \hat{x}_i) x_{i-j}\right] = 0 \tag{12}$$

for all $j > 0$. If we take the residual $y_i = x_i - \hat{x}_i$ to be the output of our transform, the LMS linear prediction gives us

$$E\left[y_i x_{i-j}\right] = 0 \tag{13}$$

for all $j > 0$, and therefore

$$E\left[y_i y_k\right] = 0 \tag{14}$$

for all $k < i$, since

$$y_k = x_k - (a_1 x_{k-1} + a_2 x_{k-2} + \cdots).$$

Thus linear predictive coding will also give us the uncorrelated outputs we need.

Srinivasan *et al.* [49] suggested that predictive coding is used in the fly's visual system to perform decorrelation. They compared measurements from the fly with theoretical results based on predictive coding of typical scenes, and found reasonably good agreement at both high and low light levels. However, they did find a slight mismatch, in that the surrounding inhibition was a little more diffuse than the theory predicted.

A possible problem with the predictive coding approach is that only the *output* noise is considered in the calculation of information: the input noise is assumed to be part of the signal. At low light levels, where the input noise is a significant proportion of the input, the noise is simply considered to change the input power spectrum, making it flatter [49]. In fact, it is possible to analyse the system for both input *and* output noise (Fig. 7). We can take a similar Lagrange multiplier approach as before, and attempt to maximise transmitted information for a fixed power cost. Omitting the details, we get the following quadratic equation to solve for this optimal filter at every frequency f [50]

$$(R_c + R_r + 1)(R_c + 1) - \frac{\gamma}{N_c} R_r = 0, \tag{15}$$

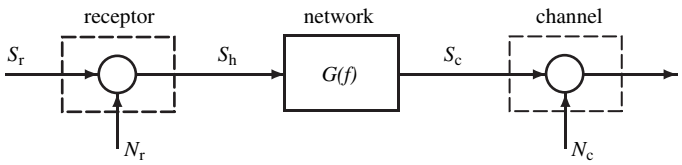


Figure 7: The signal is corrupted by both input noise before the network (or filter) and output noise after the network.

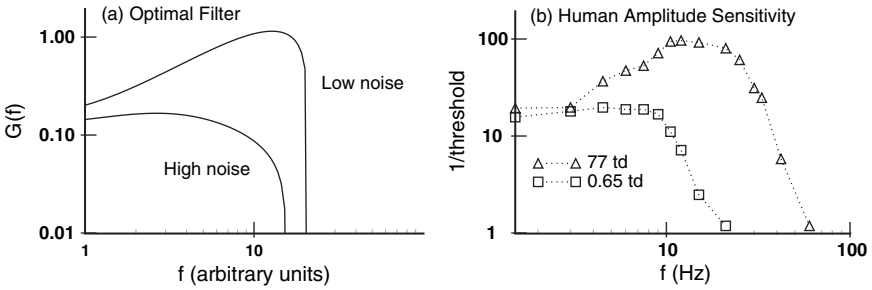


Figure 8: Optimal filter change with receptor noise level. Data in (b) adapted from [37].

where R_c is the channel signal-to-noise power spectral density ratio S_c/N_c, and R_r is the receptor signal-to-noise power spectral density ratio S_r/N_r, and γ is a Lagrange multiplier which determines the particular optimal curve to be used. This leads to a non-zero filter gain G_h whenever $R_r > [(\gamma/N_c) - 1]^{-1}$. For constant N_c (corresponding to a flat channel noise spectrum) there is therefore a certain cut-off point below which noisy input signals will be suppressed completely. A similar analysis was suggested independently by van Hateren [51].

Figure 8 compares the change in this information-optimal filter, as the input noise increases by a factor of 100 relative to the signal [50], with the effect on human visual amplitude sensitivity as the light level decreases. This approach leads to sharpening, or high-pass filtering, of a stimulus at high stimulus signal-to-noise ratios (SNRs), and 'pooling', or low-pass filtering, of a stimulus at low SNRs. Atick and Redlich [52] proposed a related approach, developed from Barlow's redundancy reduction hypothesis, but taking noise into account so that it could be applied to continuous-valued signals. Through the use of realistic stimuli and model parameters, they found a remarkable quantitative match with observed spatiotemporal response of the human visual system at various light levels [53]. Their group have since used this approach to investigate sensory coding of colour in fish and primates [54], and the behaviour of neurons with *lagged* and *non-lagged* responses of neurons in the *lateral geniculate nucleus*, the 'relay station' between the optic nerve and the visual cortex [55].

4.3 Topographic maps

So far we have looked at sensory systems where all of the channels are considered to be similar. However, biological sensory systems are often very non-uniform, concentrating resources in some areas more than others. For example, the retina is much more sensitive to detail in the centre, the *fovea*, than in the periphery. Similarly, the fingers are more sensitive to touch than the small

of the back [36]. To explore this phenomenon, we shall the consider organisation of cortical *topographic maps*.

Topographic maps are a well-known feature of sensory systems in biological organisms. These maps are organised so that nearby parts of the sensory world are represented or processed in physically nearby locations in the brain. One notable feature of these topological maps is that they can be very non-uniform in the area given over to particular parts of the sensory input. For example, the mammalian visual cortex has a very large part given over to representation of the fovea [56], and the auditory cortex of the moustached bat has a large expanded representation around 61 Hz, corresponding to the most intense harmonic of its emitted echo location pulse [57]. The ratio of cortical area to the area of sensory input which it represents is often known as the *cortical area magnification factor*, or simply the *magnification factor*.

Several approaches have been suggested for cortical map formation based on information theory. For example, Linsker [58] generated a self-organising neural map by optimising the mutual information from an input vector to the output of a network with a single layer of output neurons, arranged in a grid. Van Hulle [59] and Holthausen and Breidbach [60] also introduced information theoretic cortical map models. These models tend to use a single input value any a given time, and typically result in a magnification factor from the sensory input to the cortical map, which is proportional to the input probability density. This leads to a topographic map which has a uniform probability density for the location the single point of activation in the cortex. However, biological cortical maps are not restricted to activity in a single neuron at once, as these probability-density models tend to be. To address this, one of us [61] considered how to optimise the *information density* across a set of parallel neurons arranged in a cortical map.

We have already seen that similar Gaussian signals need to be decorrelated to make the best use of information capacity. However, we have not yet considered the possibility that these signals might have different variances, and therefore carry different amounts of information from each other. Suppose that we have n parallel sensory signals, modelled as an n-dimensional random vector Y, where each component Y_i might have a different variance. This is then transmitted via a channel which has added noise Φ, giving a received signal at the other end of the channel of

$$Z = Y + \Phi. \tag{16}$$

For simplicity, let us assume that the signal Y and noise Φ are Gaussian with covariance matrices $\mathrm{C}_Y = E(YY^T)$ and $\mathrm{C}_\Phi = E(\Phi\Phi^T)$ respectively, and that Φ is small. Then the information transmitted from Y to Z is

$$I(Z,Y) = \frac{1}{2}\log\det \mathrm{C}_Z - \frac{1}{2}\log\det \mathrm{C}_\Phi \tag{17}$$

$$\approx \frac{1}{2}\log\det \mathrm{C}_Y - \frac{1}{2}\log\det \mathrm{C}_\Phi, \tag{18}$$

with a power cost to transmit the information of $S_\mathrm{T} = \mathrm{Tr}(\mathrm{C}_Y)$ where $\mathrm{Tr}(\,\cdot\,)$ is the matrix trace operator, i.e. the sum of the diagonal entries of the matrix. Now, we wish to transmit our information as efficiently as possible, so we wish to maximise $I(Z,Y)$ for a given value of S_T. Using the technique of Lagrange multipliers as before, we therefore attempt to maximise the function

$$J = I(Z,Y) - \frac{1}{2}\lambda S_\mathrm{T}, \tag{19}$$

where λ is our Lagrange multiplier. Taking the derivative of eqn (19) with respect to C_Y and equating to zero leads to the condition [62]

$$\mathrm{C}_Y = (1/\lambda)\mathrm{I}_n, \tag{20}$$

where I_n is the identity matrix. Thus for most efficient transmission of information through the channel, we should ensure not only that the network outputs are uncorrelated, but that they also have *equal variance*, $\sigma_Y^2 = 1/\lambda$. Even if Y is non-Gaussian, we can still say that this condition maximises the information capacity $C(Z, Y) = \max_Y I(Z, Y)$ for any given power limit S_T, and that this capacity will be achieved if Y is Gaussian.

Consider now the distribution of information in these n neurons. At the optimum point, the output variance σ_Y^2 and noise variance σ_Φ^2 of all neurons is equal. Now, the information conveyed by a scalar Gaussian signal of signal variance S plus added noise variance N is $I = 0.5 \log(1 + S/N)$, so the information transmitted across the ith neuron is

$$I_i = \frac{1}{2} \log\left(1 + \sigma_Y^2/\sigma_\Phi^2\right), \tag{21}$$

with $\sigma_Y^2 = S_T/n$. Therefore, the information I_i must be equal across each neuron i.

Now suppose these neurons are arranged into a two-dimensional map, and that the neurons are all the same size, and hence packed into the map with uniform density of h neurons per unit area. Therefore the map will have information density of

$$I' = \frac{1}{2} h \log\left(1 + \sigma_Y^2/\sigma_\Phi^2\right) \tag{22}$$

$$= \frac{1}{2} h \log(1 + S_T/N_T) \tag{23}$$

bits per unit area, where $N_T = \mathrm{Tr}(C_\Phi) = n\sigma_\Phi^2$ is the total noise power. The important point here is that this optimal information density is the same everywhere. This leads to the proposal of the principle of *uniform information density* [61]: that the most efficient processing of sensory information in a topographic cortical map with uniform density of neurons will be achieved when the information density across the map is uniform.

So, if an organism needs higher sensitivity for certain ranges of its sensory input than others, the most information-efficient way to do this is to *magnify* the representation of these in the cortex, in proportion to the information density required at the sensors (Fig. 9). This fits qualitatively with the observation that cortical maps tend to have a large area allocated for sensory ranges where high accuracy or sensitivity is required, with a correspondingly smaller area where less sensitivity is required. Specifically we found that the principle of uniform information density was consistent with changes in behavioural threshold and corresponding cortical areas observed by Recanzone *et al.* [63] for owl monkeys trained on a tactile frequency discrimination task [61].

4.4 Energy efficiency and spiking neurons

It is well known that communication between neurons using spikes is widespread in biological sensory systems. These neurons communicate by firing one or more identical spikes along their axons, rather than having the type of continuous output response that we have considered up to now. As we mentioned earlier, we could simply consider the firing rate to encode a (positive) continuous variable, and use information optimisation principles developed for continuous variables. However, recently it has become increasing clear that a single spike may encode a considerable amount of information on its own [64].

Spikes might also be useful for another reason. For the early stages of a sensory system, such as the visual system, we are likely to have an *information bottleneck*, where we have a limited capacity channel, such as the optic nerve, through which all visual information has to be squeezed.

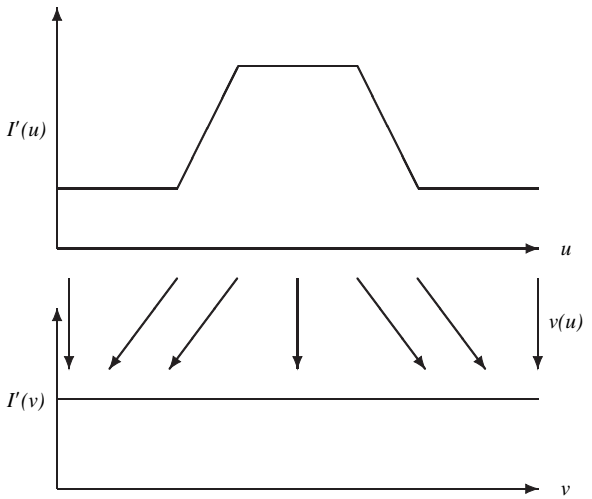


Figure 9: Magnification of input range over some regions of input u, allowing a non-uniform information density $I(u)$ over u, while we have a uniform map information density $I(v)$ over v.

It therefore seems likely that the information should be encoded as compactly as possible in order for as much information as possible to be passed on to higher processing centers. However, once the information reaches the neocortex, a huge expansion in representation is possible. This indicates that a compact representation may no longer be necessary: instead we may wish for an *energy efficient* coding, reducing the energy used by the brain to store and process the information. For example, Field [65] suggested that sensory information should be represented using *sparse* codes, in which most units are inactive, instead of compact codes: Baddeley [66] suggests that these sparse codes are essentially codes with a minimal firing rate cost, and are therefore energy-efficient. Verdú [67] considered the problem of optimising information channel capacity per unit cost, finding that the maximum can be achieved with a *binary* alphabet consisting of two symbols. One symbol should be zero cost, with the other being the lowest-cost symbol per bit of information transmitted. Thus if spikes have evolved to be the cheapest symbol available to a neuron, it may be that communication using spike-like symbols is the most energy-efficient method of encoding information.

Inspired by Barlow's [68] concept of an 'economy of impulses' in the brain, Levy and Baxter [7] considered the optimisation of the ratio of information capacity to energy, finding that the optimum ratio depended on the ratio r of energy cost for the cell to emit a spike, against that for no spike. Using values of r suggested by glucose utilisation and oxygen consumption measured in the rat, they calculated optimal firing frequencies close to those found in the rat subiculum (in the limbic system). This suggests that the brain may have evolved to be energy efficient in its handling of information, rather than simply maximising information capacity.

Balasubramanian *et al.* [69] consider two regimes for information processing in an organism: *immediate* activities such as catching prey or avoiding predators, and normal *exploratory* activity. They suggest that immediate activities require maximisation of information given the average energy available, whereas the exploratory regime requires maximisation of information per unit energy cost. They found different optimal operating conditions for the organism, depending on

the regime being used. Hence the organism concentrates its information processing capacity at critical points in time, lowering this capacity (and hence the rate of energy consumption) when this full capacity is not needed.

In an early paper considering the information transmitted by spiking neurons, MacKay and McCulloch [70] argued it is more efficient for a spiking neural system to use the time delays *between* spike firings to encode information than it is to use the presence or absence of a spike in a certain time slot. They estimated that, depending on the timing accuracy, interval modulation such as this is at least nine times more efficient per impulse. Nevertheless, a largely prevailing view in psychology was that individual spikes contained little information by themselves, and that many spikes would be needed for any useful information to be transmitted. However, using a method based on signal reconstruction, Bialek *et al.* [64] showed that a small number of individual spikes were able to reconstruct a typical sensory signal quite accurately, and that each spike contained a significant amount of information about the sensory input. They also showed that information flow in spiking neurons can be very efficient, in comparison with the theoretical limits of this type of information transmission [71]. (See also the book by Rieke *et al.* [46].)

Expanding our view to encompass sets of spiking neurons, correlations between the firing times of neurons may be important to represent sensory information [72]. Recently Deco and Schurmann [73] proposed that a system of spiking neurons should be tuned so that discrimination between different input stimuli can be made as reliable as possible as quickly as possible, a principle that they call *minimal time-maximum reliability*. Using this principle, they found that clusters of synchronised neurons emerged as representations of different stimuli, very reminiscent of the *cell assemblies* proposed by Hebb [39] many years earlier. Digging down even deeper into the mechanisms of neural information transmission, looking at the synapses that connect one neuron to another, Manwani and Koch [74] found that, while single synapses provided poor information transmission on their own, a small number of redundant synapses, as found in the typical set of axonal connections from one neuron to another, would produce robust and reliable information transmission.

Many of these more recent models make quite specific assumptions, and have different models of the biology built into their system model. Nevertheless, they all share the key issues of noise and unreliability of transmission. It will be interesting to see whether these approaches can be used to form a unified theory of biological information transmission in the not-too-distant future.

5 Discussion

5.1 Redundancy and structure

The information-theoretic idea of redundancy links back to the idea of *structure* in a perceptual stimulus that was discussed in Section 2.5. Let us model a stimulus as an n-tuple of random variables $X = (X_1, \ldots, X_n)$, where the individual components X_n can represent simultaneous observations from many receptors, or successive observations from a single receptor, or both. Hochberg's *intra-stimulus constraints* [19] imply that the observations are confined to some lower-dimensional subspace or manifold of the full space of X. In a stochastic system, this hard constraint can be replaced with a soft one: the distribution of X might not strictly occupy a lower-dimensional subspace, but it will have some sort of *shape*, and certain sorts of shape will imply statistical dependencies between the components of X. Informally then, one may characterise structure as any departure from purely 'random' or 'unstructured' statistical independence.

So, if X is unstructured, it will be impossible to predict one component from the others; that is, the conditional probabilities $P(X_i|\{X_j: j \neq i\})$ will be equal to the marginals $P(X_i)$. Conversely, structuredness implies that it is possible to make some of those predictions, or what Attneave [2] called 'better-than-chance inferences'. Some components will carry information about others, so there will be redundancy in the data. In the case of images, we have seen that structure implies that some regions of an image can be roughly predicted from others. This is also true for symmetric or periodic data, and accords with Barlow's suggestion [42] that 'structure is anything that is regular, repeating, or symmetric'.

The principle that redundancy constitutes structure agrees with many examples of what we might intuitively identify as *structure*. In images, common patterns or features, as well as being distinctive to the eye, will be a source of redundancy because the parts of the feature tend to accompany one-another: the presence of one half is a strong indicator that the other half is to be found adjacently. Objects are also a source of redundancy, since their parts always go together. Temporal coherences of many kinds, such as that found in the overall brightness of visual scenes, in the overall loudness of auditory scenes, and in the set of probable notes in tonal music, are all further forms of redundancy.

As well as the benefit of producing a more concise encoding of sensory data, the process of identifying and removing 'structuredness' requires that the structure be 'understood' in some sense. It involves an implicit probabilistic model of the data, because in an optimal code, the encoding of a message x with probability $P(x)$ is $\log_2 P(x)$ bits long. Thus, in order to optimally determine how much resource to allocate to representing the message, one must know the probability of its occurrence.

5.2 Gibson and information

The concept of *redundancy* sheds some light on Gibson's ideas about perceptual systems. One of the points he makes is that perceptual systems should not be defined in terms of sense *organs*, (that is, anatomically) but rather that different, possibly overlapping, sets of receptors work together to form perceptual systems [12]. Each perceptual system is dedicated to knowing a different aspect of the world: they are outward looking systems rather than inward looking ones. In the current context, we might say that perceptual systems are defined by statistical dependencies between receptors. For example, the vestibular and visual organs together constitute a system for the perception of orientation, because they both respond in a correlated fashion to changes of orientation. Similarly, olfactory receptors contribute both to the perception of smells and flavours, depending on whether the signals are correlated with events at the tip of the nose, such as sniffing actions, or with tactile sensations in the mouth, chewing actions, and so on. Therefore we should be interested not just in redundancy between successive symbols in the same channels, or between symbols in neighbouring channels, but also redundancy between different sensory modalities.

However, Gibson himself expresses concern about the use of information theory, as did others (see also Section 3). While he described perception as a process of 'information pick-up', he felt that the use of Shannon's mathematical framework was inappropriate [12]. He rejected the idea that 'information' is that which brings about a reduction in uncertainty, falling back on a dictionary definition: 'that which is got by word or picture', which is perhaps rather circular in the current context. He also rejected the idea that sense data can be considered as signals in a communications system ([13], p. 63), 'for these signals must be in code and therefore have to be decoded; signals are messages, and messages have to be interpreted'. In fact, he was quite dismissive about the 'vast literature nowadays of speculation about the media of communication' which he accused of being 'undisciplined and vague'.

Responding to these concerns, Barlow [42] advocates the adoption of an information theoretic approach, while at the same time acknowledging that Shannon's original exposition of the theory as a theory of *communication*, with a transmitter, a channel, and a receiver, 'is in some ways a poor analogy for the perceptual brain, partly because we must rid ourselves of the idea that there is a homunculus to receive the messages'. Even if the terminology may seem a little strained when applied to perception, and the use of the term *information* in the technical sense might not always agree precisely with our common-sense expectations, there is a case to be made that different parts of the brain are truly in 'communication' with each other, and the mathematical forms fit very well.

5.3 Noise and irrelevant information

We saw that the consideration of noise is critical to the information theoretic approach for continuous variables. Is it reasonable to assume that we know enough about the signal and noise that we are dealing with, to be able to do this?

Marr's 'fundamental hypothesis' [20] is that a perceptual system will discard information that does not impair its ability to detect 'stable features', which he defines as those which tend to co-occur. This can be imagined to be a sort of voting system amongst receptors. If only one unit reports some stimulus, it is likely to be ignored on the basis that it is probably an error, or an aberration peculiar to the internal workings of that unit. On the other hand, if many units show concerted activity, it is more likely that this is due to events in the outside world.

Marr's hypothesis amounts to an assumption about the structure of noise, namely, that noise affects each receptor independently of the others, while real signals affect many receptors together. Consequently, this hypothesis determines which features in the input are going to be important: those which are robust to such noise. It also corresponds with the idea of redundancy: stable features are encoded redundantly, whereas noise is not. Attneave [2] also suggests that uncorrelated sensory signals tend to be interpreted as noise. He observes that even though uncorrelated noise has a high entropy, and thus could potentially carry a lot of information, both visual and auditory white noise have a dull, uniform, and uninformative texture. Redlich [75] makes a similar observation: 'Desirable input information is often encoded redundantly, so redundancy can be used to distinguish true signal from noise.'

Thus we have empirical observations that noise *tends* to appear to be non-redundant and *structureless*, implying that any structuring will not be noise and therefore will be relevant. This is quite a strong statement which somehow goes against the ecological principles outlined in Section 2.6. Relevance, or attensity, is ultimately something which cannot be decided without considering the use to which the information will be put. Atick and Redlich [52] comment that 'neither *noise* nor *signal* is a universal concept, but depend on what is useful visual information to a particular organism in a particular environment'. This *usefulness* is unlikely to be determined from the sensory input data alone. Although much of the work on information theory discussed so far concentrates on bottom–up processing, organisms are likely to have evolved top–down mechanisms, such as *attention*, that allow them to extract this useful information while ignoring information that is not so useful. Nevertheless, the best that we can do for the moment appears to be to assume that information about anything with structure might be useful, so a perceptual system should aim to keep it.

5.4 Uniform information, attention, and active perception

The principle of uniform information density for cortical maps suggests that the most efficient way to hand information is if it is spread evenly across the cortical surface. The same is likely

to hold for time: the most efficient way to process information is to take it in evenly, at a steady rate. In an exploratory system, a visual system with a fovea can be used to scan over a new scene, gradually transmitting the information found like the flying-spot scanner of a television camera. If, instead, the visual system had a retina which had the acuity of the fovea everywhere, not only would a huge number of additional receptors and optic nerve fibres be needed, but we would be confronted with a huge amount of new information on first seeing a scene, with this huge information capacity being subsequently wasted. Only when something moved in the scene would more information be available, and organisms have typically evolved to move their fovea to attend to movement, ready to extract the new information available.

However, the eye is not scanned randomly or regularly over a scene. Perhaps it is 'looking' to extract from a scene the most information (or the most information relevant to some task) in the shortest possible time, even if this process may be subconscious. We can take this even farther if we physically approach an object, and turn it over to see a side which would otherwise be hidden: clearly a process of active perception [13]. This aspect is currently missing from the information theory approaches that we have discussed in this chapter. They tend to assume that information about everything in our observation X (except random noise on the receptors) is equally important, whereas for this type of active perception we would need some sort of top-down information to tell is what in X is likely to give us information about the Ω that we wish to know about, and what can be ignored.

A few researchers have begun to use information theory to investigate the efficiency of *motor output* [76]. This approach may help to explain features of skills learning or muscle operation [77]. The use of Shannon's information to investigate active perception may need to bind these two approaches together, into a theory that considers perception much more as a complete action/observation loop, with information flowing both into and out of an organism, rather than simply a process of passive observation.

6 Conclusion

People have been trying to understand sensory perception for hundreds, if not thousands, of years. Traditionally, theories of perception have been expressed in terms of objects, qualities, form, invariants, and so on. Gestalt theory, for example, proposed that perception selects interpretations with good *prägnanz*, or form, if several interpretations are possible, filling in missing information in the process.

In contrast, the ecological perspective asks the following question: What is perception *for*? This leads us to the proposal that perception must be *useful* for something, requiring us to find a way to measure that 'usefulness'. From a design viewpoint, if we have a measure of the usefulness of perception, we can start to consider what we consider perception *should* do, in order to achieve the goal of being useful.

The measure of usefulness explored here is Shannon's *information*. Optimisation of information leads to various alternative principles that a perceptual system should adhere to, including redundancy reduction, maximisation of transmitted information (Infomax), factorial coding, uniform information density, or energy efficiency, depending on the circumstances. These principles have been applied to binary and discrete-valued models, real-valued models, and more realistic models of spiking neural sensory systems. While this may seem a large range of alternative principles, they are in fact different aspects of the same underlying principle of information optimisation, albeit with different costs and constraints.

There are still many challenges remaining to be faced in the application of information theory to sensory perception. For example, more needs to be know about the implications of this approach for spiking systems. In addition, the application of information theory to active perception, where the organism is not considered to be simply a passive observer of the world, seems yet to be explored significantly, if at all. So, while information theory has already provided many insights into the design of perceptual sensory systems, there is still a long way to go before we can say that we really understand sensory perception.

Acknowledgements

S.A. was supported by a Research Studentship from the UK Engineering and Physical Sciences Research Council (EPSRC). Part of this work was supported by EPSRC research grant GR/R54620.

References

[1] Shannon, C.E., A mathematical theory of communication. *Bell System Technical Journal*, **27**, pp. 379–423, 623–656, 1948.
[2] Attneave, F., Some informational aspects of visual perception. *Psychological Review*, **61**, pp. 183–193, 1954.
[3] Barlow, H.B., Possible principles underlying the transformation of sensory messages. *Sensory Communication*, ed. W.A. Rosenblith, MIT Press: Cambridge, MA, pp. 217–234, 1961.
[4] Linsker, R., Self-organization in a perceptual network. *IEEE Computer*, **21(3)**, pp. 105–117, 1988.
[5] Plumbley, M.D. & Fallside, F., An information-theoretic approach to unsupervised connectionist models. *Proceedings of the 1988 Connectionist Models Summer School*, eds. D. Touretzky, G. Hinton & T. Sejnowski, Morgan-Kaufmann: San Mateo, CA, pp. 239–245, 1988.
[6] Barlow, H.B. & Földiák, P., Adaptation and decorrelation in the cortex. *The Computing Neuron*, eds. R. Durbin, C. Miall & G. Mitchison, Addison-Wesley: Wokingham, England, pp. 54–72, 1989.
[7] Levy, W.B. & Baxter, R.A., Energy efficient neural codes. *Neural Computation*, **8**, pp. 531–543, 1996.
[8] Fodor, J., *The Modularity of Mind*, MIT Press: Cambridge, MA, 1983.
[9] Barlow, H.B., Conditions for versatile learning, Helmholtz's unconscious inference, and the task of perception. *Vision Research*, **30**, pp. 1561–1571, 1990.
[10] Berkeley, G., *A Treatise Concerning the Principles of Human Knowledge*, Brown and Sons, 1937/1734.
[11] Southall, J.P.C. (ed.), *Helmholtz's Treatise on Physiological Optics*, Dover Publications: New York, 1962. (Translation of the 1910 original.)
[12] Gibson, J.J., *The Senses Considered as Perceptual Systems*, Houghton Mifflin: Boston, MA, 1966.
[13] Gibson, J.J., *The Ecological Approach to Visual Perception*, Houghton Mifflin: Boston, MA, 1979.
[14] Bregman, A.S., *Auditory Scene Analysis*, MIT Press: Cambridge, MA, 1990.

[15] Mach, E., *The Analysis of Sensations*, Dover Publications: New York, 1959. (Translated by C.M. Williams, 1st German edn, 1886.)
[16] Köhler, W., *Gestalt Psychology*, Liveright: New York, 1947.
[17] Deutsch, D., Grouping mechanisms in music (Chapter 9). *The Psychology of Music*, 2nd edn, ed. D. Deutsch, Academic Press: San Diego, CA, 1999.
[18] Warren, R.M., Perceptual restoration of missing speech sounds. *Science*, **167**, 1970.
[19] Hochberg, J., Levels of perceptual organization (Chapter 9). *Perceptual Organization*, eds. M. Kubovy & J.R. Pomerantz, Erlbaum: Hillsdale, NJ, 1981.
[20] Marr, D., *Vision*, Freeman: San Francisco, CA, 1982.
[21] Richardson, M.W., Multidimensional psychophysics. *Psychological Bulletin*, **35**, pp. 659–660, 1938.
[22] Davidson, M.L., *Multidimensional Scaling*, John Wiley & Sons: New York, NY, 1983.
[23] Shepard, R.N., Psychological complementarity (Chapter 10). *Perceptual Organization*, eds. M. Kubovy & J.R. Pomerantz, Erlbaum: Hillsdale, NJ, 1981.
[24] Knill, D.C. & Richards, W. (eds.), *Perception as Bayesian Inference*, Cambridge University Press: Cambridge, UK, 1996.
[25] Leman, M. & Carreras, F., The self-organization of stable perceptual maps in a realistic musical environment. *Proc. 3emes Journées d'Informatique Musicale (3rd Computer Music Conference), JIM96*, Caen, 1996.
[26] Shepard, R.N., Cognitive psychology and music (Chapter 3). *Music, Cognition, and Computerised Sound*, ed. P.R. Cook, MIT Press: Cambridge, MA, pp. 21–35, 1999.
[27] Locke, J., *An Essay Concerning Human Understanding*, Dent and Sons, 1961. (First published 1706.)
[28] Shaw, R.E., McIntyre, M. & Mace, W.M., The role of symmetry in event perception. *Perception: Essays in Honour of James J. Gibson*, eds. R.B. MacLeod & H. Pick, Cornell University Press: Ithica, NY, 1974.
[29] Windsor, W.L., *A Perceptual Approach to the Description and Analysis of Acousmatic Music*, PhD thesis, City University, London, 1995.
[30] Miller, G.A. & Frick, F.C., Statistical behaviouristics and sequences of responses. *Psychological Review*, **56**, pp. 311–324, 1949.
[31] Miller, G.A., What is information measurement? *American Psychologist*, **8**, pp. 3–11, 1953.
[32] Barlow, H.B., Sensory mechanisms, the reduction of redundancy, and intelligence. *Proceedings of a Symposium on the Mechanisation of Thought Processes*, National Physical Laboratory, Teddington, Her Majesty's Stationery Office: London, Vol. 2, pp. 537–559, 1959.
[33] Green, R.T. & Courtis, M.C., Information theory and figure perception: The metaphor that failed. *Acta Psychologica*, **25**, pp. 12–36, 1966.
[34] Shannon, C.E., Prediction and entropy of printed English. *Bell System Technical Journal*, **30**, pp. 50–64, 1951.
[35] Hubel, D.H. & Wiesel, T.N., Receptive fields of single neurones in the cat's striate cortex. *Journal of Physiology*, **148**, pp. 574–591, 1959.
[36] Kuffler, S.W., Nicholls, J.G. & Martin, A.R., *From Neuron to Brain: A Cellular Approach to the Function of the Nervous System*, 2nd edn, Sinauer Associates Inc.: Sunderland, MA, 1984.
[37] Kelly, D.H., Information capacity of a single retinal channel. *IRE Transactions on Information Theory*, **IT-8**, pp. 221–226, 1962.
[38] Oldfield, R.C., Memory mechanisms and the theory of schemata. *British Journal of Psychology*, **45**, pp. 14–23, 1954.

[39] Hebb, D.O., *The Organization of Behavior*, Wiley: New York, 1949.
[40] Harpur, G.F., *Low Entropy Coding with Unsupervised Neural Networks*, PhD thesis, Cambridge University Engineering Department, 1997.
[41] Hyvärinen, A., Karhunen, J. & Oja, E., *Independent Component Analysis*, John Wiley & Sons: New York, 2001.
[42] Barlow, H.B., Banishing the homonculus. *Perception as Bayesian Inference*, eds. D.C. Knill & W. Richards, Cambridge University Press: Cambridge, UK, p. 425, 1996.
[43] McCulloch, W.S. & Pitts, W., A logical calculus of ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics*, **5**, pp. 115–133, 1943.
[44] Wallace, C.S. & Boulton, D.M., An information measure for classification. *The Computer Journal*, **11(2)**, pp. 185–194, 1968.
[45] Rissanen, J., Modeling by shortest data description. *Automatica*, **14**, pp. 465–471, 1978.
[46] Rieke, F., Warland, D., de Ruyter van Steveninck, R. & Bialek, W., *Neural Coding*, MIT Press: Cambridge, MA, 1996.
[47] Shannon, C.E., Communication in the presence of noise. *Proceedings of the IRE*, **37**, pp. 10–21, 1949.
[48] Papoulis, A., *Probability, Random Variables and Stochastic Processes*, 2nd edn, McGraw-Hill: New York, 1984.
[49] Srinivasan, M.V., Laughlin, S.B. & Dubs, A., Predictive coding; a fresh view of inhibition in the retina. *Proceedings of the Royal Society of London, Series B*, **216**, pp. 427–459, 1982.
[50] Plumbley, M.D. & Fallside, F., The effect of receptor signal-to-noise levels on optimal filtering in a sensory system. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, ICASSP-91*, Vol. 4, pp. 2321–2324, 1991.
[51] van Hateren, J.H., A theory of maximizing sensory information. *Biological Cybernetics*, **68**, pp. 23–29, 1992.
[52] Atick, J.J. & Redlich, A.N., Towards a theory of early visual processing. *Neural Computation*, **2**, pp. 308–320, 1990.
[53] Atick, J.J., Could information theory provide an ecological theory of sensory processing? *Network: Computation in Neural Systems*, **3(2)**, pp. 213–251, 1992.
[54] Atick, J.J., Li, Z.P. & Redlich, A.N., Understanding retinal color coding from first principles. *Neural Computation*, **4**, pp. 559–572, 1992.
[55] Dong, D.W. & Atick, J.J., Temporal decorrelation: a theory of lagged and nonlagged responses in the lateral geniculate nucleus. *Network: Computation in Neural Systems*, **6**, pp. 159–179, 1995.
[56] Rovamo, J., Virsu, V. & Näsänen, R., Cortical magnification factor predicts the photopic contrast sensitivity of peripheral vision. *Nature*, **271**, pp. 54–56, 1978.
[57] Suga, N., Cortical computational maps for auditory imaging. *Neural Networks*, **3**, pp. 3–21, 1990.
[58] Linsker, R., How to generate ordered maps by maximising the mutual information between input and output signals. *Neural Computing*, **1**, pp. 402–411, 1989.
[59] Van Hulle, M.M., Topology-preserving map formation achieved with a purely local unsupervised competitive learning rule. *Neural Networks*, **10**, pp. 431–446, 1997.
[60] Holthausen, K. & Breidbach, O., Self-organized feature maps and information theory. *Network: Computation in Neural Systems*, **8**, pp. 215–227, 1997.
[61] Plumbley, M.D., Do cortical maps adapt to optimize information density? *Network: Computation in Neural Systems*, **10**, pp. 41–58, 1999.

[62] Plumbley, M.D., Efficient information transfer and anti-Hebbian neural networks. *Neural Networks*, **6**, pp. 823–833, 1993.
[63] Recanzone, G.H., Jenkins, W.M., Hradek, G.T. & Merzenich, M.M., Progressive improvement in discriminative abilities in adult owl monkeys performing a tactile frequency discrimination task. *Journal of Neurophysiology*, **67**, pp. 1015–1030, 1992.
[64] Bialek, W., Rieke, F., de Ruyter van Steveninck, R. & Warland, D., Reading a neural code. *Science*, **252**, pp. 1854–1857, 1991.
[65] Field, D.J., What is the goal of sensory coding? *Neural Computation*, **6**, pp. 559–601, 1994.
[66] Baddeley, R., An efficient code in V1? *Nature*, **381**, pp. 560–561, 1996.
[67] Verdú, S., On channel capacity per unit cost. *IEEE Transactions on Information Theory*, **36**, pp. 1019–1030, 1990.
[68] Barlow, H.B., Trigger features, adaptation and economy of impulses. *Information Processing in the Nervous System*, ed. K.N. Leibovic, Springer-Verlag: New York, pp. 209–226, 1969.
[69] Balasubramanian, V., Kimber, D. & Berry II, M.J., Metabolically efficient information processing. *Neural Computation*, **13**, pp. 799–815, 2001.
[70] MacKay, D.M. & McCulloch, W.S., The limiting information capacity of a neuronal link. *Bulletin of Mathematical Biophysics*, **14**, pp. 127–135, 1952.
[71] Bialek, W., DeWeese, M., Rieke, F. & Warland, D., Bits and brains: information flow in the nervous system. *Physica A*, **200**, pp. 581–593, 1993.
[72] Panzeri, S., Schultz, S.R., Treves, A. & Rolls, E.T., Correlations and the encoding of information in the nervous system. *Proceedings of the Royal Society Series B: Biological Sciences*, **266**, pp. 1001–1012, 1999.
[73] Deco, G. & Schürmann, B., Spatiotemporal coding in the cortex: Information flow-based learning in spiking neural networks. *Neural Computation*, **11**, pp. 919–934, 1999.
[74] Manwani, A. & Koch, C., Detecting and estimating signals over noisy and unreliable synapses: Information-theoretic analysis. *Neural Computation*, **13**, pp. 1–33, 2000.
[75] Redlich, A.N., Redundancy reduction as a strategy for unsupervised learning. *Neural Computation*, **5**, pp. 289–304, 1993.
[76] Haken, H., Information compression in biological systems. *Biological Cybernetics*, **56**, pp. 11–17, 1987.
[77] Senn, W., Wyler, K., Clamann, H.P., Kleinle, J., Lüscher, H.R. & Müller, L., Size principle and information theory. *Biological Cybernetics*, **76**, pp. 11–22, 1997.

This page intentionally left blank

Chapter 8

Flight

R.J. Wootton
School of Biosciences, University of Exeter, UK.

Abstract

Most animal species fly. Insects, birds, bats and the extinct pterosaurs evolved powered flight independently, and each group has developed a wide range of adaptations, exploiting the many special opportunities that flight provides. Differing life styles need different flight capabilities and levels of performance. These require a variety of techniques, many employing aerodynamic principles unknown in orthodox engineering; and these in turn are reflected in the animals' physiology and design: overall structure, size, wing shape and proportions. Understanding the relationships between these variables and the behaviour, life histories, ecology and evolution of the animals is a difficult but fascinating challenge, involving a network of interconnecting experimental, observational and theoretical approaches.

1 Introduction

1.1 Which organisms fly?

Flight is widespread in the living world. Powered flight has arisen at least four times: in insects, pterosaurian reptiles, birds, and once – perhaps twice – in mammals; opinion is divided whether bats have a single or a dual origin [1, 2]. Gliding flight, or at least paragliding, has appeared in two families of fish, in several frogs, in snakes, at least five times in other reptiles, three times in marsupials, and at least three times among placental mammals [1, 3]. Gliding seeds, or strictly fruits, are found in some vines (e.g. *Alsomitra*), and samaras – fruits that rotate like the vanes of autogyros – have evolved independently in conifers, in maples, and in several other plant families [4, 5]. Furthermore many aquatic animals – pteropod snails, some squid, whales, turtles, some birds and many groups of fish – 'fly' through the water using tails or paired fins as flapping hydrofoils and employing the same fluid dynamic principles as aerial fliers. Aquatic locomotion is covered elsewhere in this series; the present account is restricted to flight in air.

Flight has unquestionably contributed hugely to the evolutionary success of the groups that employ it, allowing ready, fast dispersal and access to habitats and food sources in the three dimensions of space, with consequent niche diversification. This is reflected in the diversity of the

animals themselves: in terms of described species, insects are by far the largest class of animals, birds the largest class of terrestrial vertebrates, and bats the second largest order of mammals.

1.2 What is flight?

Animals and plants are heavier than air. To stay aloft they need to elicit from the air a vertically upward, supporting force at least equal to the force that gravity is exerting on them – their weight. Newton's first law of motion tells us that this would also be true if they were falling at a constant speed. In this case the upward force would be the *drag* – defined as the *aerodynamic force acting back along the direction of movement* (Fig. 1a). The high surface/volume ratio of many tiny animals and plant propagules ensures high drag values for their weight, and since drag increases with velocity these reach force equilibrium at low sinking speeds, and are often kept aloft and sometimes carried to high altitudes and over long distances by air that is rising faster than they are falling.

Bigger animals and large seeds, however, have relatively less drag for their weight, and their terminal falling velocities are much higher. To remain up in the air and travel horizontally they need to generate *lift* – defined as the *aerodynamic force operating perpendicularly to the direction of movement.* Contrary to common usage, which is influenced by our experience of fixed wing aircraft, lift does not necessarily act vertically upwards; and this is crucial to our understanding of animal flight. Lift on a sail is nearly horizontal, that on a stooping falcon may be more horizontal than vertical. Nonetheless it is the vertically upward component of lift that makes flight possible.

Lift is provided by aerofoils – wings. In conventional aircraft, gliding animals and winged fruits, the role of wings is to generate an overall aerodynamic force with a large enough vertical component to balance the weight. For this they need to be pulled or actively driven through the air.

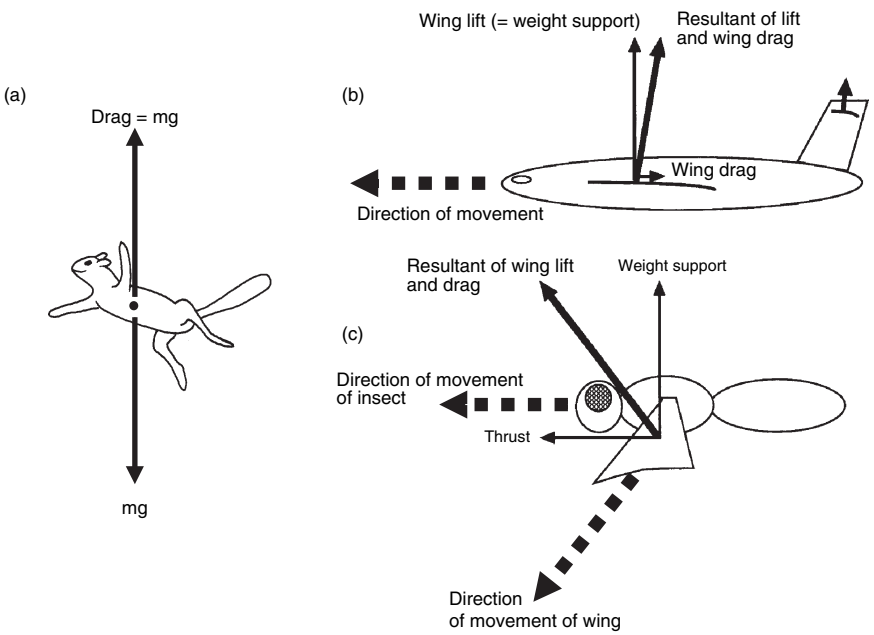


Figure 1: (a) Forces on an animal falling at a steady speed: mg, weight. Forces generated by the wings in (b) an aircraft and (c) a flapping animal in mid-downstroke.

In gliding (Fig. 3) the propulsive force is provided by gravity; in powered aircraft (Fig. 1b) by jets or propellors. In powered animal flight the wings themselves are the propulsive structures, and propulsion is achieved by active muscular flapping: by cyclically oscillating, twisting and deforming the wings so that they provide not only sufficient net upward force, but also enough *thrust* in the direction of movement to counteract overall drag (Fig. 1c).

1.3 The generation of lift

Lift is created by accelerating air. By Newton's second law of motion, the magnitude of the force F is proportional to the mass m of air accelerated, and to the acceleration a.

$$F = ma.$$

Wings generate lift by creating vorticity in the air around them. In simplest terms, a vortex is a circulation of air around a linear axis: familiar examples are whirlwinds and smoke-rings. Both examples illustrate that vortices have their own momentum; once created, they are capable of travelling when detached from the agent which formed them, and they significantly influence the flow of the surrounding air.

When an appropriately orientated wing is accelerated in air the viscosity of the air immediately adjacent to the wing causes a linear *starting vortex* to develop along the posterior margin – the trailing edge (Fig. 2a). Such a vortex cannot exist in isolation, any more than a cogwheel can

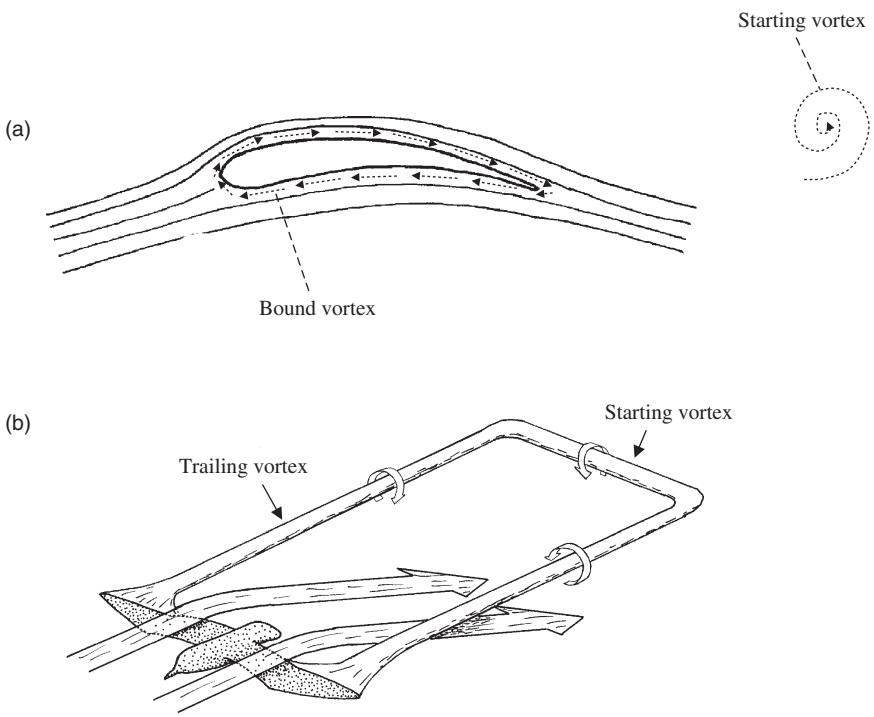


Figure 2: Diagrammatic representations of vorticity and airflow around a (a) wing section and (b) gliding bird.

rotate without turning any other that may be engaged with it. Vorticity rotating in the opposite direction must simultaneously be created; and this takes the form of a *bound vortex*, circulating round the wing, backward over the upper surface, forward under the lower. The starting vortex is left behind, but the bound vortex remains with the wing, and serves to speed up the adjacent airflow past the upper surface, and to slow down that past the lower. This results in a reduction in the pressure on the upper surface and an increase in that on the lower: the wing is both sucked and pushed upward. Simultaneously, the air flowing over the wing is deflected downward behind it; the acceleration already mentioned as essential to the generation of lift. The distribution of pressure varies along the wing, and this leads to vorticity being shed along the trailing edge and winding up at the wing tips to form a pair of *tip* or *trailing vortices* that are left behind, and ultimately link the bound vortex with the remains of the starting vortex that was shed at the onset of lift generation (Fig. 2b). The tip vortices behind high altitude aircraft are familiar to us: the low pressure in their core causes the water vapour from combustion in the engines to condense out as vapour trails.

This account describes the situation where the wings are held still, as in a gliding bird. Even here, unless the flight is entirely steady and level the strength of the bound circulation, and hence the lift, will change periodically, and further vorticity will be shed by the wings to bring this about. In flapping flight, the wings are constantly accelerating, decelerating, twisting and changing shape, and the aerodynamics of flapping flight are complex and still being actively investigated [4–13]. Many of the same principles apply. In the downstroke the wing circulation creates lift with a strong upward weight-supporting component and usually some propulsive thrust, sometimes greatly enhanced by additional *unsteady* vorticity resulting from the wings' abrupt acceleration. However, the nature of the upstroke, the form of the wake, and the strength and direction of the forces generated vary considerably: both between species, and in an individual animal according to its current speed and behaviour – see Section 3.2.2.

1.4 Stability, and the control of manoeuvres

Flight needs to be stable. Overall flight paths must follow the intentions of the animal, and any significant departures must be correctable. Stability may be *physical:* built into the design, so that momentary deviations from the flight path automatically generate restoring forces. Physical stability governs the flight of an arrow, a dart, a chuck-glider. In animals it may be provided by a long abdomen, as in most dragonflies (Odonata), or long caudal filaments, as in mayflies (Ephemeroptera) and many Palaeozoic insects, or by a long tail, as in the Jurassic protobird *Archaeopteryx* and many early pterosaurs. Early flying animals, like traditional aircraft, seem to have made extensive use of physical stability, requiring only minor active adjustments to perform simple manoeuvres and to cope with transient air currents and gusts. Physical stability, though, tends to limit agility, since the passive restoring forces oppose any changes in direction; and in many more developed animals, as in modern high-performance aircraft, it is sacrificed in favour of *active* stabilisation [14–18]. The tails of bats and advanced pterosaurs were shortened or lost. The tail skeleton of birds became reduced to a compact pygostyle supporting actively movable feathers. Manoeuvrable insects, such as advanced flies, bees, wasps and moths, have relatively compact abdomens. Flight direction and velocity are now continuously monitored by sensors, feeding back via the nervous system to the muscles controlling the movements and orientation of the wings and of other surfaces and structures: the tail feathers of birds, the caudal patagium of bats. In most advanced insects the wings are the only effective controlling surfaces, and precise, often spectacular manoeuvres are achieved wholly by changes in the amplitude, timing, shape and symmetry of the stroke, under complex neurosensory control.

Although highly agile flight may not have been possible until such neurosensory systems had evolved, the relative contributions of passive and active stabilisation are not a simple function of evolutionary 'advancement'. Flight manoeuvrability is an aspect of overall lifestyle, and the degree of physical stability may vary substantially within a single group. In butterflies, for example, it is influenced by the relative positions of the centres of wing lift and of mass along the body axis, and is significantly correlated with the degree of distastefulness to potential predators [19].

2 The origins of flight

The beginnings of flight in insects and the flying vertebrate groups have been, and still are, extensively debated.

Flight is only possible if the wings move fast enough relative to the air to generate sufficient lift. In each group, the wings of the pioneering fliers would have been small and relatively ill adapted, and the minimum airspeed for flight correspondingly high.

The most obvious way for them to achieve flight speed would be to use gravity: to climb a tree or a high rock and to jump. Some lift can then be achieved even without wings, as do parachutists in free fall. A patagium of skin between the legs, as in several groups of mammals, or between the ribs, as in the lizard *Draco* and several Mesozoic relatives; or even the toes, as in the flying frog *Rhacophorus*, increases both lift and drag, slowing the fall and allowing significant forward travel, in the continuum between parachuting and gliding. Progressive elongation and specialisation of the forelimbs and patagium or feathers of flying vertebrate ancestors, and development of the thoracic muscles and nervous system would lead to fully established flapping flight. This route – the 'trees-down' hypothesis – seems adequate to account for flight in bats [20]. *Cynocephalus*, the colugo or flying 'lemur' (order Dermoptera), which has an extensive patagium extending between fore and hind legs and, bat-like, between the fingers, appears on genetic as well as morphological evidence to be the closest relative of the macrochiropteran bats at least [1] and probably of bats as a whole [2].

For birds, for the Mesozoic Pterosauria and for insects the evidence is more equivocal. For birds, the trees-down scenario, through parachuting and gliding to flapping flight, is again a strong candidate [21] but it has serious competition. A strong case can be made for bird flight developing in fast running, feathered dinosaurs with particularly long feathers on the forelimbs [22, 23]. These might initially assist in prey capture, but would progressively come to allow short glides over the ground, and eventually enlarge to become full-scale flapping wings. A weakness in this – the 'ground-up' hypothesis – is that the propulsive force would cease, and the bird would begin to slow down, as soon as its feet left the ground, but this does not rule out the theory. Running and gliding might still be advantageous over rough terrain. The take-off speed could be increased and maintained by active flapping [24]. Recent spectacular discoveries of small, feathered, dinosaurs in early Cretaceous rocks in China [25], including one 'four-winged' example, with fully-feathered hind legs [26], show that the question is still open; and indeed the two theories are perhaps less distinct than usually considered: a cursorial protobird using gravity to gain speed in gliding from rocky outcrops would have elements of both scenarios. For pterosaurs, too, both trees-down and ground-up theories have support, though the trees-down view predominates [27]. A small Triassic patagiate reptile from the Kirghiz Republic, *Sharovipteryx mirabilis*, may be close to the ancestry of the group [28].

For insects, the ground-up scenario is less probable, as their low mass and relatively high drag would slow them rapidly after take-off. The trees-down hypothesis is still perhaps the majority view [29–31], but a recently proposed alternative theory derives from observations on some

modern stoneflies (Plecoptera) and mayflies (Ephemeroptera) that skim on the surface of water, flapping their reduced wings to propel them. This 'skimming theory' proposes that the ancestors of winged insects behaved similarly, and that their wings became progressively enlarged to the point when they were able to leave the surface and fly [32]. Both theories have authoritative support. Their implications on the later development of insect flight are rather different [33] but either hypothesis seems plausible on the existing evidence.

3 Flight roles and techniques

3.1 The functions of flight

It seems clear that flight would initially have been limited to short journeys and excursions, escaping predators and competitors, seeking food, mates, shelter, and nesting and oviposition sites. As performance and techniques evolved and diversified, new roles would progressively have become available, and modern forms exploit flight in many ways.

Many birds and insects, and most bats, catch prey on the wing: by flying continuously (hawking) like swifts, martins, falcons, insectivorous bats, and many dragonflies; or by darting from perches like flycatchers (Muscicapidae), some bats, calopterygid and libellulid dragonflies, and some wasps and predatory flies. Many flying insects, birds and bats snatch prey from surfaces, often showing great precision and fine control in avoiding collision, as do some pseudostigmatid dragonflies and vespertilionid bats that take spiders from their webs. Some birds and bats snatch fish and pelagic invertebrates from the surface waters in flight; with beaks, like petrels and skimmers, or using feet, as do ospreys, fishing eagles, and *Noctilio* and some *Myotis* species among bats. Hummingbirds, some bats, and a variety of moths, bees and flies take nectar from flowers on the wing. Advanced flying insects, particularly among dragonflies, flies and butterflies, use flight in territory surveillance and defence, courtship display, aerial combat, mate guarding. Dragonflies mate in flight and fly 'in tandem', and males of some species attack tandem pairs in competition for the females, using tactics recalling the aerial piracy behaviour used by frigate birds and skuas to steal food from other birds on the wing. Improved efficiency has allowed many birds, some bats, locusts, some dragonflies, moths and butterflies to exploit the possibilities that flight offers for long distance migration, often making use of winds and thermal updraughts. Thermals and other upcurrents allow many raptorial and carrion-feeding birds to soar for long periods in search of food on the ground, and gulls and frigate birds, albatrosses and petrels to glide at length over cliffs, islands and the sea.

3.2 Categories of flight

Such a variety of flight activities and functions require a wide range of techniques.

3.2.1 Gliding and soaring

Gliding is powered by gravity alone. Lift from the wings is required; drag, impeding the movement, is inevitable; and the path of the glide depends on the ratio between these two. Figure 3 represents an animal gliding at a constant speed in still air. Its direction relative to the horizontal is given by the *glide angle* Θ. Its weight – its mass m multiplied by the gravitational acceleration g – must be equalled by the net aerodynamic force F_A acting vertically upwards. F_A is the resultant of the drag D, acting back along the glide path, and the lift L, at right angles to the flight path. L and D are the sides of a rectangle of forces, of which F_A is the diagonal. Examination will show that the

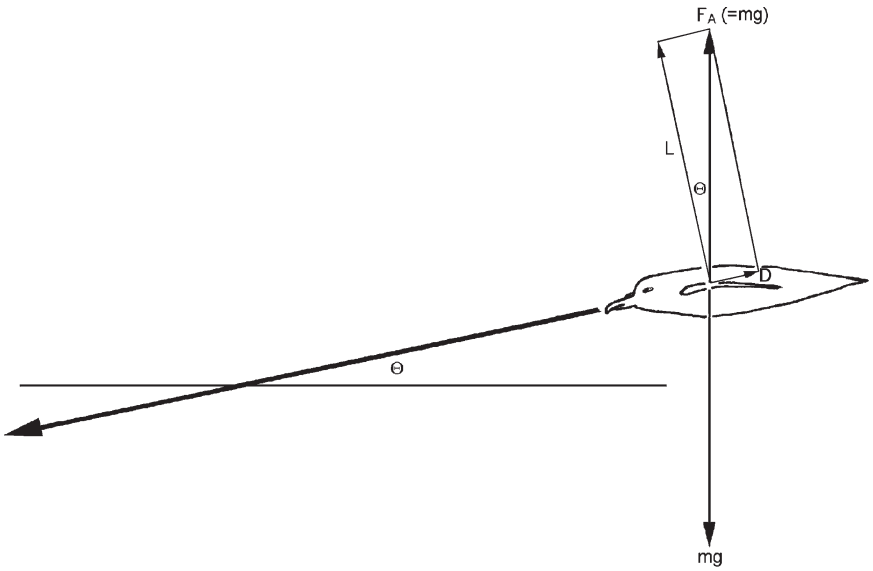


Figure 3: The forces acting on a gliding bird: F_A, net aerodynamic force; mg, weight; L, lift; D, drag; Θ, glide angle.

angle between L and F_A is equal to Θ, the glide angle, demonstrating that the latter is determined by L/D, known as the *glide number*. Using trigonometry, we find that:

$$L = mg \cos \Theta,$$
$$D = mg \sin \Theta,$$

and hence

$$L/D = \cot \Theta.$$

In consequence, a poor glider whose drag equalled the lift it was producing would fall at 45° to the horizontal. A glider with $L/D < 1$ would fall still more steeply, while $L/D > 1$ would give a more shallow glide path.

While 'flying' frogs and some reptiles glide steeply, and are better described as parachuting, truly gliding animals operate at angles far below 45° in still air. How rapidly they descend, the *sinking speed* V_s, often the most important criterion, depends both on the glide angle and the *glide speed* V_g along the flight path. Figure 4 illustrates two different glide strategies. Bird A is a fast glider, with high lift, low drag wings, and hence a low glide angle. Bird B is a slower glider with a higher glide angle. Which has the lower sinking speed depends wholly on the values of Θ and V_g:

$$V_s = V_g \sin \Theta.$$

In still air descent is inevitable, and a descending glide is appropriate for a flying squirrel moving from tree to tree or a bird coming in to land, but prolonged gliding must recruit energy from rising air. This is usually distinguished as *soaring,* and it takes many forms. Air rises when wind is deflected upward by a cliff or the face of a wave. Air passing over a rock or an island can be thrown into standing waves, rising and falling in sequence down wind. A linear updraught forms where air masses converge and meet. Columnar thermals develop above hot ground, and

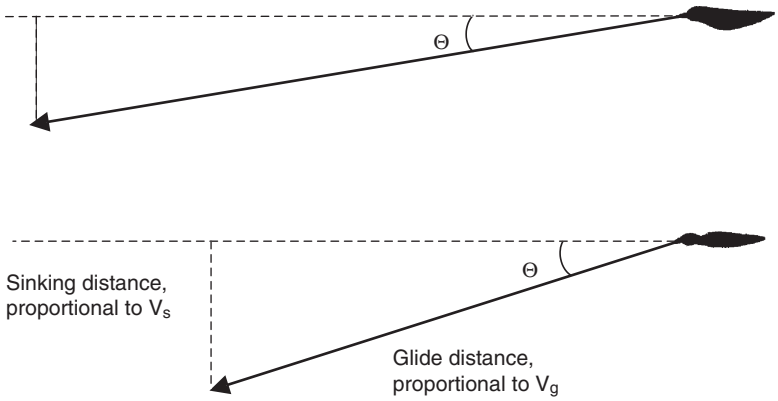


Figure 4: Two gliding strategies: V_g, glide speed; V_s, sinking speed; Θ, glide angle.

over towns. All these are exploited by birds. Many gulls, for example, *slope-soar*, gliding back and forth in the updraughts that occur when the wind meets a cliff. Condors and eagles do so in the mountains. Albatrosses and shearwaters slope-soar in the smaller upcurrents off the faces of waves. A variety of sea and land birds exploit frontal updraughts, and standing waves in the lee of large rocks. Vultures, eagles, buzzards, kites, storks and pelicans soar in thermals and travel with them downwind [34, 35], and frigate birds are able to exploit thermals that develop over the sea in the trade wind zones [36, 37]. The circumstances and lifestyle of the bird influence the gliding strategy, and hence the design; this is explored in Section 4.2.

In comparison with birds, bats make little use of gliding, flying mainly at night in situations where upcurrents are unimportant. Many pterosaurs, in contrast, must have made significant use of gliding and soaring. Several analyses of pterosaur flight have been published, and although these do not always agree in detail it seems certain that the giant Cretaceous forms at least were habitual soarers. The oceanic *Pteranodon* species, with wing spans up to 9 m, have been interpreted both as exclusively soarers with very low stalling speeds [38, 39] and as slow flappers with soaring capabilities [40], and the still more enormous terrestrial *Quetzalcoatlus northropi*, with an estimated span of around 12 m, must have been a specialised soarer; its muscles could surely not have coped with continuous powered flight [41]. By analogy with birds, many smaller pterosaurs must have made extensive use of both gliding and flapping.

Most prolonged gliders are relatively large. Many medium-sized birds, like crows, alternate between periods of flapping and gliding. The drag on smaller, lighter birds soon slows them down, and these glide, if at all, only briefly between bursts of flapping. Swifts and swallows, with their slender, high lift wings, are effective brief gliders, as are some bats, particularly those with high aspect ratio wings [42, 43]. Some large insects – locusts, some dragonflies and butterflies – similarly glide for short distances between periods of flapping, and dragonflies have been observed to soar on thermals [44]. Interestingly, gliding insects usually have moderately broad wings, apparently because on this scale slender wings are liable to stall before reaching their theoretically optimal gliding speed. The gliding seeds of *Alsomitra* and related plants also have moderate aspect ratios, probably for the same reasons [45]. The rule is not without exceptions: the huge pseudostigmatid damselflies glide well, despite their relatively slender, petiolate wings [46].

3.2.2 Powered forward flight

Provided metabolic energy is available, effective flapping allows flight to continue indefinitely, providing extra lift and hence weight-support and thrust to maintain altitude and speed. It also gives improved scope for accelerations and manoeuvres, greatly increasing the range of flight patterns and functions.

However the aerodynamic implications of flapping are considerable. The wing stroke cycle must again generate sufficient net upward force to support the weight. In moderate to fast flight, where the forward velocity of the animal relative to the flapping velocity of the wings is high, this may present no great problem, since the wings will be operating at a favourable angle of attack throughout much of the stroke. In fast flight, long-winged birds like falcons, gulls, swifts and martins, and many bats, generate weight-supporting force throughout the stroke-cycle, merely flexing their wings slightly in the upstroke [47, 48]. Visualisation of the flow behind a kestrel, flying at around 7 m/s, showed it to be leaving a wake that consisted of a pair of undulating vortices trailing behind the wing tips (Fig. 5a) [49]. Both upstroke and downstroke were generating weight support, but the upstroke produced less force – necessary as its horizontal component was

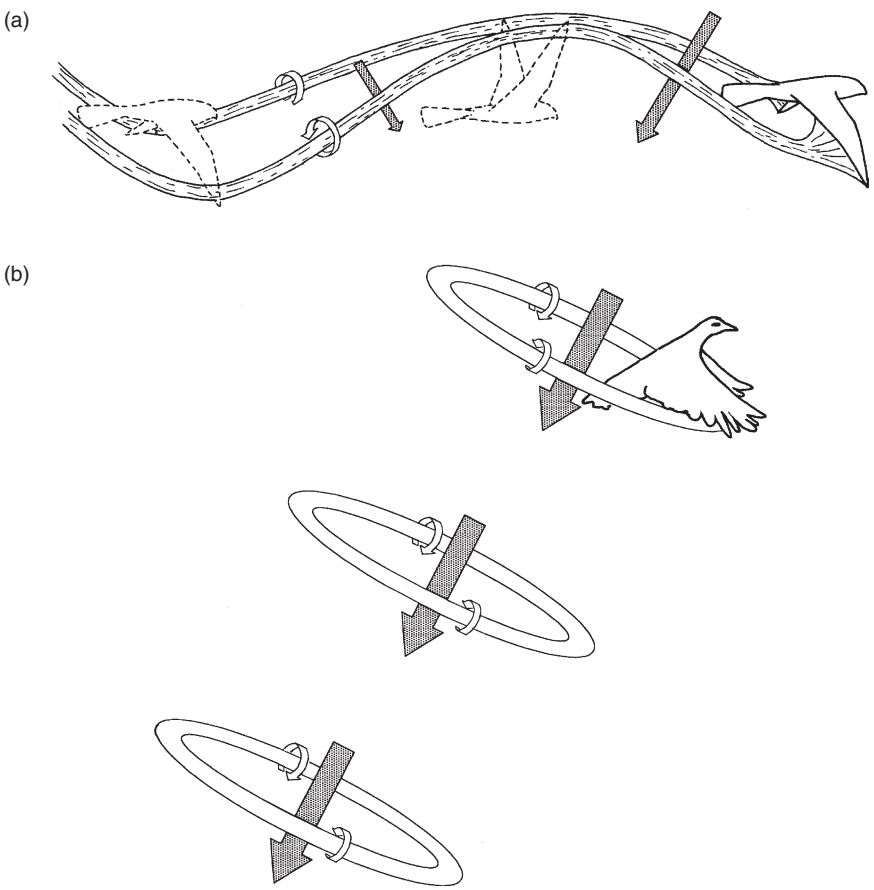


Figure 5: (a) Undulating vortex wake behind a fast-flying falcon. (b) Vortex ring wake behind a rising pigeon. Broad arrows represent the directions of mass airflow.

backward, reducing the overall thrust. Birds and bats with shorter wings, long-winged birds at low flight speeds, and most insects need to introduce far more asymmetry of shape or motion between the two halves of the stroke, so that the upstroke does not generate adverse forces countering the weight support and/or thrust provided by the downstroke. At low flight speeds, many birds and bats develop most of their useful aerodynamic force in a strong downstroke, and fold the wings close to the body in the upstroke, the birds twisting their primary feathers like the vanes of a venetian blind to minimise both lift and drag. Each downstroke generates and sheds a discrete ring vortex (Fig. 5b), so that visualisations of the animal's wake have shown it to consist of a series of such vortices projected downwards and backwards [50, 51].

These two flow patterns have long been regarded as distinct, discontinuous gaits, in the sense of the distinct walking, trotting and galloping gaits of mammals, but new work on thrush nightingales (*Luscinia*) flying freely in a giant wind tunnel, indicates that in this species at least they are actually extremes of a continuum of flow patterns, changing with increasing speed. Furthermore, even at their lowest speeds the nightingales generate a weak ring vortex in the upstroke, as well as the far stronger ring of the downstroke. If this last proves to be generally true, it solves a problem that has worried scientists for three decades: that calculated forces from the single vortex ring model are inadequate to support the bird's weight [52].

Insects do not actively fold the wings in flight, but the latter are twisted and often bent in the upstroke [53–59]. This may serve to 'feather' them, so that the upstroke generates minimal forces, but the wings of most insects are capable of being twisted enough for the upstroke to generate useful force, often including some weight support (Fig 6b and c). Since the wings have no internal muscles, this twisting, often accompanied by transverse bending, seems to be driven mainly by the inertial forces they experience as they accelerate [60]. Some insects are capable of altering the shape of the wings in the upstroke according to their flight speed, but bees, and by implication many other insects, appear to change speed simply by altering the angle of the body and with it the wings' stroke plane, tilting the body progressively more vertically and the stroke plane more horizontally as speed falls (Fig. 7).

Much attention has been paid to investigating how far flapping flight can be explained by conventional aerodynamic theory, and how far unconventional mechanisms are involved. Conventional theory was developed in association with fixed wing flight, and assumes that the airflow past the wings is relatively steady, with the wings maintaining a more or less constant velocity and orientation to the oncoming air. With oscillating wings this is clearly not so. Early investigations of insect flight [61, 62] treated the flapping cycle as a sequence of steady states, calculating the

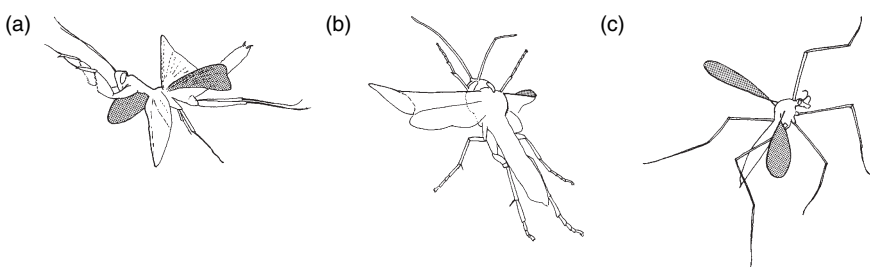


Figure 6: Three insects showing the form of the wings in the upstroke. (a) A mantis. (b) A sawfly, showing wing tip bending and twisting. (c) A cranefly showing complete twisting. Traced from photographs: (a and b) from [63]; (c) from [64].

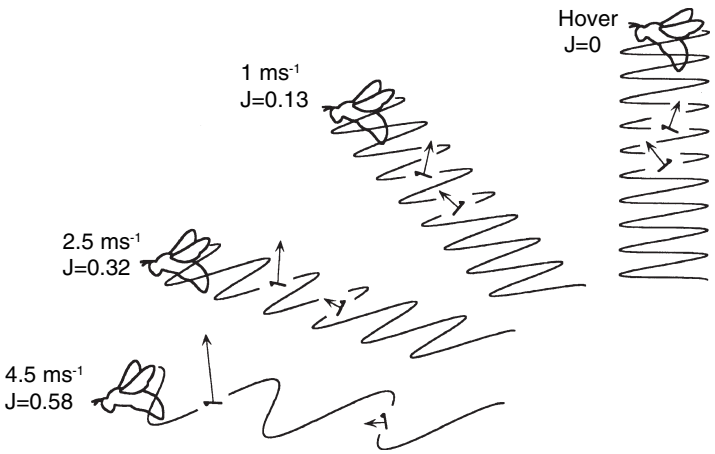


Figure 7: Wing tip paths and wing and body orientations in a bumblebee, *Bombus terrestris*, at various speeds and advance ratios (J). Net aerodynamic forces are shown for representative downstrokes and upstrokes. The underside of the leading edge is marked by a triangle [65].

forces developed at each instant of the cycle, integrating these to estimate the total weight support and thrust, and comparing these with the weight and the total drag. However it has become progressively evident that unconventional mechanisms must often be involved, and several have been identified and analysed [8, 12, 13, 66, 67]. Some of these are transient departures from the steady states of orthodox aerodynamics, operating as the wings decelerate and accelerate around the top and bottom of the stroke. Many tiny insects, and some others make extensive use of the 'clap and fling' mechanism, in which the wings clap together at the top of the stroke and fling apart into the downstroke, generating high levels of lift-creating vorticity as soon as the stroke commences [68–70]. In the fruit-fly, *Drosophila*, high, unsteady lift appears to be created by rotating the wing backwards about its long axis as it begins to decelerate at the end of the downstroke, generating a kind of back spin like that on a sharply cut tennis ball [9, 12, 67, 71]. The timing of these transient events is probably crucial in the fine control of flight manoeuvres in these adept, precise fliers. Other mechanisms operate during the main sweeping (*translational*) part of the stroke. In hawkmoths, and by implication many other insects, extra lift is generated by the development of a low-pressure vortex behind and above the anterior edge that is sustained throughout the half stroke by spiralling out along the span, a steady-state but unconventional effect with some similarity to the flow pattern over the wings of delta-winged aircraft [8, 72]. The form, behaviour and occurrence of the leading edge vortex in insects, and its possible relevance to small birds [73–75] are currently being intensively studied. Two approaches are now usual: theoretical, using computational fluid dynamic analyses [11, 70]; and experimental, using flow visualisation, particularly particle imaging velocimetry. The former uses precise measurements of the shape and movements of the wings to compute the flow, the strength and distribution of the vorticity and consequent aerodynamic forces. The latter computes similar information from video recordings of the movement of suspended particles past flapping [9] or revolving [74, 76] or fixed [73] model wings, or past actual animals in tethered flight [77, 78] or flying freely [52, 75]. Models and tethers are normally instrumented to measure the forces actually generated.

The results from these investigations are at present conflicting, reflecting the diversity of the animals, the range of experimental approaches and the relative novelty of the techniques, which are still being refined. There is still much to learn.

3.2.3 Hovering

Many insects and some small birds and bats are capable of hovering. Many male dragonflies and flies hover in territory surveillance and mate location. Hummingbirds, some phyllostomid bats, hawkmoths (Sphingidae), many bees, bombyliid and syrphid Diptera, and various other moths and flies hover to feed at flowers. Most birds over 20 g, and many weighing less, can hover only briefly, if at all, in still air, though the African pied kingfisher, weighing about 100 g, hovers sustainedly over water in search of prey [79]. Kestrels and some other falcons, terns and some kingfishers can fly at zero ground speed into a wind – usually distinguished as wind hovering.

The majority of hovering insects do so by tilting their bodies towards the vertical, and beating their wings almost horizontally, twisting them extensively so as to generate vertical weight support on both half strokes, with any horizontal forces cancelling out. So do the nectar-feeding glossophagine phyllostomid bats [80, 81]; and so do hummingbirds, their short arms and long primary feathers allowing them to twist the wings to an extent unparalleled in other birds. Hummingbird flight technique appears astonishingly insect-like, and they have often been mistaken for large moths and *vice versa*. The curious back-tilted thorax of dragonflies allows them to hover without tilting the body. Hoverflies (Diptera, family Syrphidae) can do the same; but the latter [82], and most other hovering birds [83] and bats [6, 47, 84, 85] use an oblique stroke-plane, and generate most or all the necessary vertical force on the downstroke. Some butterflies can hover briefly by beating their relatively huge wings vertically downwards [86] – in effect operating them like fans or paddles and using pure drag for weight support – but others use the more conventional horizontal stroking method [87].

3.2.4 Special techniques for flight economy

Flight is metabolically expensive; a 100 g bird in cruising flight uses energy around twice as fast as a walking or running mammal of the same body mass [88]. Birds in particular have adopted a range of flight behaviours that serve to reduce energy consumption. Many species, particularly those of medium and large size, alternate between phases of flapping, rising as they do so, and gliding, in which they lose altitude. This is known as *undulating flight* [89]. Brief alternating bursts of flapping and gliding are also characteristic of the aerial manoeuvres of the slender-winged swifts and swallows, and many butterflies do the same.

The V-formation flying of geese, swans and cranes is again an economy measure. With the exception of the leader, each bird flies just outside the wing-tip vortex of the one in front, gaining support from the upwash that it causes. Theoretically this may save up to 15% of the total aerodynamic power expended [79].

The spectacular fishing technique of skimmers (*Rynchops*) which fly with a shallow wing beat in straight lines very low over the surface of water, snatching fish with an elongate lower bill, exploits the *ground effect*, in which the high pressure in the space between the bird and the water can save as much as a fifth of the necessary mechanical power [90, 91]. Waders and some other water birds, bats [92] and insects [10] often fly close to the water surface, and probably gain similar advantages. So may a variety of satyrid and other butterflies that fly close to the ground [10]. The Jurassic pterosaur *Rhamphorhynchus* is thought to have been a fish feeder, and may well have caught them, skimmer-like, by flying low over the water.

Many small birds with rather broad, short wings – small passerines like finches and tits, and woodpeckers, some parrots, and kingfishers – use *bounding flight*, in which flapping bursts alternate with periods in which the wings are folded close to the body. The flapping period accelerates the bird, and projects it into the folded phase, like a thrown pebble. This too has formerly been considered to be an energy-saving device, but is now seen as a means by which the bird can control its power output over a wide speed range by varying the length of the flapping periods [89, 93]. Bounding flight appears to be limited to birds with a body mass below *c*. 200 g.

4 Designs for flight

4.1 Basic morphology

4.1.1 Vertebrates

The wings of birds, bats and pterosaurs are/were modified fore-limbs, with bones, muscles, tendons, ligaments, and a vascular and nervous supply. The bat wing skeleton has a full complement of greatly elongate fingers, supporting a thin patagium of skin, muscle fibres and elastic ligaments, extending back to the hind limbs, and in most families to the tail. The patagium of the pterosaurs was principally supported by the arm and one extremely long finger, the other digits being free. Interestingly, the extent and structure of the pterosaur patagium are still controversial. Early reconstructions showed it continuing behind the hind wings to the tail, but this seems not to be true, though one interpretation of *Pteranodon* postulates that it may have been attached to extended tail vertebrae, bypassing the legs [94]. In general, though, opinion is divided as to whether the patagium arose from the side of the trunk, leaving the legs free [95, 96], or extended to the hind knee, or to the ankle. This would affect the area of the wings, and hence both wing-loading and aspect ratio (see Section 4.2.2) and has implications in calculating probable flight performance. In many specimens the patagium shows a close-packed series of fine ridges and grooves. These were earlier suggested to be wrinkles, resulting from the contraction in death of elastic filaments in the membrane whose function was to keep it taut, and allow it to function, in a wide range of flight positions. This would make pterosaur flight more bird-like than bat-like, and would have major implications on their capabilities and ecology [97]. However they are now believed to have been stiffening fibres, the so-called actinofibrils [98].

Another point of controversy is the position and function of the pteroid, a slender bony rod arising from the wing carpals. In most undisturbed fossils this points back along the wing towards the body, and evidently supported the propatagium, the relatively small area of the patagium lying anterior to the arm. However it may in life have been directed forward, supporting a far larger propatagium than has usually been assumed, and to have been capable of altering the basal profile of the wing. If this is the case, the wing area would have been significantly greater than has been believed, and wind-tunnel experiments on models have shown that flight versatility and overall performance would have been considerably enhanced [99].

The bird wing skeleton has a much reduced hand, and the main aerofoil surface is provided by the feathers: the primaries, extending to the wing tip and the distal part of the trailing edge, and the secondaries, forming the posterior part of the proximal region of the wing. The bases of these are overlaid by the shorter wing covert feathers, supplying much of the curved upper surface. The vestigial thumb bears a tuft of feathers – the alula – that can be actively lifted in many species and is believed to act like the 'slot' on an aircraft wing, deflecting air down over the wing's upper surface and delaying stalling at low speeds and high angles of attack.

The tail feathers of birds form an additional aerofoil, providing some lift, and particularly stability and active control [100, 101]. The shape and length of bird tails vary enormously, reflecting different modes of flight but also showing a range of special adaptations in association with courtship and other display.

In both birds and bats, the thoracic muscles powering the wing stroke are inserted via tendons on the humerus bone of the upper arm. In both groups the downstroke is powered by the large pectoralis muscles, arising on the sternum, which in birds is broad and keeled – familiar enough to anyone who has ever carved a chicken. The sternum in bats is relatively narrow, and the pectoralis muscles pull against each other, separated by a sheet of tough membrane. The upstroke in vertebrates is powered by smaller muscles, assisted by the aerodynamic force on the wing. In birds this muscle is the supracoracoideus, inserted on the sternum and covered by the pectoralis; with a tendon which passes through a hole between three bones to attach on the upper side of the humerus. In bats the wing elevator is the deltoid muscle attached, more conventionally, to the scapula – the shoulder blade. Pterosaurian flight musculature had features of both patterns. The broad, rather bird-like sternum indicates the presence of a large pectoralis for the downstroke, but the upstroke seems to have been powered by both a supracoracoideus from the sternum and a deltoid from the scapula [95].

We have seen that flapping flight requires the wings to undergo significant, often extreme cyclical changes in shape, during the stroke-cycle, and when the wings are stowed at rest. In all three groups of flying vertebrates, these deformations are/were brought about directly by the muscles of the arm and hand, flexing and extending the arm; and spreading and apposing the feathers in birds, and the fingers in bats and pterosaurs. Not so in the insects.

4.1.2 Insects

The flight system of insects is fundamentally different from those of flying vertebrates. The skeleton of insects is topologically external, and insect wings are skeletal structures, with no internal musculature, and very little other internal tissue. They consist largely of *cuticle*, a highly structured complex of organic polymers whose physical properties – stiffness (resistance to deformation), strength, toughness (resistance to fracture and tearing), resilience (springiness) – vary greatly from place to place according to the cuticle's local chemistry and ultrastructure [102]. Typical wings consist of thin membrane, supported by and continuous with a framework of *veins*, usually hollow tubes containing blood and air-filled respiratory tubes. Most insects have two wing pairs, not one as do vertebrates, and these may beat in or out of phase. In many insects the two wing pairs are coupled together, forming a single effective aerofoil. In several groups one or other pair is lost, or modified to protective structures (the fore wings of beetles and earwigs) or special sense organs (the hind wings of flies).

The muscles powering flight are restricted to the thorax, and in most insects are remote from the wings themselves, although in dragonflies, cockroaches and Orthoptera the downstroke is wholly or partly powered by muscles inserted at the wings' extreme base. Elsewhere these 'direct' muscles are concerned only with modifying the wings' angle of attack and the shape of their base in flight, and in folding them at rest. The aerodynamically useful deformations that the wings undergo in flight are brought about by the inertial and aerodynamic forces that act on them, modulated, often with great precision, by the muscular forces at the base. How these interacting forces influence the shape of the wing through the stroke cycle depends largely on the detailed architecture of the wings themselves: the distribution of stiff and flexible components and areas, whose properties are influenced both by their macrostructure and by the chemistry and ultrastructure of the cuticle (Section 4.2.2).

4.2 Morphological variables

The extensive range of flight functions and techniques that animals adopt is reflected in a wide variety of designs, and discovering the relationships between design, performance capabilities and behaviour is a major challenge to students of flight. Conventional engineering analyses based on orthodox aerodynamic theory have been helpful, but the theoretical framework for analysing flapping flight is still incomplete, though advancing rapidly. Actual measurements and observations are essential; and the results are sometimes surprising and thought-provoking.

Some design features are readily quantifiable, and their association with particular types of flight performance and behaviour can be theoretically predicted and tested against actual observations. A variety of body characters may influence performance, adversely or favourably. A bulky head or long, unwieldy legs will increase drag. Tail size and design are very important in birds, and a recent analysis of microbats has shown that tail and ear lengths and areas, as well as wing variables, may contribute to flight performance and manoeuvrability, and assist in predicting foraging strategies and habitat preferences [103]. Nonetheless most attention has properly been focussed on wings.

4.2.1 Wing loading and aspect ratio

The wing loading of an animal is the weight carried by unit area of wings. Small wings on a heavy animal need to generate more lift per unit area, and this is most readily achieved by moving faster through the air: by more rapid flight, or by flapping at a higher frequency or at higher amplitudes, or any combination of these. Orthodox aerodynamic theory predicts that, for animals of similar design, characteristic speeds such as minimum speed, the speed at which minimum power is expended, and that at which the *cost of transport* (energy consumed in transporting unit weight of the animal through unit distance) is lowest, should vary approximately as the square root of the wing loading (see Section 5.2). Animals with high wing loading values might be expected to fly fast and/or flap rapidly, and to have high take-off and landing speeds. Since wing loadings tend to increase with size (in a range of similarly shaped animals weight will vary as length^3, wing area as length^2, so wing loading will increase directly with linear dimensions), one would expect larger animals to fly faster and to expend more effort in getting into the air, or to have disproportionately large wings.

Clearly, the designs of flying animals vary enormously, but in birds, which include the largest extant flying animals, these predictions are very broadly correct, though at any given mass or wing loading the range of speeds is wide. Many of the highest reliable estimates of sustained flight speed, between 17 and 21 m/s, are recorded for swans, geese and large ducks, whose wing loadings are among the highest, while recorded speeds for small passerines and, unexpectedly, swifts, with the lowest wing loadings, are generally below 12 m/s [79]. A peregrine falcon has been reported to exceed 70 m/s in a dive, but its wings would have been folded, and the wing loading thereby greatly increased. Flight speed data on bats are relatively scarce, but there appears again to be a broad correlation between speeds, wing loading and body mass. High wing loading can reduce manoeuvrability, as the radius of turning is proportional to wing loading, and insectivorous and carnivorous bats all have relatively low values [42, 47].

Few insects have wing loadings approaching those of even the smallest birds and bats, and most are several orders of magnitude lower [104]. Bumblebees, whose loadings are among the highest, are capable of hovering; and the ability, or lack of it, for insects to fly slowly and to hover depends on other factors of design and physiology. Low wing loadings, however, certainly limit top speeds, and although very few reliable values have been published, the majority of insect species – most of which are tiny – appear incapable of exceeding 1 m/s [10]. A significant factor allowing the evolution of small flying insects has been the origin, several times independently, of

an unique type of muscle whose operating frequency is not limited by the need to be excited by a nerve impulse at each contraction. These *asynchronous* muscles are activated by being stretched by the muscles which oppose them, so that the thorax and wings are able to operate at or near their physical resonance frequencies, which may be several hundred cycles per second. Some tiny biting midges have been recorded at over 1 kHz [105], though this needs confirmation. High frequency flapping accelerates more air, and asynchronous muscle allows insects to function with relatively smaller wings and higher wing loadings, and hence perhaps at higher flight speeds for their size [10]. However even the fastest insects seem to operate at the lower end of the speed range of flying vertebrates. A queen bumblebee of the species *Bombus terrestris* has been recorded at 7.6 m/s in a wind tunnel [106], and foraging bumblebees can fly still faster in the field [107]. Some dragonflies may fly at higher speeds, though the evidence is largely anecdotal.

The aspect ratio (AR) is a measure of the slenderness of wings. It is properly calculated as $4(\text{wing length})^2/\text{wing area}$, though often as $(\text{wingspan})^2/\text{area}$, including the width of the body; the two values may be rather different.

Aspect ratio, at least in aircraft, birds and bats, can be thought of as an approximate measure of aerodynamic efficiency. High aspect ratio wings have lower induced drag (the drag associated with lift generation, (see Section 5.1)), so that the lift/drag ratio is high, and energy costs relatively low. Virtually all man-made gliding aircraft have high aspect ratios, and so do marine gliding birds – albatrosses (AR 15–20), gannets (AR *c*. 13), fulmars (AR *c*. 11), operating in open spaces, often close to the water in situations where upcurrents are chancy and relatively slight and shallow glides are essential. So too do frigate birds (AR *c*. 15), which are also capable of thermal soaring in the trade wind zones [36, 37]. In these, and also in swifts, terns, martins and many migratory bats high aspect ratios allow long periods of sustained, economical flight. Long wings also generate large turning moments, and can hence aid agility, the ability to perform rapid banked turns on the wing. Birds and bats that catch insects on the wing in open spaces tend to have high aspect ratios. Most insects change direction without noticeable banking, but long-winged butterflies like the heliconiid *Dryas julia* and some swallowtails are relevant exceptions.

Low aspect ratio wings, though less efficient aerodynamically, are by no means uncommon. Considerable structural and mechanical strength is needed to support and flap long wings, and breadth is often necessary to limit wing loading, and hence to allow slower flying and higher manoeuvrability. Most terrestrial soaring birds have a low aspect ratio, a lower lift/drag ratio, and a correspondingly steep glide angle. Steeper glide paths matter less to birds exploiting strong upcurrents, and broad wings can have a large area and hence low wing loading without excessive mechanical stress. Slower flight aids the tight turning circles that may be necessary to stay in narrow thermals. Fast gliding is still possible when necessary, by partly flexing the wings, so that the bird has considerable control over speed, and Rüppell's griffon vultures (*Gyps rüppellii*), for example, can glide huge distances by soaring slowly in thermals and gliding fast between them [34, 35].

Long wings can also hinder flight within small spaces, and many woodland birds and insectivorous bats have short, broad wings, relatively low wing loadings, and the capacity for versatile, manoeuvrable flight in dense vegetation [42, 43, 48].

The influence of aspect ratio in insects is less clear-cut, because of the presence in most groups of two pairs of wings. Where fore and hind wings are physically coupled, or at least beat in the same phase, it is reasonable to calculate a single value for the aspect ratio of the four wings together, but this is less logical where fore and hind wings, often quite different in size and shape, beat out of phase.

For a given body size, long wings sweep out a larger area through which the air is accelerated. This should particularly favour slow flight and hovering, and indeed many slow-flying insects,

like zygopterous Odonata (damselflies), myrmeleonid Neuroptera (ant-lions) and nematoceran Diptera (midges and gnats), have high aspect ratios; but this is by no means invariable, and most skilled hoverers have medium aspect ratio wings. Recent work with model wings suggests that aspect ratio may not be particularly influential in determining the lift/drag ratio in the insect size range [76].

The relationship between flight performance and design is complex, with many interacting variables. Statistical analysis of such situations needs a multivariate technique; and for flying vertebrates principal component analysis (PCA) has proved particularly useful. PCA allows the relative overall influence of individual variables to be computed. Analysis of a wide range of birds [48] and of bats [42] shows size to be overwhelmingly important in accounting for overall variation, followed by wing loading and aspect ratio. It is then possible to examine the relationship between loading and aspect *corrected for size* in a two dimensional scatter plot of these variables for the individual species, and to interpret the results in terms of their flight capabilities and behaviour. The results are illuminating: the plots conveniently and encouragingly tend to bring together many species, often in different families, with similar flight habits.

A combination of high aspect ratio and relatively low loading minimises the power requirements and cost of transport (Section 5.1) and favours animals that fly continuously for long periods (albatrosses, swifts, terns, gulls) and long distance migrants, including some bats. In helping slow, economical flight it also assists aerial predatory birds (swifts, swallows and martins, bee-eaters, falcons, harriers, kites, skuas).

High aspect ratio with high loading still minimises cost of transport, but allows higher speeds. This favours animals that commute long distances, like the molossid Mexican free-tailed bat *Tadarida brasiliensis* which flies many kilometres each night to feed on distant moth swarms. Birds with this combination tend to be fast-flying migrants, like eiders and other arctic ducks, or shore birds, flying rapidly between feeding grounds; or wing-swimmers, like auks and diving petrels, whose wings are a compromise between hydrofoils and aerofoils [48].

Low aspect ratio with low loading would tend towards slow flight, which is potentially costly. This combination characterises a large number of woodland birds and bats operating in situations where long wings would hinder movement, and manoeuvrability is useful. Here too are the terrestrial soaring birds, where energy expenditure is minimised by gliding, and the ability to fly slowly and turn tightly is often valuable (Section 3.2.1).

Low aspect ratio with high loading tends to fast expensive flight. Birds with this combination are generally poor fliers – ground-living galliforms (grouse, pheasants, partridges), tinamous and rails, that fly mainly for escape, where speed is more important than economy. No bats live like this; bats with relatively low aspect ratio and high loadings tend to be insectivores, or fruit- and nectar-feeders, operating in close habitats, capable of fast but agile flight and also hovering. Hummingbirds too combine hovering and nectivory with fast flight between plants, but have higher aspect ratios than their bat equivalents.

A similar PCA of Pterosauria, using estimated values for mass, places all species studied in the high aspect ratio, low loading category, occupied in birds by the marine gliders and aerial predators [108]. This conforms well enough with both morphological and geological evidence. The fact that *Quetzalcoatlus* and some other late Cretaceous pterosaurs are found in continental deposits merely confirms that a high aspect ratio is excellent for terrestrial as well as marine soaring, no surprise to glider-pilots. There is no reason to suppose that all pterosaurs were soarers; active, falcon-like flight is entirely probable for smaller species.

Analyses of this kind are less easily applied to insects. The ambiguity of aspect ratio as a concept in four-winged animals, their frequent reliance on unsteady aerodynamics, the uncertain influence of very small size and low wing loadings, their far more varied wing shape, all introduce problems

of practical analysis and interpretation. In a PCA of butterflies it was far harder to distinguish the influence of wing loading and aspect ratio from that of other morphological factors, and the relationships between design, flight habit and habitat were far less clearly defined than in comparable studies on vertebrates [87].

4.2.2 Wing shape and proportions

One consequence of flapping, as opposed to flying with fixed wings, is the need to expend energy in accelerating and decelerating the wings against their inertia at the end and beginning of each half stroke. This *inertial power* (energy consumed in this way in unit time) is a function not only of the mass of the wing but also of the *moment of inertia* of the wing about its base, a measure of the distribution of mass along its length. The effect of the mass of any component of the wing increases as the square of its distance from the base, so for minimum energy expenditure it is logical to concentrate the mass of the wing close to the base, keeping the distal part as light as possible. This is evident in all animal wing designs: in the reduction of the bony component of bird wings, with the distal area supported by feathers made of the significantly lighter keratin; in the ultra-slender digits of bats, and their reduction to one supporting finger in pterosaurs; and in the tapering, thin-walled distal veins of many insect wings. It also partly accounts for some major differences in wing design within these groups. Figure 8 compares the wing proportions of three birds with high aspect ratios. The bones of the albatross (Fig. 8a) extend *c*. 65% the length of the wing, those of the hummingbird (Fig. 8b) *c*. 35%, and of a swift (Fig. 8c) a mere 28%. Albatrosses glide extensively and flap slowly, and the long bony skeleton is necessary to support the wing against the high bending moments associated with large size. Swifts alternate gliding with periods of rapid flapping. The short arm and long primary feathers are stiff enough to cope

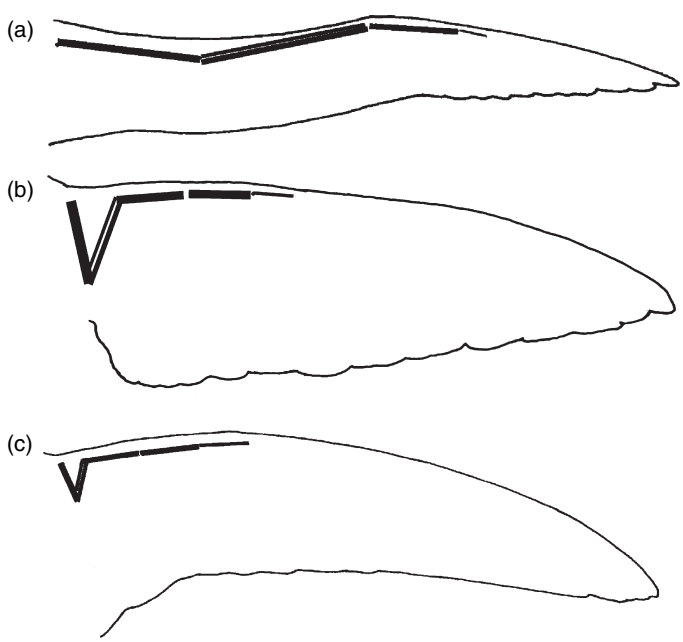


Figure 8: Three narrow bird wings, showing the extent of the bony skeleton: (a) an albatross, (b) a hummingbird and (c) a swift.

with the bending moments, and the moment of inertia is relatively far lower. Hummingbirds flap continuously at high frequencies, up to 52 cycles/s in the case of the ruby-throated hummingbird [79]. Besides ensuring a low moment of inertia, the long primaries assist in the extensive wing twisting that characterises their insect-like flight technique. Hummingbird hands are as long as their arms; and a high ratio of hand length to arm length and consequent low moment of inertia also seem advantageous for hovering bats [42, 43] though the moments of bat wings average twice those of comparable birds [109]. Minimising the moment of inertia, and hence the inertial power, may also be an incidental advantage of the relatively short, broad wings of many woodland birds and bats, and of the short, pointed, broad-based wings of many shorebirds.

A second feature of flapping flight is the presence of a velocity gradient along the span of the wing: the tip is always moving faster than the base. For a given flapping frequency the importance of this gradient increases with decreasing flight speed, and this relationship is conveniently expressed by the *advance ratio*, the ratio of forward velocity to flapping velocity. Many insects that habitually fly at very slow speeds, with a low advance ratio, have narrow-based, even stalked wings whose area is concentrated distally, in the part of the wing where the airflow is fastest. This is true of zygopterous Odonata (damselflies), some Neuroptera and nematoceran Diptera (midges and gnats), and reaches extremes in some fossil insects of the Carboniferous and Permian periods [110]. In contrast, the individual wings, or coupled fore and hind wings, of the majority of insects capable of flying fast have broad bases.

A third important consequence of flapping flight, already touched on in Section 3.2.2, is the frequent need for the wing to deform in flight. Much of this deformation is concentrated in the distal part of the wing: the spreading, closing and twisting of primary feathers and the hands of bats; torsion and bending in insect wings. The length and area of the distal, deformable part of the wing relative to the whole can significantly influence the flight capabilities of the animal, particularly its manoeuvrability and versatility.

4.2.3 Structural detail

The wings of each of the three flying vertebrate classes are rather uniform in basic structure, and their variations are essentially those of proportion. An exception is provided by the wingtip primary feathers of birds. Primaries vary slightly in number, from 9 to 12 but typically 10, but they differ more obviously in the extent to which they are emarginated, and spread in flight. Many seabirds, shore birds, falcons and aerial insectivores have pointed wings, but the majority of birds are capable of spreading their primaries at need, and in many owls and soaring birds they are habitually widely spaced in both the horizontal and the vertical plane, like an array of small individual wings. In larger birds the vanes are usually emarginated, increasing the gaps between them. The aerodynamics of spread primaries are far from straightforward, but they appear to reduce the drag by acting as a series of individual slender wings, and to delay stalling at low speeds [111]. Spread primaries are typical of low aspect ratio wings, but are not limited to them. Some storks and pelicans have aspect ratios in excess of 10 – much the same as gulls – but their tip feathers are widely spread in flight.

Insect wings are bewilderingly varied: in shape, venation density and pattern, in the relative size of fore and hind wings, in the presence and arrangement of lines of flexibility, and in many other features of detail [112]. The functional significance of these variations is progressively becoming understood, and one can recognise some design features occurring widely and convergently in association with particular flight techniques and skills. The broad, fan-like hind wings of Orthoptera and Dictyoptera (grasshoppers and crickets, cockroaches and mantises) for example, are associated with rapid forward flight, with limited versatility; most of the aerodynamic force being generated on the downstroke. Wings of this kind are incapable of significant twisting in

flight [113] (Fig. 6a). Many insect wings, though, have adaptations that allow twisting, at least of the distal areas. Twistier wings can create useful thrust and even some weight-support on the upstroke, and this facilitates flight at low speeds, and in many cases extends the insects' speed range and overall versatility. Torsion is often accompanied by some ventral bending, and the wings of many insects, like the sawfly in Fig. 6b, have a transverse, or obliquely transverse line of flexible cuticle, like a one-way hinge, that allows the tip to bend and twist in the upstroke, even when the wing base itself is incapable of much torsion. The most versatile fliers of all – dragonflies, many flies, some lacewings – have astonishingly twisty wings, with a soft posterior margin that allows the entire wing to rotate about its long axis, in effect turning upside down for the upstroke (Fig. 6c).

These deformations are facilitated and limited by the wings' architecture: the distribution of flexible and rigid components; and they are also influenced by the muscles at the extreme base – a combination of automation and remote control which is without parallel elsewhere in the animal kingdom [114]. Not all wing characters, however, are flight-related. Insect wings have come to assume many secondary functions: camouflage, signalling, armour, thermoregulation; and these may influence or even dominate their adaptive design.

5 The energetics of flight: power, speed, size and behavioural ecology

Powered flight is an expensive activity [115]. The potential for high speed means that the cost of transporting unit mass of animal over unit distance is generally lower than that of terrestrial locomotion, but the metabolic power requirements (energy consumed in unit time) are extremely high. Cruising flight in a 0.1 kg bird uses between 7 and 10 W – compare about 4 W for a walking/running mammal of the same size [88]. Bat flight may be slightly more expensive, and insect flight is more costly still: bumblebees in flight consume up to 400 W/kg body mass over a wide speed range [116].

5.1 Power, and the power curve

The *metabolic power input* is the rate at which fuel is consumed. The *mechanical power output* is the power delivered by the muscles in flight. The ratio of the latter to the former is the *efficiency* of the system, taking account of energy lost as heat in muscle contraction, and other flight-related metabolic processes.

The power output can be further divided into the *inertial power*, used in overcoming the wings' inertia in flapping, and the *aerodynamic power*, used in overcoming drag. In this context we may distinguish three kinds of drag. These are *induced drag*, the drag incurred in generating the lift-producing vorticity; the *profile drag*, the drag on the wings; and the *parasite drag*, the drag on the body and legs. The aerodynamic power spent in overcoming these three drag components can similarly be divided respectively into *induced power* (P_{ind}), *profile power* (P_{pro}) and *parasite power* (P_{par}). Figure 9 shows the relationships between these power components, in the form of a flow diagram.

The inertial power component should vary with the square of the maximum flapping velocity, with the frequency, and with the moment of inertia both of the wing and of the *virtual mass* – the mass of a volume of air around the wing that is accelerated with it. In practice the cost of overcoming inertia is lower than might be expected. In insects most or all of the inertial energy is thought to be stored elastically in the cuticle or the flight muscle and given back in the next

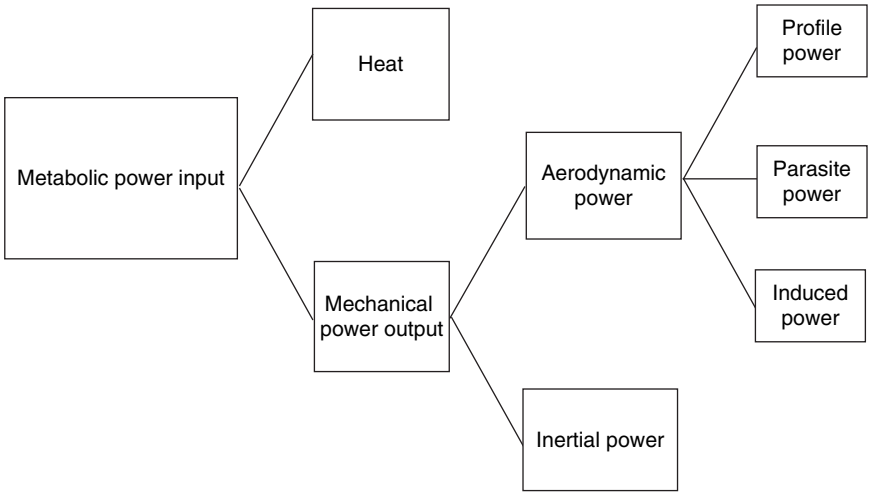


Figure 9: Flow chart of energy expenditure in flight.

half stroke, and in vertebrates too a significant proportion seems to be converted into useful aerodynamic work [47, 109].

The three components of aerodynamic power can be combined in the *power curve,* which has been very influential in the study of animal flight, particularly that of vertebrates, as it is a useful tool for comparing flight patterns and behavioural strategies. The curve and its components are extremely difficult to determine experimentally for any species, and are normally obtained by modelling. The modelling process is complex, however, as many factors are involved, and the literature is controversial [117–123]. In gliding flight the three types of aerodynamic power are expected to vary in different ways with flight speed. Induced power is highest at low speeds, and falls off thereafter. Parasite power and profile power probably increase as the third power of the animal's forward velocity. Flapping horizontal flight is expected to increase both the induced and the profile power components. Induced power will then be highest in hovering and at low speeds, and will rise again somewhat at higher speeds, and some profile power will be expended even when the forward speed is zero. In insects, at least, the flapping velocity may sometimes be more important than the forward velocity in determining profile power [10].

Figure 10 shows these general relationships, and illustrates how the total aerodynamic power P_{sum} should follow a U-shaped curve. The graphs shown have no values attached, and the curves will vary in detail between species according to their morphology and physiology and also their ecology. Information on these can be used to predict the flight capabilities at least of birds and bats, and the interacting effects of a wide range of design features on the various components of the power curve and on other aspects of performance have been extensively analysed and discussed in the literature [42, 118–120, 122, 123].

The power curve has been particularly useful in indicating certain characteristic speeds that may have behavioural significance (Fig. 10). That at the lowest value of P_{sum} is clearly the *minimum power speed* (V_{mp}), at which the animal can fly most economically. That at the point where the curve meets a tangent drawn from the origin is the *maximum range speed* (V_{mr}), where the power/speed ratio and the cost of transport are lowest, and the animal can maximise its flight distance per unit energy expended. Many useful predictions about the effect of design features on performance emerge, and these can be compared with the actual flight characteristics

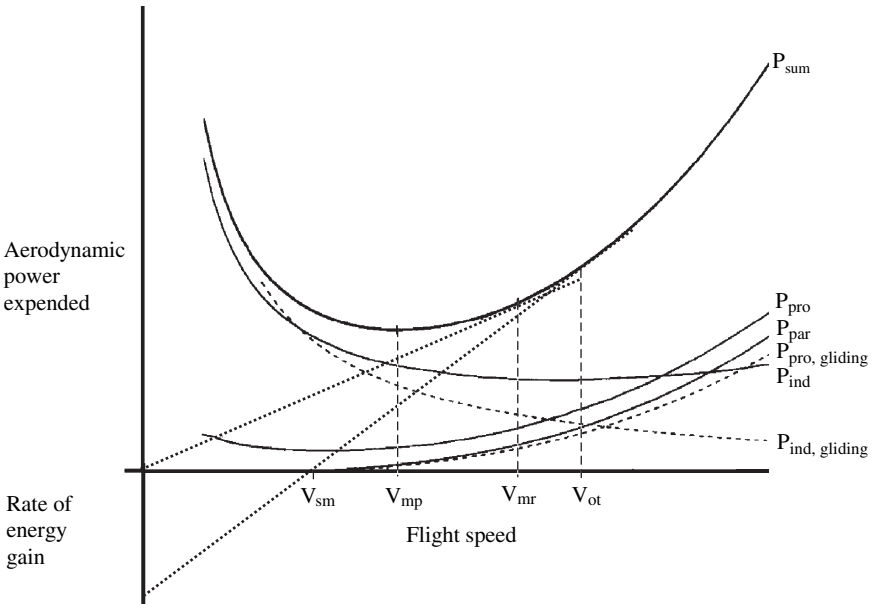


Figure 10: The power curve: how the aerodynamic power requirements for flight should vary with speed. P_{sum}, total aerodynamic power; P_{pro}, profile power; P_{par}, parasite power; P_{ind}, induced power; V_{sm}, predicted optimum speed between stopping places for migrating birds; V_{mp}, minimum power speed; V_{mr}, maximum range speed; V_{ot}, optimum transportation speed. Modified and generalised after [122] and [129].

of particular species. The high aspect ratio wings of swifts, for example, can be expected to minimise both induced and profile drag values, and favour economical flight at all speeds; those of hummingbirds, by reducing induced power expenditure, assist in sustained hovering – at zero airspeed, where P_{ind} will be highest.

It is crucial to realise, though, that these are curves of power *required* for flight at particular speeds. Equally important to the animal are the mechanical power *available* from the muscles, and the rate of energy consumed by them: the metabolic power. This last has been determined for a number of species, by measuring the mass lost in flight, or the oxygen consumed, or the rate of decline of radioactively labelled water. For some bird and bat species the plots of metabolic power against flight speed indeed appear U-shaped but others tend to be rather flat [122, 123]. Most of these results do not include data for low speeds, where P_{ind} would be highest, so the flat trajectories may represent the broad shallow part of the U-shaped curve around V_{mp}, but for some hummingbirds [124], small bats [125] and insects the curve is better regarded as J-shaped, where the power requirements of hovering, low and moderate speeds are similar, and costs only rise at high speeds [121]. Aerodynamic power output curves have been calculated for relatively few insects: bumblebees [116], two dragonflies [126], two moths [127, 128]; and these too appear best to fit a J-shaped model [10].

5.2 Speed and size

If indeed the curve of the mechanical power required for larger animals to fly is U-shaped, we may use it to revisit the relationship between size and available flight speeds. Limiting our discussion

to birds and bats: in general, the larger the animal, the higher will be its power requirements, and the faster it will tend to fly – the power curve is shifted upwards and to the right.

The power *available from the muscles*, however, depends on the morphology and physiological state of the animal, and will be largely independent of speed, though there is some evidence from starlings that the overall efficiency (mechanical power output/metabolic power input) may increase somewhat with speed [122]. Again, the larger the animal, the more power will be available; but the available power rises far more slowly than the power required, and larger birds will be progressively incapable of flight at the lower, more costly part of the speed range [117, 118]. This is very evident when one considers birds taking off. Small birds can rise directly from their perches, but many large birds need to run, or to taxi on the water surface to gain flight speed. The upper size of flying animals seems to be imposed partly by available power, but probably also by the mechanical strength of bones, tendons and muscles. The largest today – swans, bustards, albatrosses, pelicans, the Californian Condor – have a mass of around 15 kg. However fossils show the actual limit to be far higher: estimates put the mass of the Pleistocene condor *Teratornis* at around 40 kg, and the enormous Cretaceous pterosaur *Quetzalcoatlus northropi* and the Miocene condor *Argentornis magnificans* may have been twice as heavy [41, 48].

The huge diversity of insects in size, morphology and flight kinematics makes performance predictions from power curves far harder than in vertebrates. Almost all are smaller than the smallest vertebrates, and some of the largest – hawkmoths, dragonflies – are capable of sustained hovering and flying at a wide range of speeds, while others, of widely different sizes – grasshoppers, many Hemiptera, halticine beetles – need to jump to gain airspeed, and most of these are incapable of hovering. It is indeed not yet wholly clear whether the limits to flight performance in insects are determined by available metabolic power, or by aerodynamic constraints [10, 121].

5.3 Flight strategies, and appropriate speeds

Animals fly for a variety of purposes (see Section 3.1) and the power curve allows us to make some more predictions about the most appropriate speeds [129] (Fig. 10). The minimum power speed V_{mp}, allowing the animal to stay in the air for the longest time without refuelling, seems appropriate for swifts flying at night without feeding; for harriers, kestrels, ospreys and hunting wasps quartering the area in search of a single prey, or frigate birds waiting about to steal food from other seabirds [129, 130] for female insects searching for oviposition sites [10]; for nectar-feeding bats waiting for access to flowers [43]. Other kinds of foraging flight, where many prey items are taken, would be better suited by flying at a speed where the net energy gain is highest – probably rather above V_{mp}, but limited by the animal's ability to recognise prey at speed. This would apply among others to martins and swifts, and to bats [43] feeding on the wing. Speed adopted searching for prey with a patchy distribution would vary with the distance between patches, but should approximate to the maximum range speed V_{mr}. Flight at the food source would be slow, close to the minimum possible. Birds, bees and molossid bats foraging over long distances to bring back food or build up milk for their brood theoretically need to fly faster than V_{mr} [129, 131], at a speed – the optimum transportation speed (V_{ot}) , found by drawing a tangent to the power curve from a point on an extended ordinate representing the rate of net energy gain in foraging. Optimal speed between stopping points (V_{sm}) for a migrating bird is found where this same tangent crosses the abscissa (Fig. 10) [129]. For escaping a predator or pursuing prey the best speed is sensibly the highest available, but for vertebrates might be the highest anaerobic (sprint) or aerobic (stamina) speed, depending on circumstances. Insects are not known to use anaerobic metabolism in flight.

Flight speed is notoriously difficult to measure in the field, but there is a useful body of data tending to confirm some at least of these predictions. Many bird species searching for food have been observed at speeds between V_{mp} and V_{mr}. Kestrels, *Falco tinnunculus*, tend to wind-hover at airspeeds close to the predicted V_{mp} [132]. Swallows travelling between food patches operate well above their feeding speed and above V_{mr} [133]. Pipistrelle bats *Pipistrellus pipistrellus* range around their calculated V_{mr} when foraging, but are significantly faster when commuting [134]. Bumblebees operate well above their calculated maximum range speed when commuting to food patches [106, 107].

6 Conclusions

Flight is a frequent method of locomotion for the great majority of animal species. Though their designs are ultimately constrained by the laws of aerodynamics, the astonishing diversity of flying forms reflects the huge range of lifestyles that flight allows. Habitat and distribution, food preferences and feeding behaviour, generation time, breeding situation and habits, whether sedentary, migrating or commuting, and many other ecological and behavioural factors interact with each other, and indeed with phylogeny: the morphological and physiological heritage of each separate evolutionary line on which later diversification has built. Seeking general principles and common solutions in a field with so many variables is a fascinating but daunting task, and so indeed is accounting for the variability; particularly in insects – why are some groups so variable despite having apparently similar flight capabilities, while others with recognisably different flight behaviours are comparatively uniform? Theoretical and experimental aerodynamics, muscular, circulatory, respiratory and neurosensory physiology, functional morphology and structural mechanics, ecological modelling, all have parts to play, but there is still a major need for field observation and measurement of flight behaviour and performance of the greatest variety of animals under natural conditions. Whatever our models may predict – the animals still fly.

References

[1] Simmons, N.B., Bat relationships and the origin of flight. *Ecology, Evolution and Behaviour of Bats. Symp. Zool. Soc. Lond. 67*, eds. P.A. Racey & S.M. Swift, Clarendon Press: London, pp. 27–43, 1995.

[2] Pettigrew, J.D., Flying primates: crashed, or crashed through? *Ecology, Evolution and Behaviour of Bats. Symp. Zool. Soc. Lond. 67*, eds. P.A. Racey & S.M. Swift, Clarendon Press: London, pp. 3–26, 1995.

[3] Rayner, J.M.V., Flight adaptations in vertebrates. *Vertebrate Locomotion. Symp. Zool. Soc. Lond. 48*, ed. M.H. Day, Academic Press: New York, pp. 137–172, 1981.

[4] Vogel, S., *Life in Moving Fluids. The Physical Biology of Flow*, Princeton University Press: Princeton, NJ, 1994.

[5] Azuma, A., *The Biokinetics of Flying and Swimming*, Springer-Verlag: Berlin, New York, Tokyo, 1992.

[6] Norberg, U.M., *Vertebrate Flight*, Springer-Verlag: Berlin and New York, 1990.

[7] Spedding, G.R., The aerodynamics of flight. *Mechanics of Animal Locomotion*, ed. R.McN. Alexander, Springer-Verlag: Berlin and New York, pp. 51–111, 1992.

[8] Ellington, C.P., Unsteady aerodynamics of insect flight. *Biological Fluid Dynamics*, eds. C.P. Ellington & T.J. Pedley, Company of Biologists: Cambridge, UK, pp. 109–129, 1995.

[9] Dickinson, M.H., Lehmann, F.-O. & Sane, S.P., Wing rotation and the aerodynamic basis of insect flight. *Science*, **284**, pp. 1954–1960, 1999.
[10] Dudley, R., *The Biomechanics of Insect Flight.* Princeton University Press: Princeton, NJ, 2000.
[11] Ramamurti, R. & Sandberg, U.C., A three-dimensional computational study of the aerodynamic mechanisms of insect flight. *J. Exp. Biol.*, **205**, pp. 1507–1518, 2002.
[12] Dickinson, M.H., Solving the mystery of insect flight. Insects use a combination of aerodynamic effects to remain aloft. *Sci. Amer.*, **284(6)**, pp. 48–57, 2001.
[13] Wang, Z.J., Dissecting insect flight. *Ann. Rev. Fluid Mech.*, **37**, pp. 183–210, 2005.
[14] Maynard Smith, J., The importance of the nervous system in the evolution of animal flight. *Evolution*, **6**, pp. 127–129, 1952.
[15] Taylor, G.K., Mechanics and aerodynamics of insect flight control. *Biol. Revs. Camb. Phil Soc.*, **76**, pp. 449–471, 2001.
[16] Taylor, G.K. & Thomas, A.L.R., Animal flight dynamics I. Stability in gliding flight. *J. Theor. Biol.*, **212**, pp. 399–424, 2001.
[17] Taylor, G.K. & Thomas, A.L.R., Animal flight dynamics II. Longitudinal stability in flapping flight. *J. Theor. Biol.*, **214**, pp. 351–370, 2002.
[18] Dudley, R., Mechanisms and implications of insect flight maneuvrability. *Integr. Comp. Biol.*, **42**, pp. 135–140, 2002.
[19] Srygley, R.B. & Chai, P., Flight morphology of neotropical butterflies: palatability and the distribution of mass to the thorax and abdomen. *Oecologia*, **84**, pp. 491–499, 1990.
[20] Norberg, U.M., Evolution of vertebrate flight: an aerodynamic model for the transition from gliding to active flight. *Am. Nat*., **126**, pp. 303–327, 1985.
[21] Feduccia, A., *The Origin and Evolution of Birds*, 2nd edn, Yale University Press: New Haven & London, 1999.
[22] Ostrom, J., Bird flight: how did it begin? *Am. Sci*., **67**, pp. 46–56, 1979.
[23] Padian, K. & Chiappe, L.M., The origin and early evolution of birds. *Biol. Revs. Camb. Phil. Soc.*, **73**, pp. 1–42, 1998.
[24] Burgers, P. & Chiappe, L.M., The wings of *Archaeopteryx* as a primary thrust generator. *Nature*, **399**, pp. 60–62, 1999.
[25] Ji, Q., Currie, P.J., Norell, M.A. & Ji, S.-A., Two feathered dinosaurs from northeastern China. *Nature*, **393**, pp. 753–761, 2001.
[26] Xu, X., Zhou, Z., Wang, X., Kuang, X., Zhang, F. & Du, X., Four-winged dinosaurs from China. *Nature*, **421**, p. 335, 2003.
[27] Bennett, S.C., The arboreal theory of the origin of pterosaurs flight. *Hist. Biol*, **12**, pp. 265–290, 1997.
[28] Gans, C., Darevski, I.S. & Tatarinov., L.P., *Sharovipteryx*, a reptilian glider? *Paleobiology*, **13**, pp. 415–426, 1987.
[29] Wootton, R.J. & Ellington, C.P., Biomechanics and the origin of insect flight. *Biomechanics in Evolution*, eds. J.M.V. Rayner & R.J. Wootton, Cambridge University Press: Cambridge, UK, pp. 99–112, 1991.
[30] Ellington, C.P., Aerodynamics and the origin of insect flight. *Adv. Ins. Physiol.*, **23**, pp. 171–210, 1991.
[31] Kingsolver, J.G. & Koehl, M.A.R., Selective factors in the evolution of insect wings. *Ann. Rev. Ent.*, **39**, pp. 425–451, 1994.
[32] Marden, J.H. & Kramer, M.G., Surface-skimming stoneflies: a possible intermediate stage in insect flight evolution. *Science*, **266**, pp. 427–430, 1994.

[33] Wootton, R.J., How insect wings evolved. *Insect Movement: Mechanisms and Consequences*, eds. I. Woiwod & D.R. Reynolds, CABI Publishing: Oxford and New York, 2001.
[34] Pennycuick, C.J., The soaring flight of vultures. *Sci. Amer.*, **229**, pp. 102–109, 1973.
[35] Pennycuick, C.J., Soaring behaviour and performance of some East African birds, observed from a motor glider. *Ibis*, **114**, pp. 178–218, 1972.
[36] Pennycuick, C.J., Thermal soaring compared in three dissimilar tropical bird species, *Fregata magnificens*, *Pelecanus occidentalis* and *Coragyps atratus*. *J. Exp. Biol.*, **102**, pp. 307–325, 1983.
[37] Weimerskirch, H., Chastel, O., Barbraud, C. & Tostain, O., Flight performance. Frigatebirds ride high on thermals. *Nature*, **421**, pp. 333–334, 2003.
[38] Bramwell, C.D. & Whitfield, G.R., Biomechanics of *Pteranodon. Phil. Trans. R. Soc. Lond. B.*, **267**, pp. 503–592, 1974.
[39] Brower, J.C., The aerodynamics of *Pteranodon* and *Nyctosaurus*, two large pterosaurs from the Upper Cretaceous of Kansas. *J. Vert. Palaeont.*, **3**, pp. 84–124, 1983.
[40] Stein, R.S., Dynamic analysis of *Pteranodon ingens*: a reptilian adaptation to flight. *J. Paleont.*, **49**, pp. 534–548, 1975.
[41] Langston, Jr., W., Pterosaurs. *Sci. Amer.*, **244**, pp. 122–136, 1981.
[42] Norberg, U.M. & Rayner, J.M.V., Ecological morphology and flight in bats (Mammalia; Chiroptera): wing adaptations, flight performance, foraging strategy and echolocation. *Phil. Trans. R. Soc. Lond. B.*, **316**, pp. 335–427, 1987.
[43] Norberg, U.M., Wing form and flight mode in bats. *Recent Advances in the Study of Bats*, eds. M.B. Fenton, P.A. Racey & J.M.V. Rayner, Cambridge University Press: Cambridge, UK, pp. 43–56, 1987.
[44] Hankin, E.H., The soaring flight of dragonflies. *Proc. Camb. Phil. Soc.*, **20**, pp. 460–465, 1921.
[45] Ennos, A.R., The effect of size on the optimal shape of gliding insects and seeds. *J. Zool. Lond.*, **219**, pp. 61–69, 1989.
[46] Rüppell, G., Kinematic analysis of asymmetric manoeuvres of Odonata. *J. Exp. Biol.*, **144**, pp. 13–42, 1989.
[47] Rayner, J.M.V., The mechanics of flapping flight in bats. *Recent Advances in the Study of Bats*, eds. M.B. Fenton, P.A. Racey & J.M.V. Rayner, Cambridge University Press: Cambridge, UK, pp. 23–42, 1987.
[48] Rayner, J.M.V., Form and function in avian flight. *Curr. Orn.*, **5**, pp. 1–77, 1988.
[49] Spedding, G.R., The wake of a kestrel (*Falco tinnunculus*) in flapping flight. *J. Exp. Biol.*, **126**, pp. 59–78, 1987.
[50] Kokshaysky, N.V., Tracing the wake of a flying bird. *Nature Lond.*, **279**, pp. 146–148, 1979.
[51] Rayner, J.M.V., Vorticity and animal flight. *Aspects of animal movement*, eds. H.Y. Elder & E.R. Trueman, *Semin. Ser. Soc. Exp. Biol.*, **5**, Cambridge University Press: Cambridge, UK, pp. 177–199, 1980.
[52] Spedding, G.R., Rosen, M. & Hedenstrom, A., A family of vortex wakes generated by a thrush nightingale in free flight in a wind tunnel over its entire natural range of flight speeds. *J. Exp. Biol.*, **206**, pp. 2313–2344, 2003.
[53] Wootton, R.J., Support and deformability in insect wings. *J. Zool.*, **193**, pp. 447–468, 1981.
[54] Wootton, R.J., Functional morphology of insect wings. *Ann. Rev. Ent. Palo Alto*, **37**, pp. 113–140, 1992.

[55] Brodsky, A.K. & Ivanov, V.P., Functional assessment of wing structure in insects. *Ent. Rev. Wash.*, **62**, pp. 35–52, 1983.
[56] Brodsky, A.K., *The Evolution of Insect Flight*, Oxford University Press: Oxford, 1994.
[57] Grodnitsky, D.L., *Form and Function in Insect Wings*, Johns Hopkins University Press: Baltimore, Maryland, 1999.
[58] Nachtigall, W., Insect wing bending and folding during flight without and with an additional prey load. *Entomologia Generalis*, **25**, pp. 1–16, 2000.
[59] Combes, S.A. & Daniel, T.L., Shape, flapping and flexion: wing and fin design for forward flight. *J. Exp. Biol.*, **204**, pp. 2073–2085, 2001.
[60] Ennos, A.R., The inertial cause of wing rotation in Diptera. *J. Exp. Biol.*, **140**, pp. 161–169, 1988.
[61] Osborne, M.F.M., Aerodynamics of flapping flight with application to insects. *J. Exp. Biol.*, **28**, pp. 221–245, 1951.
[62] Jensen, M., Biology and physics of locust flight. III. The aerodynamics of locust flight. *Phil. Trans. R. Soc. B*, **239**, pp. 511–552, 1956.
[63] Brackenbury, J., *Insects in Flight*, Blandford: London, 1992.
[64] Dalton, S., *Caught in Motion*, Weidenfeld and Nicholson: London, 1982.
[65] Ellington, C.P., The novel aerodynamics of insect flight. Application to micro-air vehicles. *J. Exp. Biol.*, **202**, pp. 3431–3438, 1999.
[66] Ellington, C.P., The aerodynamics of hovering insect flight. IV. Aerodynamic mechanisms. *Phil. Trans. R. Soc. B*, **305**, pp. 79–113, 1984.
[67] Lehmann, F.O., The mechanisms of lift enhancement in insect flight. *Naturwissenschaften*, **91**, pp. 101–122, 2004.
[68] Weis-Fogh, T., Quick estimates of flight fitness in hovering animals, including novel mechanisms for lift production. *J. Exp. Biol.*, **59**, pp. 79–104, 1973.
[69] Maxworthy, P., Experiments on the Weis-Fogh mechanism of lift generation by insects in hovering flight. Part 1. Dynamics of the 'fling'. *J. Exp. Biol.*, **93**, pp. 47–63, 1979.
[70] Miller, L.A. & Peskin, C.S., A computational fluid dynamics model of 'clap and fling' in the smallest insects'. *J. Exp. Biol.*, **208**, pp. 195–202, 2005.
[71] Sun, M. & Tang, J., Unsteady aerodynamic force generation by a model fruitfly in flapping motion. *J. Exp. Biol.*, **205**, pp. 55–70, 2002.
[72] Ellington, C.P., van den Berg, C., Willmott, A.P. & Thomas, A.L.R., Leading edge vortices in insect flight. *Nature Lond.*, **384**, pp. 626–630, 1996.
[73] Videler, J.J., Stamhuis, E.J. & Povel, G.D.E., Leading edge vortex lifts swifts. *Science*, **306**, pp. 1960–1962, 2004.
[74] Altshuler, D.L., Dudley, R. & Ellington, C.P., Aerodynamic forces of revolving hummingbird wings and wing models. *J. Zool.*, **264**, pp. 327–332, 2004.
[75] Warrick, D.R., Tobalske, B.W. & Powers, D.R., Aerodynamics of the hovering hummingbird. *Nature*, **435**, pp. 1094–1097, 2005.
[76] Usherwood, J.R. & Ellington, C.P., The aerodynamics of revolving wings. II. Propellor force coefficients from mayfly to quail. *J. Exp. Biol.*, **205**, pp. 1565–1576, 2002.
[77] Thomas, A.L.R., Taylor, G.K., Srygley, R.B., Nudds, R.L. & Bomphrey, R.J., Dragonfly flight: free-flight and tethered flow visualizations reveal a diverse array of unsteady lift-generating mechanisms, controlled primarily by angle of attack. *J. Exp. Biol.*, **207**, pp. 4299–4323, 2004.
[78] Bomphrey, R.J., Lawson, N.J., Harding, R.J., Taylor, G.K. & Thomas, A.L.R., The aerodynamics of *Manduca sexta*: digital particle imaging velocimetry analysis of the leading edge vortex. *J. Exp. Biol.*, **208**, pp. 1079–1094, 2005.

[79] Norberg, U.M. & Rayner, J.M.V., Movement. Flight. *The Cambridge Encyclopaedia of Ornithology*., eds. M. Brooke & T. Birkhead, Cambridge University Press: Cambridge, UK, pp. 53–66, 1991.
[80] von Helversen, O., Blütenbesuch bei Blumenfledermäusen: Kinematik des Schwirrfluges und Energiebudget im Freiland. *Biona Report 5, Bat flight – Fledermausflug*, ed. W. Nachtigall, Gustav Fischer Verlag: Stuttgart, pp. 107–126, 1986.
[81] Dudley, R. & Winter, Y., Hovering flight mechanics of neotropical flower bats (Phyllostomidae: Glossophaginae) in normodense and hyperdense gas mixtures. *J. Exp. Biol.*, **205**, pp. 3669–3677, 2002.
[82] Ellington, C.P., The aerodynamics of hovering insect flight. III. Kinematics. *Phil. Trans. R. Soc. B*, **305**, pp. 41–78, 1984.
[83] Norberg, U.M., Hovering flight in the pied flycatcher. *Swimming and flying in nature*, eds. T.Y.-T. Wu, C.J. Brokaw & C. Brennen, Plenum Press: New York, pp. 869–880, 1975.
[84] Norberg, U.M., Aerodynamics of hovering flight in the long-eared bat *Plecotus auritus*. *J. Exp. Biol.*, **65**, pp. 459–470, 1976.
[85] Aldridge, H.D.J.N., Kinematics and aerodynamics of the greater horseshoe bat, *Rhinolophus ferrumequinum*, in horizontal flight at various flight speeds. *J. Exp. Biol.*, **126**, pp. 479–497, 1986.
[86] Ellington, C.P., The aerodynamics of hovering insect flight. I. The quasi-steady analysis. *Phil. Trans. R. Soc. Lond. B*, **305**, pp. 1–15, 1984.
[87] Bunker, S.J., *Form, Flight and Performance in Butterflies (Lepidoptera: Papilionoidea and Hesperioidea.* PhD thesis, University of Exeter, 1993.
[88] Schmidt-Nielsen, K., Locomotion. Energy cost of swimming, flying and running. *Science*, **177**, pp. 222–228, 1972.
[89] Rayner, J.M.V., Bounding and undulating flight in birds. *J. Theor. Biol*., **117**, pp. 47–77, 1985.
[90] Withers, P.C. & Timko. P.L., The significance of ground effect to the aerodynamic cost of flight and energetics of the black skimmer [*Rhyncops nigra*]. *J. Exp. Biol*., **70**, pp. 13–26, 1977.
[91] Rayner, J.M.V., On the aerodynamics of animal flight in ground effect. *Phil. Trans. R. Soc. Lond. B*, **334**, pp. 119–128, 1991.
[92] Altringham, J.D., *Bats. Biology and Behaviour*, Oxford University Press: Oxford, New York, Tokyo, 1996.
[93] Pennycuick, C., Speeds and wing-beat frequencies of migrating birds compared with calculated benchmarks. *J. Exp. Biol.*, **204**, pp. 3283–3294, 2001.
[94] Bennett, S.C., New evidence on the tail of the pterosaur *Pteranodon* (Archosauria, Pterosauria). *Short Papers of the Fourth Symposium on Mesozoic Terrestrial Ecosystems*, eds. P.M. Currie & E.H. Koster, pp. 18–23, 1987.
[95] Padian, K., A functional analysis of flying and walking in pterosaurs. *Palaeobiology*, **9**, pp. 218–239, 1983.
[96] Padian, K., Pterosaurs: were they functional birds or functional bats? *Biomechanics in Evolution*, eds. J.M.V. Rayner & R.J. Wootton, *Seminar Series of the Society for Experimental Biology*, Cambridge University Press, pp. 145–160, 1990.
[97] Pennycuick, C.J., On the reconstruction of pterosaurs and their manner of flight, with notes on vortex wakes. *Biol. Rev.*, **63**, pp. 209–231, 1988.
[98] Bennett, S.C., Pterosaur flight: the role of actinofibrils in wing function. *Hist. Biol.*, **14**, pp. 255–284, 2000.

[99] Wilkinson, M.T., Unwin, D.M. & Ellington, C.P., High lift function of the pteroid bone and forewing of pterosaurs. *Proc. R. Soc. Lond*., **273**, pp. 119–126, 2006.
[100] Thomas, A.L.R., On the aerodynamics of birds' tails. *Phil. Trans. R. Soc. Lond. B*, **340**, pp. 361–380, 1993.
[101] Evans, M.R., Birds' tails do act like delta wings but delta-wing theory does not always predict the forces they generate. *Proc. R. Soc. Lond. B*, **270**, pp. 1379–1385, 2003.
[102] Smith, C.W., Herbert, R.C., Wootton, R.J. & Evans, K.E., The hind wing of the desert locust (*Schistocerca gregaria* Forskål). II. Mechanical properties and functioning of the membrane. *J. Exp. Biol.*, **203**, pp. 2933–2943, 2000.
[103] Bullen, R. & McKenzie, N.L., Bat airframe design: flight performance, stability and control in relation to foraging ecology. *Austral. J. Zool.*, **49**, pp. 235–261, 2001.
[104] Byrne, D.N., Buchmann, S.L. & Spengler, H.G., Relationship between wing loading, wing beat frequency and body mass in homopterous insects. *J. Exp. Biol.*, **135**, pp. 9–23, 1988.
[105] Sotavalta, O., The flight-tone (wing-stroke frequency) of insects. *Acta Ent. Fenn.*, **4**, pp. 1–117, 1947.
[106] Cooper, A.J., *Limitations on Bumblebee Flight Performance.* PhD thesis, University of Cambridge, 1993.
[107] Riley, J.R. & Osborne, J.L., Flight trajectories of foraging insects: observations using harmonic radar. *Insect Movement: Mechanisms and Consequences*, eds. I. Woiwod & D.R. Reynolds, CABI Publishing: Oxford and New York, pp. 129–157, 2001.
[108] Hazlehurst, G.A. & Rayner, J.M.V., Flight characteristics of Triassic and Jurassic Pterosauria: an appraisal based on wing shape. *Paleobiology*, **18**, pp. 447–463, 1992.
[109] van den Berg, C. & Rayner, J.M.V., The moment of inertia of bird wings and the inertial power requirement in flapping flight. *J. Exp. Biol.*, **198**, pp. 1655–1664, 1995.
[110] Wootton, R.J. & Kukalova-Peck, J., Flight adaptations in Palaeozoic Palaeoptera (Insecta). *Biol. Revs. Camb. Phil. Soc.*, **75**, pp. 129–167, 2000.
[111] Tucker, V.A., Gliding birds: reduction of induced drag by wingtip slots between the primary feathers. *J. Exp. Biol.*, **180**, pp. 285–310, 1993.
[112] Wootton, R.J., Wings. *Encyclopaedia of Entomology*, eds. V.H. Resh & R.T. Cardé, Elsevier: Amsterdam and New York, 2003.
[113] Wootton, R.J., Evans, K.E., Herbert, R.C. & Smith, C.W., The hind wing of the desert locust (*Schistocerca gregaria* Forskal). I. Functional morphology and mode of operation. *J. Exp. Biol.*, **203**, pp. 2921–2931, 2000.
[114] Wootton, R.J., Invertebrate paraxial locomotory appendages: design, deformation and control. *J. Exp. Biol.*, **202**, pp. 3333–3345, 1999.
[115] Harrison, J.F. & Roberts, S.P., Flight respiration and energetics. *Ann. Rev. Physiol. Palo Alto*, **62**, pp. 179–205, 2000.
[116] Dudley, R. & Ellington, C.P., Mechanics of forward flight in bumblebees. II. Quasi-steady lift and power requirements. *J. Exp. Biol.*, **148**, pp. 53–88, 1990.
[117] Pennycuick, C.J., Power requirements for horizontal flight in the pigeon *Columba livia*. *J. Exp. Biol.*, **49**, pp. 527–555, 1968.
[118] Pennycuick, C.J., Mechanics of flight. Avian Biology (Vol. 5), eds. D.S. Farner & J.R. King, Academic Press: London, pp. 1–75, 1975.
[119] Rayner, J.M.V., A new approach to animal flight mechanics. *J. Exp. Biol.*, **80**, pp. 17–54, 1979.
[120] Pennycuick, C.J., *Bird Flight Performance: A Practical Calculation Manual*. Oxford University Press: Oxford, New York, Tokyo, 1989.

[121] Ellington, C.P., Limitations on animal flight performance. *J. Exp. Biol.*, **160**, pp. 71–91, 1991.
[122] Rayner, J.M.V., Estimating power curves for flying vertebrates. *J. Exp. Biol.*, **202**, pp. 3449–3461, 1999.
[123] Tobalske, B.W., Hedrick, T.L ., Dial, K.P. & Biewener, A.A., Comparative power curves in bird flight. *Nature*, **421**, pp. 363–366, 2003.
[124] Berger, M., Sauerstoffverbrauch von Kolibris (*Colibri coruscans* und *C. thalassinus*) beim Horizontalflug. *BIONA Report 3*, eds. W. Nachtigall, Gustav Fischer Verlag: Stuttgart, pp. 307–314, 1985.
[125] Voigt, C.C. & Winter, Y., Energetic cost of hovering flight in nectar-feeding bats (Phyllostomidae: Glossophaginae) and its scaling in moths, birds and bats. *J. Comp. Physiol. B*, **169**, pp. 38–48, 1999.
[126] Wakeling, J.M. & Ellington, C.P., Dragonfly flight. III. Lift and power requirements. *J. Exp. Biol.*, **200**, pp. 583–600, 1997.
[127] Dudley, R. & de Vries, P.J., Flight physiology of migrating *Urania fulgens* (Uraniidae) moths: kinematics and aerodynamics of natural free flight. *J. Comp. Physiol. A.*, **167**, pp. 145–154, 1990.
[128] Willmott, A.P. & Ellington, C.P., The mechanics of flight in the hawkmoth *Manduca sexta.* II. Aerodynamic consequences of kinematic and morphological variation. *J. Exp. Biol.*, **200**, pp. 2723–2745, 1997.
[129] Hedenstrom, A. & Alerstam, T., Optimal flight speed of birds. *Phil. Trans. R. Soc. Lond. B*, **348**, pp. 471–487, 1995.
[130] Alerstam, T., Gudmundsson, G.A. & Larsson, B., Flight tracks and speeds of Antarctic and Atlantic seabirds: radar and optical measurements. *Phil. Trans. R. Soc. Lond. B*, **340**, pp. 55–67, 1993.
[131] Norberg, R.Å., Optimal flight speeds in birds when feeding young. *J. Anim. Ecol.*, **50**, pp. 573–477, 1981.
[132] Videler, J.J., Weihs, D. & Daan, S., Intermittent gliding in the hunting flight of the kestrel, *Falco tinnunculus* L. *J. Exp. Biol.*, **102**, pp. 1–12, 1983.
[133] Blake, R.W., Kolotylo, R. & De La Cueva., Flight speeds of the barn swallow, *Hirundo rustica. Can. J. Zool.*, **68**, pp. 1–5, 1990.
[134] Jones, G. & Rayner, J.M.V., Optimal flight speed in pipistrelle bats, *Pipistrellus pipistrellus. Eur. Bat Res.*, **1987**, pp. 247–253, 1989.

Chapter 9

Insect observations and hexapod design

M. Randall*
Faculty of Engineering, University of West of England, Bristol, UK.

Abstract

The hexapod leg configuration for walking robots is described in relation to the stick insect. The role of hairs, sensory organs and receptors in leg functions such as mediating load compensation and obstacle avoidance is explained. Observations on gait reveal important factors for informing models of hexapod walking and leads on to strategies for dealing with rough terrains.

1 Introduction

Nature has presented us with many 'blueprints' for hexapod machines in the form of millions of species of insects. The goal is not to re-create an insect in mechatronic form. Instead, in considering insect walking, the structure and nature of sensory machinery and the control processes employed by insects when walking over rough terrain, much can be learned that will inform the engineer when designing a walking machine, and particularly a hexapod. Here, a large volume of biological study is reviewed. The last section summarises the relevant findings for the design of the hexapod used in this research. In [1], we coined the term 'intelligent hexapod bio-robotics' to describe this field of research.

There is a huge amount of literature on the subject of insect neurophysiology and neurobiology. For extensive reviews, readers are referred to [2, 3] and more recently [4, 5].

2 Justification for biologically inspired engineering

There are many reasons for considering a biological approach to robotics a useful direction for making a more sophisticated hexapod. It has been the experience of a number of researchers, particularly Beer and his colleagues at Case Western Reserve University, that 'copying' strategies employed by insects has only been beneficial to their work [6]. Some of the reasons follow.

---

*Mark Randall was only 29 years old when he died on 21 September 2000. Mark had been keen to contribute a chapter to this book and his wife, Emma, offered this chapter from Mark's own book *Adaptive Neural Control of Walking Robots*, Professional Engineering Publishing, 2001.

When an insect loses a leg, it immediately adopts a different gait more appropriate to the new configuration [3].

> They do this with a relatively slow and noisy collection of nerve cells that, at first glance, would seem to be no match for even the slowest modern microprocessor. Yet even the lowly cockroach can outperform the most sophisticated autonomous robot at nearly every turn [6].

Animals use only as much as 10% of the energy used by self-powered wheeled or tracked vehicles when travelling over rough terrain [7]. Natural environments may not be even or level. They could be slippery or provide poor support. They could have significant vertical variations, large obstacles or sparse footholds. Legged locomotion is, however, well suited to such irregular environments but requires continuous adaptation to the environmental conditions. These conditions could be slowly changing as during growth, or rapidly changing as in the negotiation of rough terrain at high speed. Changes could even occur while a step is underway, for instance the weight of a leg could cause small branches or twigs to bend significantly. However, all insects are adept at rough terrain walking, avoiding obstacles and some can even climb upside down. Some stick insects (Fig. 1) can carry up to four times their own body weight on their backs [3],

Figure 1: The stick insect, *Carausius morosus.*

.

and they have their power source on board. Furthermore, a hexapod can have the advantage of being statically stable, because it can always hold its weight on at least three legs. Delcomyn [8] describes the versatility of insect walking:

> Walking insects are exceptionally adaptable in the way they move and use their legs. They can walk on irregular surfaces like leaves and branches almost as well as they can on flat, level ground. They can walk forward and backward, and sometimes even sideways, and do so right side up or upside down. Insects can also walk after injury or damage to their legs or even after suffering the complete loss of one or more legs.

At a lower level, the motor machinery used to move individual limbs faces many demands. In the locust, for instance,

> It must maintain posture for long periods whilst allowing for adjustments of tone, perform delicate and intricate movements that require graded changes of force, and at the same time be capable of producing quickly considerable amounts of force that are required for ballistic movements. It must participate in co-ordinated movements with the other limbs and other parts of the body, and at the same time be able to make independent adjustments as part of the local reflexes [9].

Insects have adapted to live in different environments. Cockroaches, earwigs and crickets prefer flat surfaces or the inside of crevices. Mantids and stick insects, on the other hand, climb on shrubs, grasses and trees. Locusts and bush crickets are less specialised, so they can climb or run on horizontal surfaces equally well [10].

The stick insect, *Carausius morosus* (Fig. 1) was chosen as a model for the hexapod robot built as the test platform used in this research. At the beginning of this research, the goal was to produce the control system for an 'autonomous' or 'semi-autonomous' hexapod that walks over extremely irregular terrain with sparse footholds. In research conducted by NASA, Dante II was developed for similar purposes, but had to be lifted out of a volcano because it tipped over when negotiating 'unanticipated contingency' [6]. The choice of *Carausius* then, was due to a number of reasons. First, the environment in which this stick insect lives and walks has few footholds, which may even change from step to step, for instance as a result of wind moving a twig. The insect has evolved very accurate foot placement capabilities and a number of its strategies, described later, could readily be employed in a robot. This accuracy of foot placement has a trade-off: the insect is a very slow walker as compared to a 'running' insect like a cockroach. Since most hexapod robots are slow, this puts any machine developed as a result of this research at no particular disadvantage, if for instance it were to be commercialised. Second, there is a vast amount of literature describing stick insect behaviours, and those of *Carausius* in particular. Although work concerning other insects is also reported here, partly for comparative reasons, '*Carausius* is an ideal object for the study of walking' [2].

Finally, although not part of this research, the relatively small number of neurones in the stick insect (around 6000) makes it potentially feasible that an exhaustive emulation of each one would result in a replication of the full stick insect control system, given the correct sensory, motor and power apparatus. However, Delcomyn [8] argues against this approach because the millions of years of evolutionary pressure were not focused entirely on achieving an efficient walking system.

There is a proviso for these observations. *Carausius* is a nocturnal climbing animal and most of the observations of its walking have been conducted by stimulating escape walks. Results from one insect like the relatively slow walking stick insect may not be transferable to models of other faster walking insects, like cockroaches, or vice versa. This would be a critical point for an engineer trying to emulate the whole control and mechanical structure of a single insect.

However, this is not the purpose here. The intention is merely to draw from biological observations in order to develop technological strategies for implementation on the final hexapod.

3 Anatomy and leg structure of insects

3.1 Body segments

The body consists of three jointed segments: the prothorax, the mesothorax and the metathorax (Fig. 2). The joint between the pro- and mesothorax is very flexible and in normal behaviour is moved both horizontally and vertically. It is possible to bend passively the meso-metathorax joint by up to 40°. However, when walking over large obstacles, this joint is maintained almost rigidly. In the extreme case of a horizontal-to-vertical transition, this joint has been observed to move actively through angles of up to 30° [2].

Location of the segments along the body and their average heights above the ground in different walking situations can be seen in Table 1.

3.2 Leg structure

The legs need to provide support for the body during standing and walking, and to maintain balance during turns or other procedures. In this way, walking is different from other locomotor systems, for instance swimming, since the legs must always support the body during walking as well as provide propulsion. This requirement can put severe constraints upon the leg movement, especially during climbing or walking upside down, where contact with the surface is required at all times [13].

Insects use their legs as struts and levers during standing and walking. In some insects, the leg structure is specialised for particular functions, for instance digging (the mole cricket), prey capture (the praying mantis) or jumping or kicking (the locust) [9, 13].

Insects have three pairs of legs, one pair on each body segment. Each leg itself consists of four segments (Fig. 3): the coxa, which joins the leg to the body, the femur, the tibia, and the tarsus (or foot). Sometimes, an additional segment can be located between the coxa and femur, called the 'trochanter' [14], but in terms of joint control, it works as part of the coxa. The joints of the leg are simple hinge joints, but often the coxa-body joint is more like a ball and socket joint. The respective measurements of these joint links are given in Table 1.

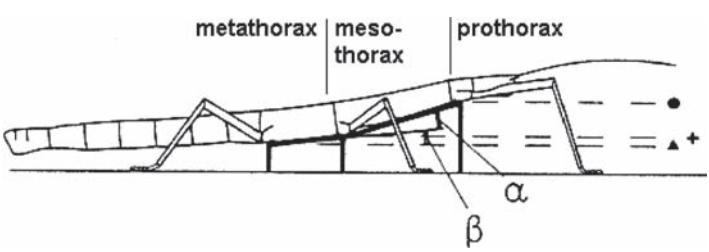


Figure 2: The body segments and intersegmental joints of the stick insect *Carausius morosus*, showing the height of the prothoracic coxae (•), the height of the mesothoracic coxae (▲) and the height of the metathoracic coxae (+). Also shown are the pro-mesothoracic joint angle (α) and the meso-metathoracic joint angle (β). Adapted from [12] (fig. 1, p. 26, © Springer-Verlag, 1976).

Table 1: Stick insect measurements [11, 12]. The body segments and segment heights relate to Fig. 2, while the leg segment lengths relate to Fig. 3. Measurements are taken from 20 animals for the vast majority of the above values (number used in the last four sets of parameters not given [12]).

Parameters	Measurement
Total length of body	73.2 (±4.6) mm
Distance between frontal part of head and …	
Coxa of prothorax	7.1 (±0.8) mm
Coxa of mesothorax	24.6 (±2.1) mm
Coxa of metathorax	35.5 (±2.8) mm
Animal's centre-of-gravity	35.8 (±3.0) mm
Foreleg	
Length of coxa	1.6 (±0.2) mm
Length of trochanterofemur	14.7 (±1.3) mm
Length of tibia	13.7 (±1.4) mm
Orientation angle λ	75° (±10°)
Orientation angle μ	40° (±10°)
Middle leg	
Length of coxa	1.5 (±0.2) mm
Length of trochanterofemur	11.4 (±0.9) mm
Length of tibia	10.7 (±0.9) mm
Orientation angle λ	75° (±10°)
Orientation angle μ	40° (±10°)
Hind leg	
Length of coxa	1.5 (±0.2) mm
Length of trochanterofemur	12.3 (±1.5) mm
Length of tibia	11.8 (±1.7) mm
Orientation angle λ	135° (±10°)
Orientation angle μ	50° (±10°)
Prothorax	
Distance from above horizontal plane during walking	9.1 (±1.4) mm
Distance from horizontal beam when walking upside down	19.0 (±2.0) mm
Distance from substrate when walking up vertical path	7.6 (±2.3) mm
Mesothorax	
Distance from above horizontal plane during walking	5.9 (±0.9) mm
Distance from horizontal beam when walking upside down	19.0 (±2.0) mm
Distance from substrate when walking up vertical path	7.3 (±2.1) mm
Metathorax	
Distance from above horizontal plane during walking	4.3 (±0.6) mm
Distance from horizontal beam when walking upside down	20.0 (±2.5) mm
Distance from substrate when walking up vertical path	7.8 (±2.2) mm
Total Weight	820 (±160) mg
Pro-mesothoracic angle (α)	
Range during active movements	−20° to +30°
Extreme range	−40° to +40°
Meso-metathoracic angle (β)	
While walking over horizontal plane	7.6° (±4.6°)
Extreme range	−12.6° (±5.0°) to +31.6° (±7.7°)

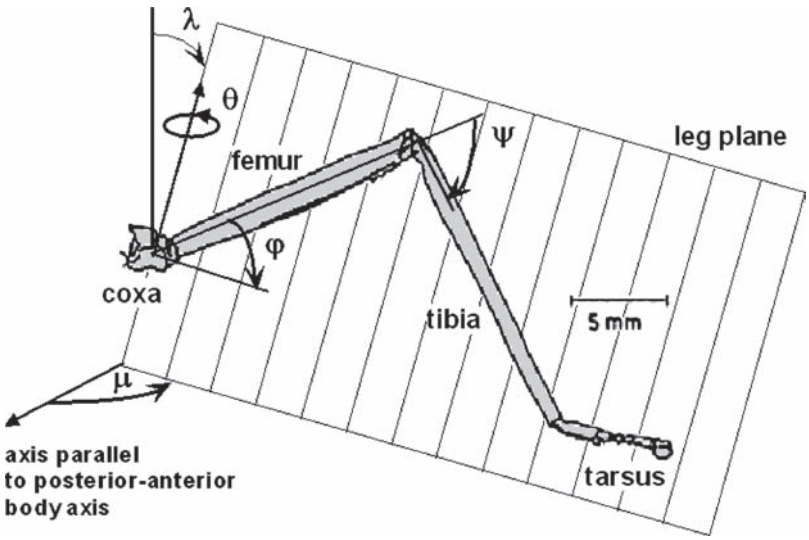


Figure 3: The structure of a leg in the stick insect *Carausius morosus*. The orientation of the leg relative to the longitudinal axis of the body and the vertical can be described by two angles λ and μ. Each of the main joints of the leg exists in a common plane called the 'leg plane'. The angle around which the coxa moves the leg forwards and backwards is here denoted as θ, the coxa-trochanteris is φ and the femorotibial joint is ψ. Note the multi-jointed tarsus.

The resting position of a leg in a stick insect is not orthogonal to the body in either the vertical or horizontal directions. The three actively powered joints can however be considered to exist in a common plane, the orientation of which with respect to the body can be described by two angles λ and μ, as can be seen in Fig. 3. This plane has been called the 'leg plane' [15].

Measurements by Burrows [9] in the locust show that the neuronal control structure of the leg is highly complex. There are very few neurones specialised for one task, for instance local non-spiking interneurones are not limited to motor neurones affecting just one joint and can be involved in both the extension of the tibia, depression of the tarsus or rotation of the coxa.

Cruse [11] showed how legs are used in different walking situations. For instance, in horizontal walks, the forelegs are used as sensors and for braking, the middle legs carry the body weight and the hind legs provide support and extra thrust. When walking upside down by hanging from a horizontal bar, all the legs are used to carry the body weight. In vertical walks, all the legs are used to provide support and thrust. Hence, the motor neurone output is heavily dependent on global sensory information. Finally, in terms of leg function, lateral stability is provided by passive forces in the leg joints [3].

3.3 Leg joints

The joint angles in a resting animal can be determined by a number of factors, but in the main they depend upon muscle properties and physiological effects. Fourtner [16] describes these as: first, passive resting tension in the muscles; second, slow excitations in the motor neurones; third, residual tensions after the cessation of neuronal control, and finally hormonal influences which can induce contractions in the absence of motor activity.

In the complete absence of neural input in a wide variety of arthropods studied by Fourtner [16], it was shown that there is sufficient tension in the muscles to hold a joint at a given position in the resting animals. However, there are also active feedback systems in each of the joints and these are described in the next subsections.

3.3.1 Body-coxa joint

The coxa joint (or basal joint) attaches the leg to the body. It can be considered as a ball joint. There is evidence to show that both degrees of freedom of this ball joint are used. However, the up–down movement is so small, it can be ignored to a first approximation [15]. Thus, although the angles λ and μ (Fig. 3) remain almost unchanged during stance, it has been shown that their slight variation allows for a smaller range of angular displacement in the coxa joint. The profile of these changes has been described by Cruse and Bartling [15]. Most of the muscle power is directed to changing the coxal joint through the angle θ (Fig. 3) and so the variations in the other angles are likely to be a result of passive compliance rather than due to control of the step.

Graham [17] reported that the coxa joint is under servo-control with position feedback.

3.3.2 Coxa-trochanteris

The coxa-trochanteris (or trochanter) is responsible for the main up–down movement of the leg (through angle φ in Fig. 3). This joint possesses a feedback loop with a hair plate as a receptor organ providing the motor feedback. This feedback loop must have a dynamic component in it since a stick insect can support up to four times its normal weight without collapsing. Bässler [2] reported that the dynamic component is sensitive to low velocities.

3.3.3 Femorotibial joint

The femur-tibia joint is under reflex control during walking if the leg is disturbed from its expected position by externally imposed forces. The joint uses servo-control with velocity feedback [3, 17]. Cruse [18] and Cruse and Pflüger [19] were able to show that the control mechanism of the femorotibial joint in the middle leg of the stick insect uses negative feedback in fixed animals, in free-standing animals with tarsal contact and in free-walking insects during retraction. Also, for a fixed animal making active movements, they concluded that the servomechanism for adaptive control of the joint had either zero or positive feedback.

The gains of the reflex loop for a standing leg in a stationary animal and the supporting leg of a walking animal are different. Cruse and Schmitz [20] were able to conclude from this that the state of an individual leg (‘walking’ or ‘standing’) is not controlled by the neuronal subsystem belonging to the leg, but by the behaviour of the organism as a whole.

The femur-tibia joint can flex and extend at maximum rates of 1800°/s and 1200°/s, respectively, but usually these movements are unidirectional and do not persist for very long [2].

3.3.4 The tarsus

The movement of the tibia-tarsus joint is achieved by three muscles, which control raising and lowering of the tarsus, lateral movements and even possibly rotation around the longitudinal axis [2].

There are a number of joints in the tarsus. They are all controlled by one muscle, the *retractor unguis*. This muscle is attached to a single tendon that extends through most of the femur and the tibia. It has no antagonistic muscle but works against ‘elastic bands’ in the tarsus [2]. Normally the tarsus is held in a slightly curved manner when the leg has no tarsal contact. Once the tarsus makes contact with the substrate (normally a branch for a stick insect), it grips it by tightening the tarsal joints around it. This is done by the contraction of the *retractor unguis* on detection of

the surface. When the tarsus does contact the ground, the forces are directed against a series of sticky gripping pads (*euplantae*) on the underside of the tarsus [2, 3].

The underside of the first tarsal segment has a row of very sturdy hairs which enable the insect to grip smooth as well as rough surfaces. Most of the tarsus is densely covered with large tactile hairs [2].

3.4 Leg sense organs and proprioceptors

For most sensory cells, there is an inverse relationship between sensitivity and range. It is usual that a very sensitive cell will only respond to a restricted range of stimuli and vice versa. Hence, for a large range of stimuli, most animals will have an array of cells for responding, where each group responds to a different portion of the range. This is known as 'range fractionation'. When considering sense organs, it is important not to assume that all sensory cells on the leg are part of the walking system. Legs are also used for testing surfaces and vibrations, for grooming and in some cases for fending off predators. Also, there is a huge redundancy in the number of cells that each leg has, to the extent that whole regions of the sensory apparatus can be removed without there being any significant effect on walking co-ordination or other behaviours that rely on the cells [14].

The sense organs of the leg provide position feedback control for each of the joint angles to enable body stabilisation, correct body support and probably to assist propulsion ([3] for stick insects; [10] for locusts). There are several types of sense organs. Most are known as mechanoreceptors which respond to mechanical stimuli and these are shown in Fig. 4 [14]. Others include

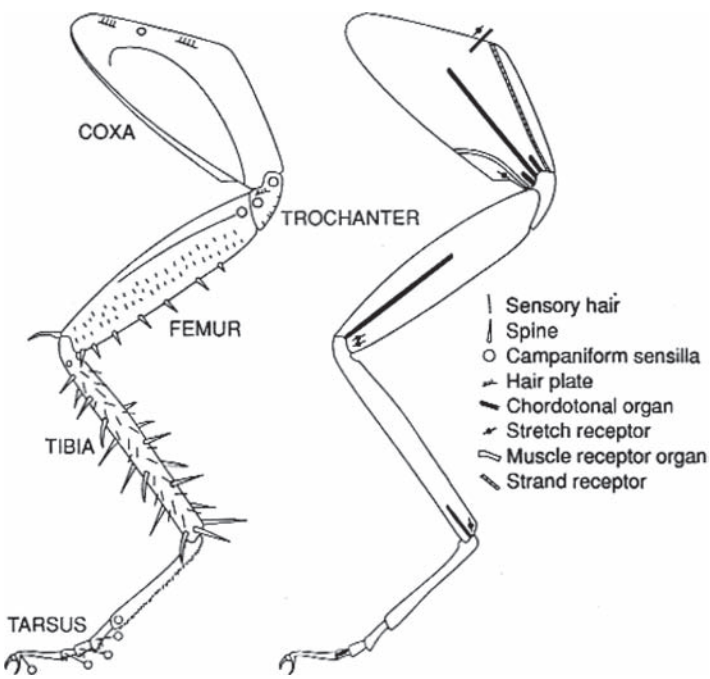


Figure 4: The inside view of the right leg of typical insects. The left image shows the sense organs in or on the cuticle (exoskeleton) and right image shows the organs inside the leg. Not all insects have the full array of organs depicted.

chemoreceptors which respond to specific chemicals in the insect's vicinity. Some insects also have sensors in their legs which respond to vibrations in the ground or sound waves. Since it is only mechanoreceptors that are thought to take part in agile locomotion, only these are described, of which there are around eight types. Four types, the muscle receptor organs, strand receptors, stretch receptors and chordotonal organs are internal to the leg musculoskeletal. The next three types – hair plates, spines and sensory hairs – are found on the surface of the leg and provide tactile information to the central nervous system. The last type, called *campaniform sensilla*, are stress receptors that are embedded in the cuticular exoskeleton of the insect. These mechanoreceptors can be categorised according to their function: proprioceptors, tactile receptors and stress receptors [14]. An example of each kind of mechanoreceptor will be described below. Proprioceptors are perhaps the most important, and a more detailed description of them will be provided first.

A proprioceptor is defined as a 'sense organ that encode[s] information about the position and the movements of parts of the body relative to one another' [14]. Such proprioceptors include the chordotonal organ, strand receptors, stretch receptors, muscle receptors and hair plates. Proprioceptive information from such sensors can influence both position and timing of leg movements during walking [21]. Typically proprioceptive feedback can have a number of roles [22]:

1. It can determine walking direction.
2. It can accelerate or delay the transition from one phase of the step cycle to the following one, especially the transition from stance to swing. These kinds of influences establish the temporal co-ordination of the movements of the legs.
3. It can guide the movement of the other legs during the swing phase and thus influence spatial co-ordination.
4. It can influence the fine structure of the other leg's motor output, either directly, or indirectly by influencing its motor programme . . . Such influences cannot alter the temporal co-ordination. The spatial co-ordination is only influenced during the swing phase.

Insects have a copious supply of such proprioceptors along the whole length of the leg, and they function as part of the control loops of the different behavioural patterns by feeding back the motor output [23].

Since stick insects are relatively slowly walking and climbing animals, enough time is available for sensory feedback, and so it is likely they depend on it highly. Faster walking animals, like cockroaches, probably rely a lot less on peripheral (sensory) mechanisms [22]. Certainly, for cockroaches, sensory information is critical to proper timing signals between the legs during slow walking, but is not as important during fast walks [8].

3.4.1 Chordotonal organ

The chordotonal organ – a proprioceptor – measures position, local bending and leg strain. There is roughly one chordotonal organ per leg segment. A chordotonal organ consists of as many as 100 sensory cells or 'receptors', each of which sends a separate sensory signal to the central nervous system. Most of these signals are caused by flexion of the more distal segment, for instance the femoral chordotonal organ is activated by flexion of the tibia [14]. In the hind legs, it is possibly the major proprioceptor for determining how large a step is and for tarsus position, where the leg's action is parallel to the body. Therefore, the chordotonal organ in the hind leg is the sensor used in lateral stabilisation of the body [3]. Bässler [24] showed that the chordotonal organ acts as a velocity transducer during walking, and when it signals a rearward movement, it is enough to excite the whole animal to walk. In the locust, the chordotonal organ is involved in the control loop of the leg movement to the point of providing trajectory-specific motor co-ordination [10].

Hofmann *et al.* [23] and Hofmann and Koch [25] showed that in the stick insect *Cuniculina impigra*, the sensory apparatus of the femoral chordotonal organ contains position sensors, position-velocity sensors, unidirectionally-sensitive velocity receptors, as well as velocity-acceleration sensors and acceleration sensitive units. Some of the sensors had range fractionation providing precise information about the movement as well as the position at which it occurs. No bi-directional position-velocity sensors were found and no range fractionation was found in the acceleration sensors. A possible role for the acceleration sensors would be in an alarm system. This is because the acceleration sensors' response is asymmetric, peaking earlier than both velocity and position, thus they are able to detect changes in mechanical conditions faster [25].

Hofmann *et al.* [23] were unable to determine how the chordotonal organ worked in the femorotibial control loop. However, Weiland and Koch [26] were able to show that in *Carausius* the chordotonal organ was involved in controlling the end-stops of the femur-tibia joint, and that it reacted to changes in both velocity and acceleration.

The chordotonal organ plays a major role in the control loop of the stick insect behaviour known as 'catelepsy' (see Section 4.6).

3.4.2 Campaniform sensilla

Campaniform sensilla are cuticular stress receptors that measure local stress in leg segments and mediate load compensation during slow walking [3]. In the locust, Burrows [9] reports that there are approximately 450 sensilla on the hind leg tarsus alone. It is not at all clear how the central nervous system processes this vast flood of information.

Dean and Schmitz [27] showed that the two groups of campaniform sensilla on the ventral coxal hair plate in *Carausius* have different roles during walking. The first limits how far forward a leg can swing, while the second group provides targeting information used to end the swing of the leg so that it touches the substrate close to the tarsus of the leg in front (the 'targeting behaviour' described in Section 7.1).

The tibial campaniform sensilla seem to aid in load compensation and in limiting muscle tensions during walking. The trochanteral campaniform sensilla in *Carausius* compensate for changes in cuticular stress due to movements in the muscles that act in the forward-backward plane of motion. This is the case in both walking and standing animals for both active movements (counteracting bending moments) and passive movements (if the leg's position is passively changed due to, for instance, a moving twig). Thus, Schmitz [28] demonstrated that cuticular stress was a controlled variable in the leg of the stick insect in order to determine the end of a stance and start of a swing phase, as well as in maintaining proper co-ordination of legs during walking.

In the locust, the trochanteral campaniform sensilla in the hind legs are used in order to assist co-contraction of muscles in preparation for a jump, but this is likely to be a specialisation [28].

Finally, it seems that the campaniform sensilla have a critical role in the generation of a normal walking pattern [2, 14].

3.4.3 Hairs and spines

Hair rows respond to touch stimuli and inform the central nervous system that a specific part of the leg has encountered an external object [3]. The hairs themselves are slender and can range in length from a fraction of a millimetre to about a centimetre [14]. The hair rows provide a continuous analogue representation of leg position which may be used in the targeting behaviour (see Section 7.1). Secondly, they contribute to the decision-making process involved in determining the end-points for swing and stance (see Section 6) by measuring joint angles [21].

Spines are also tactile receptors used in detection of the physical environment. Compared with hairs, they are usually much stouter and are intermediate in length. Spines tend to be located

towards the distal end of the femur and along the entire length of the tibia. Hairs tend to have greater freedom of movement than spines, which point at an acute angle to the leg, away from the body. For both hairs and spines, a single receptor is attached to the movable element at its base [14].

4 Insect behaviours

For biologists, the central issue in understanding insect behaviours is to understand how the nervous system can produce the correct muscle activations to achieve the observed behaviours [8]. For control engineers, the issues are similar with respect to appropriate algorithms and control techniques for 'copying' such behaviours in an artificial hexapod. The purpose of this section is to outline some of the higher-level observed behaviours in insects.

Graham [3] describes a hierarchy of behavioural control in insects, which can be mapped to the control architecture described in chapter 2 [29]. At the highest level, there is sensory input provided by the eyes and antennae which affect decisions on the whole body movement. At the next level, support, stabilisation and propulsion are provided by the individual legs coupled through a series of co-ordination mechanisms. At the lowest level, leg muscles are controlled in order to drive both standing and walking legs.

4.1 Height control

Adding a weight to an insect in no way changes the height of the body above the surface unless it is many times heavier than the normal load on the legs [30]. The distance between the substrate and the body of a stick insect is maintained in both standing and hanging animals. There are a number of hair fields between the leg joints that respond to angular position, angular velocity, and angular acceleration and control the overall range of movement to prevent damage by over-extension or over-flexion. These hair fields are involved in height control [3]. It has also been observed that the stick insect effectively keeps its body rigid when walking over obstacles, thus there is no change in the meso-metathoracic joint (see Fig. 2) of the animal (the most important intersegmental joint for height control) [12]. Bässler [2] describes the stick insect's height control system as follows: 'The forces are directed as if the animal were 'trying' to bring all the tarsi to the same level.'

A number of studies have examined the way in which a stick insect modulates its body height above the substrate [12, 31–33]. The problem is that when walking over uneven surfaces, each leg has to span a different distance between the body and the ground in order to maintain the rigid body above the ground. Some investigations have shown that the animal as a whole acts as a height controller with PD properties (position and velocity) [32]. A number of earlier models [12] were rejected in favour of each leg acting as an individual height controller, with no intersegmental neuronal connections [32]. Each leg, then, is modelled as a proportional controller with non-linear characteristics, the details of which have been reported in [33]. Cruse *et al.* [32] describe their model as follows:

> It was assumed that each leg acts as an elastic system that controls the distance between the body and surface independent of the other legs. This elastic system might be represented simply by the elasticity of the muscles of the leg or it might include the resistance reflex systems found in the standing animal to control the angles of the individual joints . . . Only mechanical connection between the legs was necessary to perform the height control . . . The system can be described as adopting that body position which requires the minimum total energy.

Having only a 'mechanical connection' between the legs implies that height control of the whole body does not require neural connections between the legs: the interaction between the

legs occurs through the substrate. Cruse *et al.* [32] also suggested that the distance between the tarsus and the coxa would be a better approximation to the controlled parameter rather than the vertical height between the coxa and the substrate. However, it cannot be the only parameter, since the perpendicular force component does not depend solely on this parameter.

There is no reduction in force on the legs of a stick insect towards the end of the swing phase, which would allow for a soft touchdown. There is thus no transfer of information between neighbouring legs with respect to an 'expected' ground height at the end of the swing movement, which could lead to a decrease in force before likely ground contact. This simplifies the neuronal controller: in fact, it may be totally unnecessary to have a force controller for an object of such little inertia – the stick insect leg weighs about 11 mg [15]. (This last assumption is obviously not true of the much larger artificial hexapod developed so far.)

4.2 Posture

A stick insect is capable of maintaining its body in a relatively stable equilibrium position. Insect legs tend to grip the substrate and this provides a stable support structure, even if the body's centre-of-mass is outside the polygon of support (the area enclosed by the legs where the feet are at the vertices). Hence, balance is generally not a control problem for the stick insect [2].

Rocking, which occurs when the animal is slightly disturbed, acts perpendicular to the body axis and aids posture. It leads to oscillations in the φ and ψ angles of the leg (Fig. 3). This behaviour is described in Section 7.5, because it is sometimes used as a strategy in negotiating rough terrain.

4.3 Orientation

A stick insect demonstrates negative geotaxis, which means that it will tend to climb vertically upwards. This temporarily switches to a positive geotaxis when the end of a branch is encountered. This behaviour is probably to find more favourable feeding sites. Precht [34] showed that a blinded stick insect will walk up a vertical branch. When it bifurcates, it will take one of the forks and climb to its end, then turn around and go up the other fork. It will do this several times before climbing back down the branch. Both sensory hairs and the femoral chordotonal organs have been shown to participate in the detection of the gravity vector during this behaviour [2].

The resting orientation of the stick insect assumes a resultant of the gravity vector (negative geotaxis) and the direction of the principal light source (positive phototaxis). Hence, light and gravity act as references for setting the walking direction or body position. When the turning due to negative geotaxis (D_S) exactly counterbalances the turning due to positive phototaxis (D_L), the stick insect is in equilibrium, that is, when:

$$D_S = -D_L. \tag{1}$$

The signals from proprioceptors on the subcoxal joints change only the sign (not the magnitude) of D_S. The weight of the body (G) is taken into account in calculating the components of the force. The sense hairs on the subcoxal joints measure the force in the direction of the longitudinal axis of the body. On the other hand, the sense organs of the femur-tibia joint measure forces along the transverse axis (i.e. the sensors that measure the $G\cos\alpha$ component). Removal of these sense organs shows that the value of D_S increases, hence, it is given by [2]:

$$D_S = K\frac{G\sin\alpha}{G\cos\alpha} = K\tan\alpha, \tag{2}$$

where K is a constant and α is the angle of the body with respect to the gravity vector.

The only stable position for the insect is then 0°, because although the function $y = \tan x$ crosses the x axis at both 0° and 180°, the head-down position for the insect (180°) has the opposite sign to the head-up position 0°. Temporary positive geotaxis is achieved by reversing the signs of the proprioceptive inputs from both the leg joints and the antennae as the insect turns [2].

4.4 Use of antennae

The antennae have a number of roles in stick insects. They are used to detect the height of the ground in front of the insect and provide information to the front legs as to when to end the swinging of the leg in forward walking. If the antennae strike an object, this can cause the insect to walk backwards. Finally, they are also used in the detection of the gravity vector [3].

4.5 Vision

Vision is not essential to the stick insect. It can navigate, walk, climb and feed without its eyes [3]. The eye of the stick insect *Carausius morosus* is highly adapted to dark conditions and they do not possess colour vision. However, they are able to register the degree of turning elicited by an external cue and to compensate for it when the external cue is removed [2].

4.6 Other stick insect behaviours

Stick insects exhibit a number of behaviours related to their habitat. Since they are nocturnal creatures, during daylight they normally show no spontaneous activity at all. They can remain so still for such long periods that such behaviour has been called 'death-feigning'. A second common behaviour is to camouflage themselves by 'twig mimesis'. This is achieved by aligning their legs parallel with the longitudinal axis of the body. A third behaviour, called 'catalepsy', is a state when a leg joint of an inactive stick insect is bent slowly by an experimenter and held at a new position. The joint will return to its original position at a rate of 0.1°/min which is five times slower than the hour hand on a clock [2].

5 Insect walking

5.1 Stopping and starting

When all the legs are touching the ground, a stick insect temporarily stops. When walking is discontinued, this leads to a rearrangement of the legs [3]. This resting position between walks is not maintained by elastic or friction properties of the leg muscles but as a result of the active use of particular muscles [2]. In observations of stick insects on a mercury surface, where the legs are mechanically decoupled through the substrate, it was noted that the insect was able to stand still rather than be in a continuous state of slipping. This implies that

> there must be an equilibrium position for the legs when standing, in which no reaction is expected and against which no continuously maintained force need be applied [35].

Once an insect starts to walk, it can take a number of steps to achieve correct co-ordination of walking [3]. If a standing animal is given a tactile stimulus in order to elicit walking, all the legs on the ground start to propel it with rearward-directed forces. One interpretation is that the neural

mechanisms in the legs always start in a power stroke when activated after a pause, at least in stick insects. Reports for other arthropods like the jumping spider suggest this is not a general result.

5.2 Gait terminology

Before presenting the observations of insect gaits as reported in the literature, it would be helpful to provide a series of related definitions. Delcomyn [36] mentions that the parameters of interest when considering insect walking are 'protraction, retraction, cycle duration, protraction/retraction ratios, and relative phase positions.' Graham [37] also described the 'lag' between steps.

5.2.1 Anterior and posterior extreme positions

The posterior extreme position (PEP) is the furthest back that a leg moves relative to a fixed point on the insect, say for instance the tip of the head, before it begins a swinging of the leg in the air, in order to begin another step.

The anterior extreme position (AEP) of a leg is the furthest forward to which a leg normally swings before it begins to make contact with the ground and push against it to propel the insect forward.

The AEP and PEP may vary from step to step (see [3] for summary). Typical values of the AEP and PEP for animals in different walking situations are given in Table 2.

5.2.2 Protraction and retraction

The protraction of a leg occurs when it reaches its PEP. It is also called the 'swing phase' of the leg or its 'recovery phase'.

The retraction of the leg occurs when the leg reaches the AEP after the swing phase. It is also known as the 'stance phase' or the 'support phase' of the leg.

5.2.3 Period of a step

A step period is the total time that a leg spends swinging followed by the total time it spends in stance phase, as measured from the onset of the swing phase [2]. The period is also called the 'cycle duration' or the 'duration of a step'.

Table 2: AEP and PEP measurements [11]. All measurements are in mm. The minimum sample size for any measurement is 63 animals.

	Walk on horizontal path	Walk on horizontal plane	Walk hanging from horizontal beam	Walk up vertical path
Foreleg				
AEP	11 (±3)	17 (±4)	14 (±3)	18 (±9)
PEP	−7 (±4)	2 (±8)	−5 (±3)	−7 (±5)
Middle leg				
AEP	−17 (±2)	−16 (±4)	−11 (±4)	−11 (±6)
PEP	−35 (±4)	−34 (±5)	−31 (±4)	−33 (±12)
Hind leg				
AEP	−40 (±3)	−39 (±3)	−34 (±4)	−34 (±12)
PEP	−58 (±4)	−58 (±4)	−52 (±5)	−56 (±15)

5.2.4 Lag

The lag can be defined between any two legs. It is the time from the onset of protraction in one leg, to the onset of protraction in the leg for which the lag is being defined [36]. For instance, the lag between the rear right leg (labelled 'R3') and the rear left leg (L3) is written as ${}_{R3}L_{L3}$ [37]. When referring to lags on the same side of an animal, for instance between the middle and hind legs, the symbol used is often reduced to ${}_2L_3$ [38].

5.2.5 Phase

The phase of two legs is the lag between the two legs divided by the period of the first leg. Unlike the lag, the phase is independent of period duration [2]. The phase of the middle left leg (L2) relative to the right middle leg (R2) is written L2:R2.

5.2.6 Gait

The gait of a walk refers to the temporal pattern of leg movements, not just the order of the leg movements [13]. Since the gait is dependent on the stepping pattern, there is a correlation between the gait and the forward speed of the insect. The general rule for a gait is that the lag across the body is relatively constant and is less than half the step period [3].

5.2.7 Duty factor and *p/r* ratio

The duty factor and p/r ratio has also been used in an attempt to quantify gait patterns [36, 39]. They are defined as follows. The duty factor is the fraction of the step period that the leg spends in contact with the ground, i.e. in retraction [3]. The p/r ratio is the protraction duration to retraction duration ratio. It turns out that this quantity is not useful in comparative studies of insect gaits because not all insects have constant protraction duration and the parameter does not give any information as to how one leg steps relative to another [13].

5.2.8 Step amplitude

The step amplitude of a leg is the distance between the AEP and PEP measured parallel to the longitudinal axis of the body, or along a straight line between the two points.

5.3 Gait observations

In 1966, Wilson published his foundational paper [30] on insect walking, in which he presents his observations of insect gaits. He attempted to make the observations as general as possible, and then to describe a general model to explain them. His observations are presented in Fig. 5. He hypothesised that each side of the insect produced a metachronal wave starting with the hind leg protraction and followed by the middle and front legs protracting in turn. The slowest gait he observed involved an alternation between a right-sided metachronal wave and a left-sided metachronal wave (Fig. 5a). As the insect moved faster, the left- and right-sided metachronal waves overlapped more and more until the typical tripod gait was observed, where not only do the left and right side waves overlap but one wave overlaps with its subsequent wave on one side.

This gait is shown in Fig. 5e. Wilson proposed a set of rules to explain these observations:

1. A wave of protractions (forward movements of the legs relative to the body) runs from posterior to anterior (and no leg protracts until the one behind is placed in a supporting position).
2. Contralateral legs of the same segment alternate in phase.
3. Protraction time is constant.
4. Frequency varies (retraction time decreases as frequency increases).

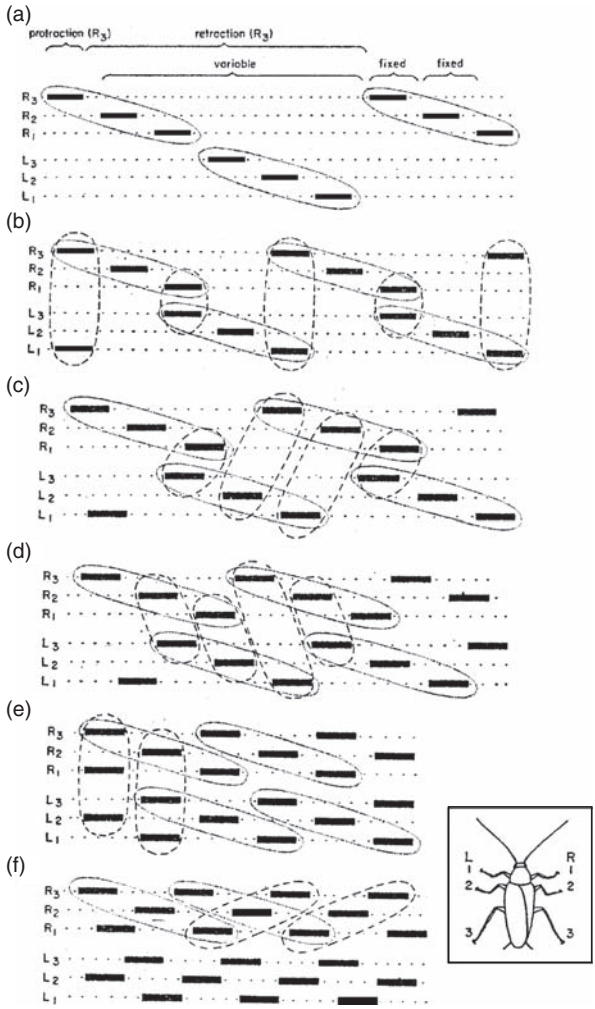


Figure 5: Biologically observed gaits. Solid bars represent protraction, i.e. when a leg is swinging forward. (a) Each leg steps alone in a 'metachronal wave' on each side of the body. (b–d) The basic patterns overlap more as the stepping frequency increases. (e) The tripod gait. (f) The highest frequency with leg 3 stepping before previous metachronal wave for the same side has been completed. Legs are labelled front to back as numbers 1–3, and the left and right sides are 'L' and 'R', respectively. Therefore, 'L2' refers to the middle left leg (see inset). Reprinted with permission from [30] (© 1966 Annual Reviews, www.annualreviews.org).

5. The intervals between steps of the hind leg and middle leg and between the middle leg and foreleg are constant, while the interval between the foreleg and hind leg steps varies inversely with frequency.

Wilson's rules enabled the description of gaits to be in terms of only one parameter, namely the retraction frequency. However, as Wilson himself knew, nearly every assumption is violated in one case or another. The most important deviation from the rules he presents is that in many

insects, including cockroaches, locusts and beetles, neither protraction nor ipsilateral lags are constant (rules 3 and 5, respectively), but instead vary as the period of the leg movements [13].

Nearly all species of insect adopt the alternating tripod gait (Fig. 5e), but no insect uses it to the total exclusion of other gaits. The tripod has been reported in 12 species of true insects (representing six different orders) and one Dipluran, and thus it is believed to be almost universal [13]. It is also clear that this gait is used at very long step periods (600–700 ms, equivalent to 1.6–1.4 steps per second). Sometimes, deviations from a strict phase of 0.5 can be observed for short periods in adjacent ipsilateral legs, but this is probably due to slight variations in the durations of swing and stance phases between the front, middle and hind legs [8]. Two insects, the Praying Mantis and the Water Strider, have never been observed using the tripod gait under any conditions [13].

In the cockroach, *Periplaneta americana*, the main observation is that the alternating tripod gait is used almost exclusively over the entire range of walking speeds (2–80 cm/s). Only at very low speeds is there any significant difference to this finding [36]. As for most insects, the leg movements in generating the alternating tripods are rarely synchronous, that is, the three legs of the tripod do not begin protraction exactly simultaneously. In particular, the lag ${}_3L_1$ is on average 0.954 for both right and left sides, indicating that the front legs begin swing phase slightly earlier than the hind legs. At very low speeds this ratio is reduced [36] which indicates a tetrapod-like gait similar to that of Fig. 5d. Phase is constant over nearly the entire range of locomotor speeds. Wilson's rules 3 and 5 are broken in the case of *Periplaneta* because the gait is unchanged over its range of walking speeds. Delcomyn [13] thus preferred to replace these two rules with another: 'No front or middle leg steps before the one behind it has finished its forward movement', which seems self-evident but he qualifies it by noting that high speed motion pictures of fast-moving cockroaches have shown that it can be violated. This implies that sensory input associated with leg placement cannot be essential for triggering protraction of the anterior leg.

A number of comparisons can be made between the locomotion observations of the cockroach *Periplaneta* and the stick insect *Carausius* [36]. As with *Periplaneta*, the retraction duration decreases as walking speed increases, but unlike in the cockroach, protraction duration is always constant in *Carausius*. As with the cockroach, contralateral phases (e.g. L2:R2) are independent of walking speed but ipsilateral phases exhibit drift, increasing as the stick insect's speed increases. Hence, in *Carausius* gait is a function of walking speed.

In the cockroach *Blatta*, gait is independent of speed for medium and fast walks, but is a continuous function of speed during very slow locomotion [36]. For all three of the insects, i.e. *Carausius* and the two cockroaches, the tripod gait occurs at speeds of approximately 7 cm/s [36]. *Carausius* is a much slower walking animal and only uses the tripod gait at its highest walking speeds, whereas the two cockroaches are 'running' animals, and so most of their walk is above 7 cm/s, hence the observations. Therefore, despite the observed differences concerning gait as a function of speed, there appears to be an underlying similarity, or locomotion mechanism, which is of course why Wilson could suggest 'general' rules in the first place.

The most common gaits of the first instar (an insect in the first stage of its development, in contrast to a fully mature or adult insect) stick insect *Carausius morosus* have been identified by the labels Gait I (the tripod gait, Fig. 5e) and Gait II (the tetrapod gait, Fig. 5d) [37]. In contrast to the tripod gait, the tetrapod gait swing phase duration is independent of the period. Further, the coupling between the left and right sides is not in strict antiphase but is very labile and given to irregular co-ordination [2]. The animal rarely maintains a straight course for very long when it uses the tetrapod gait. The two gaits can be clearly identified from a plot of lag between ipsilateral hind- and forelegs versus period (see Fig. 6). In the adult stick insect, walks are much more regular than the tetrapod gait of the first instar. Long, straight or slightly curved walks are frequent and the lag between ipsilateral hind- and forelegs can be given as a function of the period for both

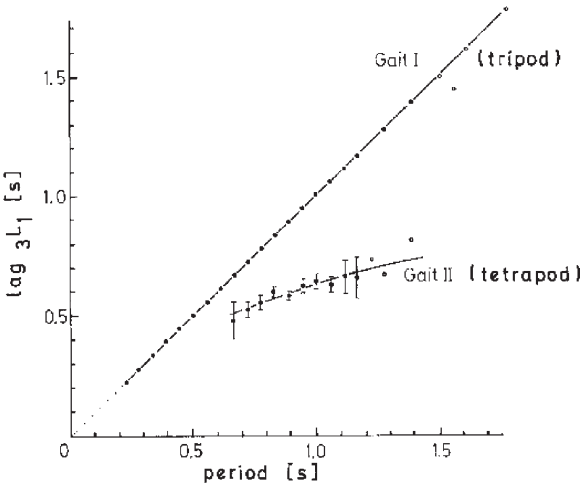


Figure 6: Metachronal lag ($_3L_1$) against period for 10 first instar stick insects. Closed circles show the mean value of $_3L_1$ for a given mean period with error bars if the value is greater than the symbol. From [40] (fig. 3, p. 190, © Springer-Verlag, 1977).

sides of the body. For slow walks, the phase R:L is about 0.35. The temporal asymmetry is almost always in this direction. For the tripod gait, right and left legs in the same segment have a phase of 0.5 [37].

Graham [38, 41] presents a number of observations of step patterns in the walking grasshopper, *Neoconocephalus robostus*. In this animal, the rear legs, which are twice as long as the front or middle legs, often stepped at half the frequency of the other leg pairs. Absences of rear leg protractions were found more frequently when the animal had been walking for some minutes. Unlike the stick insect or cockroach, the grasshopper exhibited lags between adjacent ipsilateral legs where $_3L_2$ is only about 80% of $_2L_1$, a phenomenon called 'relative co-ordination' by von Holst [42]. In most other insects (locusts, stick insects and cockroaches), these lags are approximately equal. The speed range of these animals is closer to that of stick insects and locusts rather than cockroaches. The gait used by this animal most closely resembles Gait II of the stick insect, except that the right and left sides of the grasshopper are locked into antiphase (R3 : L3 $\approx$ 0.5).

It can be shown that the gaits observed by Wilson [30] are part of a continuum that depends on speed. One insect, the silverfish *Petrobius brevistylis*, employs a unique gait which cannot be represented by such a continuum [43]. Its 'jumping gait' occurs when it moves each pair of legs in synchrony with pauses between each set of protractions. Consequently, the animal is supported by only a single pair of legs at a time. This jumping 'gallop' gives the insect a forward speed of 16–21 cm/s, which is more than two body lengths per second and more than twice the maximum speed of most other insects [8]. Sometimes, contralateral leg pairs can step in synchrony in the stick insect, for one or two steps. However, this is never the main form of locomotion and is usually a result of high loads on the legs [44].

5.4 Co-ordination

In order to establish the underlying co-ordination mechanisms that lead to the gaits described above, a number of experiments have been performed by insect physiologists on a variety of insects.

Some of these results will be reported here. They have been used to establish a number of models for co-ordination that are described in detail in chapter 4 in ref. [29]. The experiments involve altering the environment of the insect to see how this affects walking behaviour, amputating various limbs or parts of limbs to investigate how co-ordination is changed, and interrupting normal walking to examine how recovery occurs.

The legs in the stick insect are mechanically coupled through the substrate upon which it walks [22, 45]. All the legs on the ground move with the same velocity: a result of such mechanical coupling rather than a finely structured pattern of motor outputs. The spatial relationships of the legs are also partly determined by this mechanical coupling. As a result of this coupling, it is not possible to change the retraction speed of a single leg without affecting the speeds of the other supporting legs. In order to see whether this mechanical coupling is responsible for co-ordination, Graham and Cruse [35] observed stick insects walking on a mercury substrate which removes the mechanical coupling. The fundamental temporal co-ordination between the legs was unchanged as a result of the removal of the mechanical coupling. They were also able to show that feedback from the sense organs determines the step period.

The moment-by-moment irregularity of the forward velocity of stick insects and locusts show that the purpose of co-ordination is not to achieve smooth forward propulsion. Instead, the precision of co-ordination is aimed at providing adequate support for the body during the time when legs must be lifted and moved forward to propel the animal forward [41].

Graham [41] also pointed out that 'relative co-ordination' between adjacent legs on the same side of an insect is rare. Even if it is observed, there is still a strong tendency to avoid simultaneous protraction of two adjacent legs, which gives the insect an inherently stable walking system. The term 'relative co-ordination' refers to the behaviour of two weakly coupled oscillators with slightly different inherent frequencies. A preferred phase relationship is maintained for only a short time, after which the two frequencies exhibit 'beating' until they have a meta-stable in-phase relationship again [42].

It may seem surprising, but the exact neuronal mechanisms responsible for co-ordination have not yet been identified. However, several models of co-ordination which rely on influences between legs (passed via the nervous system) are described in chapter 4 in ref. [29].

5.5 Turning

Five possible mechanisms could be employed by an insect to turn. The first is to vary the step frequency on the right and left sides, while maintaining co-ordination. Second, a backward walk on one side at the same time as a forward walk on the other would result in an 'on-the-spot' turn. Third, some insects use the middle leg on one side to push against the substrate in order to thrust the insect into a different direction, especially during fast escape turns as in the cockroach [46–48]. Fourth, the two sides of the insect can become uncoupled with the legs on the outside of the curve walking with a higher frequency. Finally, while the co-ordination between right and left sides is maintained, an increase in the step amplitude for legs on the outside of the curve with a corresponding decrease for legs on the inside of the curve will result in curved walking [49].

Examples for the first and fourth possibilities have been found in stick insects [3], although Bässler [2] holds that turns are more commonly due to changes in step amplitude rather than step frequency (the last scenario). In moderate curves (that is with radius of curvature greater than 12.5 cm), the fourth possibility is seen [50]. When the curve is narrow, however, right and left legs become uncoupled and the legs on the outside of the curve walk with higher frequency, in both the stick insect [50] and the cockroach [46]. The inner middle and hind leg frequencies decrease relative to the outside leg frequencies, since they seem to be used to support the body. This turning

behaviour is built into the central nervous system and is adapted to the centre of gravity. It depends on force loading on the legs, mainly because stick insects are climbers [50].

During turning, the targeting behaviour (see Section 7.1) of the middle and hind legs is only maintained for the legs that are on the outside of the curve. The distance between the points of contact for the tarsi that walk on the inside of the curve are further apart than in straight walks even though they move along a narrower arc [50]. Three further observations were noted by Jander [50]:

1. The movement paths of the inner legs move close to one another and approach the body whilst the legs on the outside of the curve do exactly the opposite.
2. The length of the path covered by each inner leg is reduced.
3. The direction of both the power and return strokes changes sequentially from straight to curved.

A simple way to quantify the heading of an insect is to sum the angles between the coxa joint and the body (Cruse, personal communication).

5.6 Backward walking

In the stick insect, backward walking occurs when the antenna hits an object. During backward walking, the influence of the posterior legs on the protraction timing of the anterior legs is reversed but the walking is not as well co-ordinated [3].

A stick insect can be stimulated to walk backwards by pulling on its antennae gently [2, 51]. Such walks can be relatively long lasting and regular. During the stance, the leg was moved from rear to front and vice versa for swing, although steps were not as regular as for forward walking. Many 'reflexes' reverse their direction, for instance the searching reflex (see Section 7.2), as well as muscle activations [2]. It was clear from the experiments performed on backward walking insects, that the neuronal control of walking in the stick insect is not entirely forward–backward symmetric since Schmitz and Haßfield [51] were able to elicit a 'treading-on-tarsus' reflex (see Section 7.1) in the middle leg by stimulating a front leg during backwards walking. This serves no functional purpose.

6 The swing/stance phases

The results described in this section all refer to experiments carried out on stick insects.

6.1 Swing and stance as a two-state system

Bässler [45] shows that the movement of a tarsus during walking is in one of two states – either swing or stance – rather than in a relative phase of one with respect to the other. These states are separated by sharp transitions and the system can be modelled by a relaxation oscillator or bi-stable. Since tarsal contact normally signals the end of a swing phase, proprioceptors therefore influence the timing of walking. Sense organs can delay the step rhythm of a leg, for instance leaving out a step, or increase the step frequency. A leg terminates a particular phase of motor activity only when the organs signal completion of that phase.

In extreme cases, several actions can be observed during the swing phase: raise leg, pause, protraction, pause, swing down, tarsal contact. The stance phase can also consist of more than one component: a gripping phase when the tarsus makes ground contact between swing and stance and a releasing phase between stance and swing [2]. However, if there are any pauses in the

swing/stance cycle of a walking insect, they are most likely to occur at the transition points – the AEP between swing and stance, and the PEP between stance and swing [3].

Co-ordinating mechanisms between legs are known that determine the transitions from power to return stroke and from return stroke to power stroke [49].

6.2 Velocity

Figure 7 shows the positions of the leg tarsi for normal walking in an intact animal. The open circles indicate that the swing phase has an approximately constant speed, and its duration is between 100 and 200 ms [21, 37, 52] and is independent of walking speed. This constancy is observed even in insects climbing upwards, but an increase in stance phase duration is observed in climbing. In downwards vertical walks, protraction activity takes up most of the step period [10].

Speed of locomotion is directly related to the force generated by the retraction stroke [3] and the average speed of leg retraction is equal to the forward speed of locomotion. Velocity is fairly constant throughout protraction as well as retraction. The velocity profile for all the legs is similar, but there is no clear uniformity over successful cycles [35].

6.3 Factors affecting the stance phase

The motor neurones which control the force which propels the body during stance depends on the position and loading on the leg, upon co-ordinating influences and upon velocity, and probably acceleration [53].

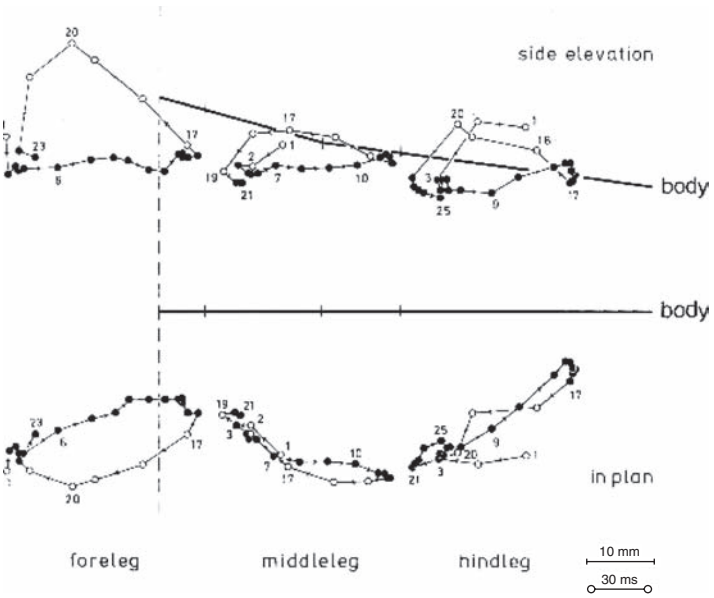


Figure 7: The tarsi movements in the stick insect *Carausius morosus* relative to body position for an insect walking on a horizontal plane. The closed circles denote stance phase. The dashed line represents the tip of the stick insect head and the tick marks are the positions of the coxae. From [11] (fig.5, p. 242, © Springer-Verlag, 1976).

At the beginning of stance phase, high activity in the protractor muscles show that the first stage of retraction is passive. Hence, the leg is moved backwards due to mechanical coupling with the other legs. There is even a resistance opposing this passive movement. The high protractor muscle activity may act as the resistance reflex, or it may be used to help the tarsus to grasp the substrate following the swing [2].

Velocity is the controlled variable during stance phase where the femur-tibia reflex acts a velocity-transducer. There is also evidence that the control system is affected by an acceleration-sensitive mechanism [54]. Acceleration sensors have been located in the femoral chordotonal organ of the stick insect by Hofmann and Koch [25].

Position is an important parameter for the decision to finish the stance phase, but is not used as the controlled variable during stance. Load is used for the decision to finish stance and seems to be used in the control of the legs during stance [55]. If a leg is under too much load (as registered by the trochanteral campaniform sensilla) then the leg will not lift off the ground. The PEP shifts forward under large loads because the load on the leg muscles prevents the continuation of the stance phase. In contrast, under small to medium loads, the muscle forces can be enough for retraction to continue past the usual PEP [44]. Dean [44] also showed that the decision to end stance does not simply depend on the leg's loading but on its efficiency in contributing to propulsion. Hence, position (relative to the PEP as indicated by sense organs) and load decrease are two important factors in determining the end of stance phase. In addition, there are a number of other co-ordinating influences that may directly or indirectly have an affect on the transition point between stance and swing [53].

During stance phase, the middle leg is oriented more nearly perpendicular to the longitudinal axis of the body than either the front or hind legs [18].

6.4 Factors affecting the swing phase

The timing of protraction of a leg is influenced by the position of the leg in front, that is, a leg will swing when the tarsus in front is in the range of the AEP. For hind and middle legs, the end of the swing phase is determined by geometrical parameters (i.e. the position of the tarsus) because of the targeting behaviour (described in Section 7.1). Clearly, the geometry of the legs cannot be the only factor, since the foreleg does not aim at a leg in front of it [53].

Dean [56] showed that the leg movement of the stick insect is also velocity-controlled during the swing phase. Since the swing phase is of constant duration independent of walking speed, this implies that a leg begins to swing while its anterior neighbour is farther forward in the case of faster walking [21]. The results of Weiland and Koch [57] confirm both that velocity is the controlled parameter during swing and that position is a key factor in ending the swing phase.

The swing phase is affected by load variations in much the same way as the stance phase, indicated by changes in the AEP of a similar amount to the changes mentioned in the PEP (probably as a consequence of the targeting mechanism). Under higher loads, the AEP shifts forward. In general, an increase in loading causes the swing phase duration to decrease relative to the stance phase for a given step period, thus extending the time that the leg contributes to forward propulsion. 'This is true for loads of different kinds, including loading parallel or perpendicular to the body axis, weight added to the body, and frictional and inertial loads' [44]. When loading assists propulsion, the swing phase duration increases but only slightly. Small to moderate assisting loads led to an unexpected increase in velocity possibly as a result of an increase in muscle activity needed to brake at the end of stance. For large assisting loads, swing duration increased, amplitude changed little and velocity decreased [44].

The front leg swings higher than the middle, which swings higher than the hind leg [15]. This is probably due to the fact that if the front leg does not hit an obstacle the middle and hind legs can afford to reduce their highest point of swing with respect to that of the front leg, thus saving energy.

7 Rough terrain strategies

Insects use many strategies for coping with rough terrain, and environments with sparse footholds. A number of these strategies will be presented here, and it is intended that some will be emulated on the hexapod.

Pearson and Franklin [58] claimed that there were no general concepts for describing how insects adapt their walking to rough terrain. In many cases, there are not even any detailed descriptions of what insects do when they walk on rough terrain. They investigated the locust *Locusta migratoria*, and discovered that it does not adopt a fundamentally different gait when walking on rough terrain, as compared with flat surface walks. However, each leg seems to act independently in finding a support site, while preserving the basic intrasegmental anti-phase between contralateral legs, as well as the posterior-to-anterior metachronal sequence. In one or two situations they observed simultaneous stepping of adjacent ipsilateral legs when negotiating a ditch but only when the supporting legs ensured stability.

A useful definition for quantifying a hexapod's response to rough terrain has been given by McGhee and Iswandhi [59]. They describe the 'reachable area of a leg' as being the sector of an annulus whose base point is the point of articulation of a given leg with the body. If enough of the ground within the reachable area of a leg is forbidden due to lack of a firm foothold then there is an obstacle that needs to be surmounted in order to establish a firm point of contact for the tarsus. If the foothold is too elastic, in that there may be some slippage between the foot and the uneven substrate, then this will affect the leg's load-bearing capacity [60].

7.1 Targeting of foot placements

When coming to the end of a swing phase, the middle and hind leg tarsi of the stick insect home in on the PEP of the standing tarsus of the leg in front. The actual foot placement is usually a few millimetres behind and outside the tarsus of the anterior leg [61]. In the case where an insect was made to walk over a 15 mm wide ditch, the front legs would step into the ditch with probability 65%. However, the middle legs stepped into it with a 48% probability and the hind legs 25% of the time [61]. This 'targeting' behaviour is clearly useful where there are sparse footholds, and little sensory information is available about the substrate, for instance from the vision system. This type of foot placement is sometimes called a 'follow-the-leader' or FTL gait [58]. As well as existing between hind and middle ipsilateral legs, and between the middle and front legs, similar targeting information is passed between the antennae and the front legs [49].

The type of substrate affects the precision of the targeting behaviour. A flat, solid surface decreases the need for precise targeting, whereas an uneven surface, such as a wire mesh, greatly increases it [27].

Assuming that vision plays no significant role in selection of individual footholds, the use of targeting behaviour implies that there is a functional specialisation of the forelegs (and possibly middle legs) for this task [60]. The use of the forelegs as sensors in this way has been suggested by several studies [3, 8, 58]. Further, a transfer of information regarding foothold position from the

anterior leg to its posterior neighbour is required, and has been clearly demonstrated in the stick insect [61]. It is likely that this information is not bidirectional, i.e. it is not similarly transferred from a posterior leg to its anterior leg in backward walking. There is some suggestion of this in [51].

Cruse [61] found that information as to leg placement must be passed from one leg to the one behind since the actual foot positions relative to the body between, say, one middle leg step and the next, might vary and yet the hind leg would on average step at a similar point relative to it. It is interesting to consider how this might occur. It is highly unlikely that the stick insect 'records' its leg end-points in Cartesian co-ordinates and subsequently performs two co-ordinate transforms: the direct kinematic calculation to work out the position for the tarsus of the target foot, then the inverse kinematic transform into the joint angles of the anterior foot.

As an attempt to explain what alternative mechanisms may exists for the targeting behaviour, Dean [62] modelled the targeting information in terms of an artificial neural network. He was able to show that it is straightforward with a simple network to associate the joint angles of the middle legs with joint angles of the hind legs in such a way that the targeting is performed, and thus no co-ordinate transform is necessary. There were large errors for the margins of the test field, however. Further, there is no known source in the stick insect for the required precision of error information to perform targeting, whereas the source of the targeting behaviour itself has been identified with sense organs in the anterior leg [51]. However, the accuracy of the targeting behaviour does vary depending on the substrate and this could reflect a learning mechanism that continues to modulate neural connections for its control. Perhaps, the error information is provided by the 'treading-on-tarsus' (TOT) reflex.

The TOT reflex occurs when a leg accidentally steps on the leg in front when performing the targeting behaviour. It has been investigated thoroughly by Schmitz and Haßfield [51]. If the hind leg steps onto the tarsus of the middle leg, the hind leg often lifts off again and is repositioned slightly posteriorly without any interruption to the walking except for a prolonged step when the TOT reflex occurs. Its occurrence depends on a number of factors including the phase between adjacent ipsilateral leg pairs and the position of the posterior leg relative to its transition points (AEP and PEP). However, the reflex does not seem to depend upon whether the anterior leg is in contact with the ground nor whether the animal was actually walking. (It can be stimulated by brushing the top of the anterior leg tarsus.)

It is not known whether the 'targeting' strategy is employed by other insects. Even if it were, it could not be the only strategy to be used for rough terrain walking since the front legs need to find support sites. The behaviour was certainly not visible in locusts [60]. Instead, Pearson and Franklin [58] noted that individual legs have the capacity for finding support sites without input from the eyes or the antennae. Unlike with the stick insect targeting behaviour, there seemed to be no evidence that information passed between legs about support sites. Their conclusion was that this is due to the less varied surfaces in the locust habitat (e.g. leaves or blades of grass) as compared to the intricate branches and twigs negotiated by the stick insect. Also, the locust can jump from one location to another.

7.2 Searching reflex

If a leg comes to the end of its swing phase and fails to make ground contact, it is common for the leg to begin a 'whole leg search'. The reason for the failure to make ground contact is that the leg could be stepping into a hole, or the substrate slips away during the swing phase. This reflex has been observed during locust walking in all legs [58, 60]; and in the stick insect it primarily (but not exclusively) occurs in the front legs [2, 61].

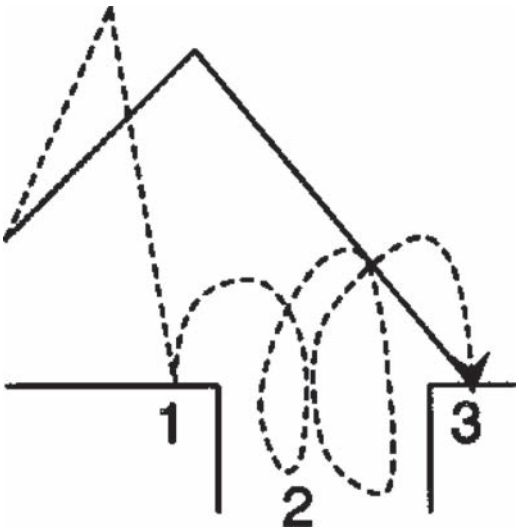


Figure 8: The searching reflex. 1 is the original foothold before the leg swings, 2 is the expected foothold which is not found and so a searching reflex is initiated, and 3 is the secured foothold. (Reprinted from [63], © 1996, with permission from Elsevier Science.)

The most detailed description of the searching reflex is given in [58] for the locust, as follows:

1. Rapid elevation and depression movements of the leg at frequencies of up to 8 cycles/second, resulting in the exploration of a wide space around the point of articulation of the leg with the body.
2. Usually, marked extension at distal joints effectively increasing the range of the search relative to the body.
3. Continuation for many cycles (the most we have observed is eight).
4. Termination when the animal stopped walking or the leg struck and found support on an object.

Pearson and Franklin [58] also found that searching movements caused the locust to pause during walking, or in extreme cases to stop altogether. This was also the case in stick insects [64]. Searching movements were not limited to walking over rough terrain. In standing insects, they have been observed during postural adjustments (but these were much slower at about 1–2 Hz) [58]. Figure 8 shows the main features of the searching reflex.

The precise information for triggering a searching reflex is not known, but it is probably some combination of leg position relative to the body and lack of tarsal contact with the substrate [60]. In the stick insect, it was found that at the end of a swing, the depressor muscle motor neurones' activity is briefly interrupted followed by activity in the levator muscle motor neurones for a shorter duration [65]. At the same time, the protractor is briefly activated [11, 66]. The case is the same for the end of searching, when the leg grasps an object [67], while touching the ground causes other motor neurone activity [68]. For a review of these muscle and motor neurone activities, see [65]. In summary, sense organs and muscle activity seems to be monitored to indicate ground contact, and when it does not occur as expected the searching reflex is elicited.

7.3 Elevator reflex

If a leg contacts an obstacle during the swing phase, a 'motor program' termed 'the elevator reflex' is immediately initiated [60]. On hitting the obstacle, the leg is not simply set down. In locusts, the leg is lifted, swung higher and set down on top of the object [58]. A similar behaviour exists in stick insects where the swing of the leg is sometimes reversed before swinging it up on top of the obstacle [2, 22]. This whole sequence can be repeated with a frequency of 3–4 Hz until the obstacle is surmounted, providing the obstacle is not too high [2]. Where the obstacle is very narrow, the effect is to swing the leg over the obstacle to a foothold on the other side over the obstacle rather than to set the limb on top of the object. The elevator has been observed in all three pairs of legs, but is most often seen in the front and middle legs. It has also been observed during searching movements but usually led to the termination of the search [58]. The elevator reflex is shown in Fig. 9.

7.4 Local searching

In addition to the whole leg search and the elevator reflex, individual legs in the locust were also seen to perform what has been called 'local searching' [58], also known as the 'stepping reflex' [63]. This movement, which occurs once a secure foothold has been located, involves repeated small stepping movements by the tarsus near the location of the initial foothold.

> Local searching was most obvious when the potential support surface was smooth and the leg action required was to propel the animal forward . . . Thus the probability of local searching is related to the smoothness of the surface. We have not determined what information is used to signal a suitable support site and terminate the local search, but we hypothesise it is a signal related to the load carried by the leg . . . This process continues until the critical load is borne by the leg [58].

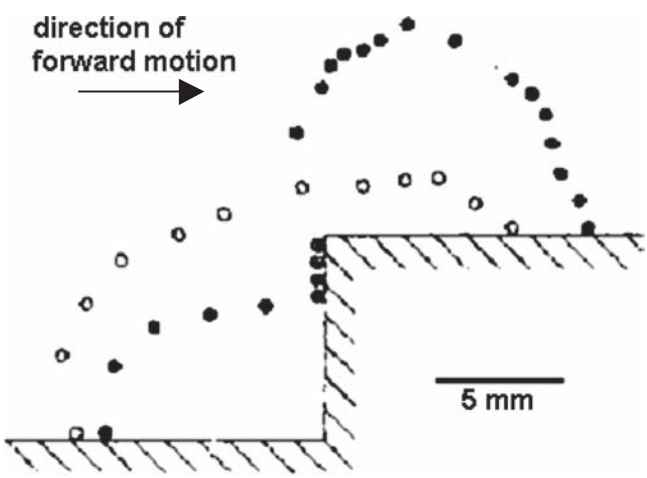


Figure 9: The elevator reflex of the left middle leg of the locust showing a normal step (open circles) and the elevator reflex (closed circles). The circles represent successive tarsi positions taken frame by frame (at 50 frames/s) [60].

Franklin [60] noted that the local search often terminated the whole leg search, and that in the locust, the whole leg search, elevator reflex and local searching reflex, 'while often occurring in sequence, also occurred independently'.

No such mechanism has been reported by observers of the stick insect. However, this is likely to be because the stick insect grasps the substrate using sticky pads and claws of the tarsus which provide forces to counteract gravity [3]. Bässler *et al.* [67] related this action to the end of searching where changes in leg muscle loading were responsible for ending the search and causing the tarsus to grasp the object.

7.5 Swaying and stepping

'Swaying' [69] or 'rocking' [2, 70] is a behaviour employed by insects for coping with mechanical perturbations of the legs. In small perturbations, it works by opposing the change of angle in the perturbed joint. It occurs perpendicular to the long axis of the body. Rocking leads to oscillation of the coxa-femur, femur-tibia and tibia-tarsus angles. A central oscillator has been shown to exist in stick insects which has influence over rocking, and to modulate appropriate motor neurons of all six legs [70]. Rocking can be stimulated in a small number of cases (around 1%) by stimulation of a tarsus [71]. Grasshoppers use swaying strategies predominantly in the middle leg, but also often in the hind leg in response to movements of the substrate. Despite the fact that the leg muscles are located at various angles, during swaying they all work to decrease the joint angles, thus pulling the insect towards the substrate as it was pulled away. Hence, one factor that is controlled during swaying reactions is the rate of change of joint angle [69]. Zill [69] also found that a cockroach is able to counter perturbations within a single step cycle using swaying, at slow to moderate walking speeds.

Thus, if the perturbations are small, an insect can compensate by swaying. However, for larger perturbations, the insect leg is lifted and moved away from the perturbation in order to ensure that the insect is balanced and that the leg is again capable of supporting load. This reaction is called a 'stepping response' [69]. The stepping response is most often used in grasshoppers when the femorotibial joint is close to its full flexion. As the displacement occurred, the leg is lifted from the substrate and moved, most often to full flexion of the femorotibial joint. Zill [69] suggested the grasshopper uses the following two rules for compensatory reactions to perturbations:

1. [Grasshoppers] use swaying strategies to hold the leg in place at joint angles that are stable and in which muscle tensions can be effectively developed to counter perturbations.
2. They use stepping strategies to move the leg out of ranges of joint angle that are unstable and in which mechanical advantages of muscles are low.

Zill [69] also noted that cockroaches use stepping strategies when a perturbation is applied late in the stance phase of a leg in moderately fast walking animals: an extra small step is taken but the overall rhythm of walking is unchanged. At higher speeds, the perturbations were considerably more disruptive. Interestingly, the same paper shows that humans also use these same two strategies in response to certain perturbations.

Related behaviours to the swaying and stepping response have been described for the stick insect by Kittmann *et al.* [71]. These behaviours include (a) a rapid leg withdrawal in response to a tarsal stimulus, (b) compensatory leg placement where, following the stimulation of a standing leg, a lifted leg is lowered to support the body; and (c) the leg set down reflex, which ensures that balance is maintained by shifting the loads on standing legs.

7.6 Avoiding obstacles and the use of vision

The use of vision by insects for negotiating rough terrain is limited. In locusts, vision was clearly used to stop the animal walking when approaching changes in terrain. However, this sense was not used always, even if it would have been advantageous to do so: while walking up a vertical rod, the front legs were used almost exclusively to locate lateral projections following elevator reflexes. There is no indication that locusts use vision for directing foreleg movements (for instance to visible footholds). However, it does seem that vision changes the animal's orientation with respect to an object within the environment [58]. Graham [3] noted that the influence of sensory inputs from the head of the stick insect (its eyes and antennae) is primarily inhibitory, i.e. without a head, the animal would walk continuously.

Franklin [60] showed that both the locust and the eight-legged opilonid were able to negotiate rough ground without visual sense, finding footholds by touch. He reasoned that this is because the use of

> visual determination requires high visual acuity and sophisticated inference, where testing the foothold by touch is more simple, direct and results in an essential, final answer.

7.7 Negotiating steps, ditches and barriers

Experiments conducted by Pearson and Franklin [58] on the locust and Cruse [11] with the stick insect show that insects can readily negotiate a variety of rough terrain environments. For instance, locusts have been observed walking over a wire mesh, an irregular surface of wooden blocks where the height differential between neighbouring blocks could be as much as 10% of the body length; a hexagonal surface of flat head nails (with only 40% of the walking surface solid and minimum gaps between nails of 10% the body length); a ditch width up to 16% of the body length; steps of height approximately 20% of the body length and a vertical rod with projecting side branches [58]. Similarly, the stick insect has been observed negotiating ditches, barriers and steps, as well as upside down on a horizontal beam, and straight up a vertical path [11].

Pearson and Franklin [58] describe a series of observations of locusts walking over ditches and up steps. In the case of ditches which were not too wide or for steps that were not too high, the insects walked over them with little change in co-ordination or speed. The compensations for these obstacles were typically one cycle of a whole leg search in the case of ditches, or a single elevator reflex in the case of steps. For bigger obstacles (ditches and steps of magnitude greater than 1 cm wide), a stereotyped strategy was employed: first, detection of the obstacle either visually or because the antennae or forelegs touched the obstacle; second, posture adjustment to ensure a symmetric stance; third, the forelegs established support sites on the opposite side of the ditch following extensive searching or on top of the step following an elevator reflex; fourth, *both* middle legs stepped simultaneously over the ditch or onto the step; finally, the animal continued the walk with a similar gait to that observed before the obstacle. A key component of this strategy was the stable tetrapod of support provided by the hind legs and forelegs when both middle legs were stepping: if this support was unavailable, the simultaneous stepping of legs was never seen [58]. In rare cases, a complete sequence of hind-middle-front in-phase stepping of legs was observed when crossing ditches [60].

Similar observations have been recorded for stick insects walking on similar terrain. Bässler [2] describes an important aspect of the stick insect in these circumstances: except in cases where the obstacles are very high and the insect is effectively making a horizontal-to-vertical transition, the body remains virtually rigid (in particular the meso-metathoracic joint) when negotiating

an obstacle. Normal body posture with the prothorax highest and the metathorax lowest is maintained even on the obstacle. Further, the longitudinal axis of the body forms approximately the same angle with the average substrate slope as it would form with a flat surface. Thus, Bässler [2] concludes that the same height control model described in Section 4.1 can be used for rough terrain as for flat terrain. Some quantitative observations are provided by Graham [3] concerning stick insects walking over obstacles. He describes three situations. When obstacles are up to 10% of the body height, the obstacle may interfere with foot placements, and the insect sometimes waits briefly to see if the tarsus slides or rolls. For obstacles that range between 10% and 100% of body height, the insect tends to modify its support position and to step close to the foreleg footprints. Larger obstacles are climbed over or avoided.

8 Compliance

Most artificial systems are still engineered with rigidity and accurate tolerances. This 'tradition' may have emerged predominantly because such structures minimise 'hard-to-control' characteristics given classical approaches, e.g. the non-linear dynamics introduced by passive compliance. It seems in contrast that natural systems cope with many non-linearities effortlessly (think for instance of the dynamic complexity of our own arms), which implies that these classical approaches could be misguided for applications like walking systems. Biological systems are passively compliant (reducing 'wear and tear' and the likelihood of excessive forces on joints in fault conditions etc.).

One suggestion is that such passive compliance is the result of the evolutionary pressure to minimise energy consumption during locomotion [72]. This energy minimisation could be achieved by alternating kinetic energy and potential energy during the course of a step thus ensuring that hardly any of the kinetic energy required to power the legs degrades to heat loss. Alexander [72] calls this the 'pogo-stick principle' and shows that it is used by humans, kangaroos and quadrupeds. Because stepping is periodic and legs are required to move backwards and forwards, the accelerating and decelerating forces will cause internal losses of kinetic energy to heat if inelastic actuators are driving the legs. However, animals use spring-like tendons to reduce such losses.

There is a third way in which animals use 'springs' during locomotion. A rigid foot landing on a rigid surface would experience a large force on impact, so animal feet typically have passive compliant properties. It is possible to model these properties as an elastic pad with rate-independent damping (i.e. it returns in its elastic recoil a constant fraction of the work done in deforming it). However, ideally, a foot should have non-linear elastic properties. It should be soft under low loads and become stiffer as load increases. Pads made of rubber-like polymers behave like this in compression, as do mammals' paws. If feet have these properties, as well as being light but with some mass, then it is likely that 'chattering' will be eliminated. This is the situation where the spring constants in the leg are such that for certain critical values of the ground contact force, there results an oscillation of the foot which occurs at the point of contact [72].

9 Dynamic considerations

The dynamics of insect standing and walking are rather complex. A number of different forces form part of the dynamical picture at different times: inertial and frictional forces due to velocity changes, gravitational forces dependent only upon the orientation and direction of motion, and spring-like forces dependent upon some displacement parameter. The effect of gravity is probably

the most important consideration in the control of walking, from the insect's point of view. Force components perpendicular and torque components parallel to the gravity vector affect the need to maintain a sufficient height from the substrate and cause the supporting legs to act as struts. During locomotion, forces parallel to the gravity vector will affect the propulsion of the body and thus cause the legs in contact with the substrate to act as levers [44].

Thus, the simple characterisation of a leg's motion in terms of swing and stance assumes that the kinetics (time developments of the forces) of all legs in stance phase is equivalent but this is not the case because it depends on what the leg is doing at the time. Many of the studies of insect walking and the corresponding models of leg function have not considered the passive dynamic properties of the insect musculoskeletal system. Also, many of the studies have been conducted on slowly moving animals where the dynamic effects are likely to be less noticeable [73].

There are essentially two problems in considering the dynamics of walking insects. First, in trying to relate neural signals to musculoskeletal responses, there are a series of difficulties. Often the same neural output will result in several different behaviours. Furthermore, when two or more legs are on the ground, the forward dynamics are indeterminate because of the closed kinematic chains involved. Secondly, very few biological investigations have correlated motor neurone output, muscle activity and kinematics with the actual kinetics [74].

9.1 Force measurements

Cruse [11] measured the forces on the legs of stick insects walking on a flat surface, a horizontal bar, vertically and upside down hanging from a horizontal beam. Some of these measurements are shown in Fig. 10. He showed that the timing and amount of force exerted by a given leg depends on the walking situation. Their magnitude is on the order of the body weight of the insect. When walking upright, legs always produce downward forces in contact with the ground, and middle

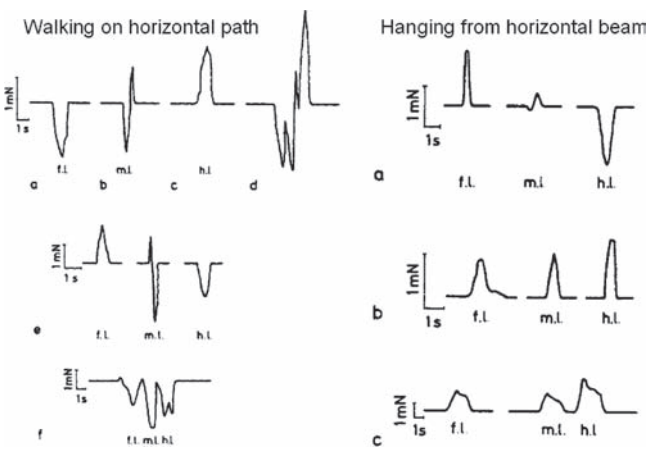


Figure 10: The time course of forces in the legs of the stick insect *Carausius morosus* when walking on a horizontal path (left) and when walking upside down hanging from a horizontal beam (right). Top row: longitudinal forces (backwards = positive), middle row: transverse forces (against the body = positive), and bottom row: the vertical axis (downward = positive); f.l. = front leg, m.l. = middle leg, h.l. = hind leg. Graph d shows all three legs stepping on the force platform. From [11] (fig. 6, p. 244, fig. 8, p. 247, © Springer-Verlag, 1976).

and hind legs provide most of the support for the body. The horizontal component of the force perpendicular to the body acts outward in the middle and hind legs, but is roughly zero for the front legs. The force measurements along the body axis, sometimes show a negative or forward-directed component. In front legs, the forces act positively towards the rear but in the hind and middle legs, applied forces oppose the direction of motion in the first half of the stance phases for these legs.

In stick insects, the front and middle legs normally provide the greatest propulsion. Sometimes hind legs develop forces that thwart forward propulsion, probably to ensure body support. However, on the whole, hind legs provide a small amount of propulsion. Either the front or hind legs provide a large decelerating force [54]. A not too dissimilar profile of forces and leg function has been observed in the cockroach. For example, in the tripod gait, the front leg decelerates the centre-of-mass in the horizontal direction, whereas the hind leg is used to accelerate the body. The middle leg does both, much like legs in bipedal runners and quadrupedal trotters. Vertical force peaks for each leg are equal in magnitude and significant lateral forces are present [73].

Cruse [75] describes how forces between neighbouring legs in walking stick insects vary under different situations. If the stance movement of a leg is interrupted, the forces in the other legs increase in order to increase the total force propelling the body. This effect was observed between all neighbouring legs except the two hind legs. Similar co-activating influences were found to act between legs when the load on the insect was increased.

9.2 Force and velocity observations

The maximum velocity is achieved close to the midpoint of the stance phase. At the end of the stance, the body or leg movement is almost stopped. These large accelerations and decelerations which occur twice in each leg cycle result in jerky lurching movements which are easily observed during the forward horizontal walk of stick insects. This results in a waste of kinetic energy, and represents a fundamental difference between insects and quadrupeds. Graham [3] suggests this mechanism is used for climbing.

Increased horizontal or propulsive loads can slow down leg movement during the power stroke because of the mechanical influences of the system. This can explain why return stroke duration was constant for all speeds in some experiments, whereas in others it was correlated with speed. Graham and Cruse [35] describe a possible muscle excitation mechanism whereby an increase in the load produces a strong excitation of the muscles involved in the return stroke. Their model predicts that for small loads, return stroke duration decreases with increasing speed, whereas it is short and constant for high loads.

Delcomyn [74] describes the effect of increased loads on one side of an insect. The effect is that legs on both sides of the body slow down and remain synchronised. However, with an increase in the load differential, the legs begin to step at different frequencies: the less heavily loaded legs stepping twice for every step of the others. In no case did the leg frequencies occur totally independently of each other (i.e. out of synchrony).

The magnitude of a frictional load varies with velocity. Hence, the frictional loads exert a larger effect during velocity peaks, which occur in the middle of the stance phase, and a smaller effect during the minima which are at either end of the stance phase [44].

9.3 Load-carrying capacity

Stick insects can carry up to four times their own body weight [3]. Pfeiffer and Cruse [76] studied the requirements for maximising load-bearing capability in a hexapod. The fact that the

coxa-trochanteris carries the greatest load, the geometry of the stick insect, limitations of driving moments, centre of mass, foot positions and cycle time were entered as inequality constraints as part of the optimisation criterion. As a result of this optimisation study, the Technical University of Münich built a hexapod with some of the following specifications for load-bearing capacity. The ratio of maximum moments in the leg joints $\tau_\theta : \tau_\varphi : \tau_\psi$ is 2 : 3.6 : 1 and reflects almost exactly the average value of the ratio of moments in the three legs of one side of the stick insect. The hexapod has a load-carrying capacity of 3.5–4 times its own weight over a long duration when friction is considered. When the load is applied for only a very short time, a maximum possible load-bearing capacity of 6 times the body weight is possible.

9.4 Effect of load changes

Increased load on the legs tend to shift the step patterns of an alternating tripod to a tetrapod gait or metachronal gait. Underlying this change in gait is the result that increased inertial and frictional loads change ipsilateral step co-ordination by shortening the lag for a given period. The change from a tripod gait puts more legs in contact with the ground at any one time and therefore increases the propulsive force [44].

The performance of the neuromuscular system is little affected by load changes corresponding to different orientations relative to gravity. This is evidenced by a more or less constant step pattern. It seems from the experiments of Dean [44] that the control system does attempt to maintain a set velocity during stance.

9.5 Motion of the centre-of-mass

The motion of the whole body cannot easily be predicted even if a good dynamical model of the legs is known. Full [73] maintains that despite the numerous studies in the kinetics of animal leg movements which demonstrate massive variation in phase and other leg behaviours, his results have consistently shown a remarkably regular movement of the body's centre-of-mass. Even in multi-legged arthropods, the motion of the body centre is not 'wheel-like' (i.e. at a constant height throughout the gait period) but more often demonstrates antiphase between the gravitational potential energy and kinetic energy, like an inverted pendulum or spring-mass system. In fast cockroach walks, the observed forces have been described by Full [73] as follows:

> Distinct maxima and minima in the whole ground reaction forces are apparent. Each vertical force peak is correlated with a step of an alternating set of legs. As the animal's body comes down on a tripod, it decelerates in the horizontal direction. Its vertical force increases above body weight. As the body lifts up, it is accelerated and the vertical force decreases below body weight. This pattern is repeated for the next step of the tripod. In contrast to a pendulum-like walking gait, potential energy and kinetic energy fluctuate in phase during cockroach locomotion.

9.6 Static versus dynamic stability

One advantage that hexapods have is that they can always be statically stable. This means that there can always be a tripod of support found during walking. However, contrary to the earlier assumption by Hughes [77] that this benefit is the result of the 'end-product of evolution', Full [73] has shown that dynamic stability is crucial even in small multi-legged agents. A number of researchers in Full's group have demonstrated that fast animal locomotion requires dynamic stability in the following series of observations [73]. First, running and bouncing gaits in crabs

and cockroaches are dynamically similar to quadrupedal trotting or biped running. Second, the cockroach *Periplaneta americana* can run at high speeds quadrupedally and bipedally. Third, cockroaches with their middle legs removed can run with a duty factor of less than 0.75 without falling. Fourth, the stability margin (i.e. the distance from the centre of mass to the polygon of support) decreases linearly with speed and becomes negative at the lowest speeds (i.e. statically unstable). The stability margin is essentially related to momentum, because as the animal attempts to stop, it 'tosses' itself into the tripod gait as a result of its forward momentum. Finally, the resultant of ground reaction forces during a single time-slice of the runs of a cockroach generate moments such that the animal would flip over.

The stick insect is a much slower animal than the cockroach. In modelling a robot on the stick insect, it has been considered 'safe' to assume that a statically stable model is sufficient for the purposes of this research. However, it is relatively straightforward to introduce dynamically stable walking into the dynamical model of the hexapod presented in chapter 6 in ref. [29].

10 Biological principles for hexapod design

Delcomyn [8] provides some tips to engineers in drawing inspiration from biological systems for hexapod control:

> First and perhaps trivially, the biological system uses specific feedback mechanisms to ensure co-ordination ... Second, in fast moving insects like cockroaches, the control mechanism is sensitive to the speed of progression of the animal ... The dependence of a co-ordinated pattern on sensory feedback during slow walking leads to a third characteristic, that of flexibility or adaptability of response. If normal feedback is interfered with, such as by loss of a leg, the locomotor control system adapts automatically to the loss. This adaptation ensures that the insect will still be able to make reasonable forward progress.

Delcomyn [8] makes some further helpful observations. He suggests that slow walking insects and fast walking insects use different control strategies and that perhaps a hexapod with a wide performance range should have two different control strategies for fast and slow walking. He further suggests there are several ways of coping with performance after damage without repair. In cars, this is achieved by replacement, and in space vehicles by redundancy. Insects use a third approach: if a part of the body is damaged, the control system automatically adapts to allow the remaining parts to continue to function reasonably normally. Hence, there is a need for adaptive control – thus

> Variability of response, for example, a concept largely foreign to mechanical devices, may be an important component in allowing insects to produce appropriate and adaptive responses to unforeseen situations [8].

To summarise this chapter, there are therefore a number of principles that should be considered in designing a hexapod, where it is viable and practical:

1. Legs should be as articulate as possible, with as many as possible of the degrees of freedom employed by insects.
2. Compliant structures are an important part of reducing 'wear-and-tear' on the system as well as realising more energy-efficient solutions to hexapod walking.
3. Engineered sensors tend to be linear over a large range and have little hysteresis in their response characteristics. On the other hand, insect sensors are frequently non-linear, with small range response and show hysteresis. This indicates that the control algorithm in the

central nervous system does not depend on the linear characteristics of its sensory inputs, and thus will not have an analytical kinematic or dynamic reference model. It sometimes seems that biological systems exploit their non-linearities to provide adequate control of agile locomotion [14]. In practice, providing an integration of sensors as numerous as found on an insect is likely to be impossible on a hexapod robot.

4. The sensory input to a hexapod controller should provide at least enough information for the hexapod to 'copy' the strategies that an insect uses in negotiating rough terrain.
5. For very fast walking hexapods, the dynamic stability is important despite the fact that a hexapod can always maintain a statically stable (or quasi-static) base of support.
6. The gait observations of insects serve as a useful model for walking over a wide range of speeds, and it has been shown that just one parameter (stepping frequency) can account for this. Gait generator models and other factors in the co-ordination of walking that result in locomotion over a large range of speeds, and which can be implemented in a hexapod robot, are the subject of further study [29].

References

[1] Randall, M.J., Pipe, A.G., Winfield, A.F.T. & Muttalib, A.Md.Gh., Intelligent hexapod bio-robotics. *Proceedings Euromech 375: Biology and Technology of Walking*, Munich, pp. 124–132, 1998.

[2] Bässler, U., *Neural Basis of Elementary Behaviour in Stick Insects*, Springer-Verlag, 1983.

[3] Graham, D., Pattern and control of walking in insects. *Advances in Insect Physiology*, **18**, pp. 31–140, 1985.

[4] Burrows, M., *The Neurobiology of An Insect Brain*, Oxford University Press: Oxford, 1996.

[5] Bässler, U. & Büschges, A., Pattern generation for stick insect walking movements—multisensory control of a locomotor program. *Brain Research Reviews*, **27**, pp. 65–88, 1998.

[6] Beer, R.D., Quinn, R.D., Chiel, H.J. & Ritzmann, R.E., Biologically-inspired approaches to robotics. *Communications of the ACM*, 1996.

[7] Song, S.M., Vohnout, V.J., Waldron, K.J. & Kinzel, G.L., Computer-aided design of a leg for an energy efficient walking machine. *Proceedings of the 7th Applied Mechanisms Conference*, Part VII, Kansas City, pp. 1–7, 1981.

[8] Delcomyn, F., The walking of cockroaches—deceptive simplicity. *Biological Neural Networks in Invertebrate Neuroethology and Robotics*, R.D. Beer, R.E. Ritzmann & T. McKenna, Academic Press, pp. 21–41, 1993.

[9] Burrows, M., Local interneurons and the control of movement in insects. *Neuroethology and Behavioral Physiology: Roots and Growing Points*, eds. F. Huber & H. Markl, Springer-Verlag, pp. 26–41, 1983.

[10] Hustert, R., The contribution of proprioceptors to the control of motor patterns of legs in orthopterous insects—the locust example. *Insect Locomotion*, eds. M. Gewecke & G. Wendler, Paul Parey, pp. 59–67, 1985.

[11] Cruse, H., The function of the legs in the free walking stick insect, *Carausius morosus*. *Journal of Comparative Physiology A*, **112**, pp. 235–262, 1976.

[12] Cruse, H., The control of body position in the stick insect (*Carausius Morosus*) when walking over uneven surfaces. *Biological Cybernetics*, **24**, pp. 25–33, 1976.

[13] Delcomyn, F., Insect locomotion on land. *Locomotion and Energetics in Arthropods*, eds. C.F. Herried II & C.R. Fourtner, Plenum: New York, pp. 103–125, 1981.

[14] Delcomyn, F., Nelson, M.E. & Cocatre-Zilgien, J.H., Sense organs of insect legs and the selection of sensors for agile walking robots. *International Journal of Robotics Research*, **15(2)**, pp. 113–127, 1996.
[15] Cruse, H. & Bartling, Ch., Movement of joint angles in the legs of a walking insect, *Carausius morosus. Journal Insect Physiology*, **41(9)**, pp 761–771, 1995.
[16] Fourtner, C.R., Role of muscle in insect posture and locomotion. *Locomotion and Energetics in Arthropods*, eds. C.F. Herried II & C.R. Fourtner, Plenum: New York, pp. 195–213, 1981.
[17] Graham, D., Influence of coxa-thorax joint receptors on retractor motor output during walking in *Carausius morossus*. *Journal of Experimental Biology*, **114**, pp. 131–139, 1985.
[18] Cruse, H., Is the position of the femur-tibia joint under feedback control in the walking stick insect? I. Force Measurements. *Journal of Experimental Biology*, **92**, pp. 87–95, 1981.
[19] Cruse, H. & Pflüger, H.J., Is the position of the femur-tibia joint under feedback control in the walking stick insect? II. Electrophysiological recordings. *Journal of Experimental Biology*, **92**, pp. 97–107, 1981.
[20] Cruse, H. & Schmitz, J., The control of the femur-tibia joint in the standing leg of a walking stick insect *Carausius morossus*. *Journal of Experimental Biology*, **102**, pp. 175–185, 1983.
[21] Dean, J., A simulation of proprioceptive input form the coxal hair rows of the stick insect: possible effect of step velocity on the representation of joint angle. *Insect Locomotion*, eds. M. Gewecke & G. Wendler, Paul Parey, pp. 49–57, 1985.
[22] Bässler, U., Proprioceptive control of stick insect walking. *Coordination of Motor Behaviour*, eds. B.M.H. Bush & F. Clarac, Society for Experimental Biology, Seminar Series, Vol. 24, Cambridge University Press, pp. 271–281, 1985.
[23] Hofmann, T., Koch, U.T. & Bässler, U., Physiology of the femoral chordotonal organ in the stick insect *Cuniculina impigra*. *Journal of Experimental Biology*, **114**, pp. 207–223, 1985.
[24] Bässler, U., Functional principles of pattern generation for walking movements of stick insect forelegs: the role of the femoral chordotonal organ afferences. *Journal of Experimental Biology*, **136**, pp. 125–147, 1988.
[25] Hofmann, T. & Koch, U.T., Acceleration receptors in the femoral chordotonal organ in the stick insect *Cuniculina impigra*. *Journal of Experimental Biology*, **114**, pp. 225–237, 1985.
[26] Weiland, G. & Koch, U.T., Sensory feedback during active movements of stick insects. *Journal of Experimental Biology*, **133**, pp. 137–156, 1987.
[27] Dean, J. & Schmitz, J., The two groups of sensilla in the ventral coxal hairplate of Carausius morosus have different roles during walking. *Physiological Entomology*, **17**, pp. 331–341, 1992.
[28] Schmitz, J., Load-compensating reactions in the proximal leg joints of stick insects during standing and walking. *Journal of Experimental Biology*, **183**, pp. 15–33, 1993.
[29] Randall, M.J., *Adaptive Neural Control of Walking Robots*, Professional Engineering Publishing: Suffolk, UK, 2001.
[30] Wilson, D.M., Insect Walking. *Annual Review of Entomology*, **11**, pp. 103–122, 1966.
[31] Bässler, U., Sensory control of leg movement in the stick insect *Carausius morosus*. *Biological Cybernetics*, **25**, pp. 61–72, 1977.
[32] Cruse, H., Riemenschneider, D. & Stammer, W., Control of the body position of a stick insect standing on uneven surfaces. *Biological Cybernetics*, **61**, pp. 71–77, 1989.
[33] Cruse, H., Schmitz, J., Braun, U. & Schweins, A., Control of body height in a stick insect walking on a treadwheel. *Journal of Experimental Biology*, **181**, pp. 141–155, 1993.
[34] Precht, H., Das taxisproblem in der zoologie. *Z. Wiss. Zool.*, **156**, p. 1, 1942.

[35] Graham, D. & Cruse, H., Coordinated walking of stick insects on a mercury surface. *Journal of Experimental Biology*, **92**, pp. 229–241, 1981.
[36] Delcomyn, F., The locomotion of the cockroach *Periplaneta americana*. *Journal of Experimental Biology*, **54**, pp. 443–452, 1971.
[37] Graham, D., A behavioural analysis of the temporal organisation of walking movements in the first instar and adult stick insects (*Carausius morosus*). *Journal of Comparative Physiology*, **116**, pp. 91–116, 1972.
[38] Graham, D., Unusual step patterns in the free walking grasshopper *Neoconocephalus robostus*, I. General features of the step patterns. *Journal of Experimental Biology*, **73**, pp. 147–157, 1978.
[39] Song, S.M., *Kinematic Optimal Design of a Six-Legged Walking Machine*, PhD Thesis, Ohio State University, 1984.
[40] Graham, D., Simulation of a model for the coordination of leg movement in free walking insects. *Biological Cybernetics*, **26**, pp. 187–198, 1977.
[41] Graham, D., Unusual step patterns in the free walking grasshopper *Neoconocephalus robostus*, II. A critical test of the leg interactions underlying the different models of hexapod co-ordination. *Journal of Experimental Biology*, **73**, pp. 149–172, 1978.
[42] Holst, E., von, Über relative koordination bei arthropoden. *Pflügers Arch. Ges. Physiol.*, **246**, pp. 847–865, 1943.
[43] Manton, S.M., The evolution of arthropodan locomotory mechanisms Part 10: locomotory habits, morphology and evolution of hexapod classes. *Zoological Journal of the Linnean Society*, **51**, pp. 203–400, 1972.
[44] Dean, J., Effect of load on leg movement and step coordination of the stick insect *Carausius morosus*. *Journal of Experimental Biology*, **159**, pp. 449–471, 1991.
[45] Bässler, U., Proprioceptive control of stick insect walking. *Insect Locomotion*, eds. M. Gewecke, & G. Wendler, Paul Parey, pp. 43–48, 1985.
[46] Camhi, J.M., Escape behavior in the cockroach: distributed neural processing. *Experientia*, **44**, pp. 401–408, 1988.
[47] Beer, R.D., Kacmarcik, G.J., Ritzmann, R.E. & Chiel, H.J., A model of distributed sensorimotor control in the cockroach escape turn. *Neural Information Processing Systems*, **3**, eds. R.P. Lippmann, J.E. Moody & D.S. Touretzky, Morgan Kaufmann, pp. 436–442, 1991.
[48] Beer, R.D. & Chiel, H.J., Simulations of cockroach locomotion and escape. *Biological Neural Networks in Invertebrate Neuroethology and Robotics*, eds. R.D. Beer, R.E. Ritzmann & T. McKenna, Academic Press, pp. 267–285, 1993.
[49] Cruse, H., What mechanisms coordinate leg movement in walking arthropods? *Trends in the Neurosciences*, **13**, pp. 15–21, 1990.
[50] Jander, J.P., Mechanical stability in stick insects when walking straight and around curves. *Insect Locomotion*, eds. M. Gewecke & G. Wendler, Paul Parey, pp. 33–42, 1985.
[51] Schmitz, J. & Haßfield, G., The treading-on-tarsus reflex in stick insects: phase-dependence and modifications of motor output during walking. *Journal of Experimental Biology*, **143**, pp. 373–388, 1989.
[52] Cruse, H. & Saxler, G., The co-ordination of force oscillations and of leg movement in a walking insect (*Carausius morosus*). *Biological Cybernetics*, **25**, pp. 143–153, 1980.
[53] Cruse, H., The influence of load, position and velocity on the control of leg movement of a walking insect. *Insect Locomotion*, eds. M. Gewecke & G. Wendler, Paul Parey, pp. 19–26, 1985.
[54] Cruse, H., Which parameters control the leg movement of a walking insect? I. Velocity control during the stance phase. *Journal Experimental Biology*, **116**, pp 343–355, 1985.

[55] Cruse, H., Which parameters control the leg movement of a walking insect? II. The start of the swing phase. *Journal Experimental Biology*, **116**, pp. 357–362, 1985.

[56] Dean, J., Control of leg protraction in the stick insect: a targeted movement showing compensation for externally applied forces. *Journal of Comparative Physiology A*, **155**, pp. 771–781, 1984.

[57] Weiland, G. & Koch, U.T., Sensory feedback during active movements of stick insects. *Journal of Experimental Biology*, **133**, pp. 137–156, 1987.

[58] Pearson, K.G. & Franklin, R., Characteristics of leg movements and patterns of coordination in locusts walking on rough terrain. *The International Journal of Robotics Research*, **3**, pp. 101–112, 1984.

[59] McGhee, R.B. & Iswandhi, G.I., Adaptive locomotion of a multilegged robot over rough terrain. *IEEE Transactions of Systems, Man, and Cybernetics*, **SMC-9(4)**, 1979.

[60] Franklin, R.F., The locomotion of hexapods on rough ground. *Insect Locomotion*, eds. M. Gewecke & G. Wendler, Paul Parey, pp. 69–78, 1985.

[61] Cruse, H., The control of the anterior extreme position of the hindleg of a walking insect, *Carausius Morosus*. *Physiological Entomology*, **4**, pp. 121–124, 1979.

[62] Dean, J., Coding proprioceptive information to control movement to a target: simulation with a simple neural network. *Biological Cybernetics*, **63**, pp.115–120, 1990.

[63] Espenschied, K.S., Quinn, R.D., Beer, R.D. & Chiel, H.J., Biologically based distributed control and local reflexes improve rough terrain locomotion in a hexapod robot. *Robotics and Autonomous Systems*, **18**, pp. 59–64, 1996.

[64] Graham, D. & Bässler, U., Effects of afference sign reversal on motor activity in walking stick insects (*Carausius morosus*). *Journal of Experimental Biology*, 1981.

[65] Büschges, A., Schmitz, J. & Bässler, U., Rhythmic patterns in the thoracic nerve cord of the stick insect induced by Pilocarpine. *Journal of Experimental Biology*, **198**, pp. 435–456, 1995.

[66] Godden, D.H. & Graham, D., A preparation of the stick insect *Carausius morosus* for recording intracellularly from identified neurons during walking. *Physiological Entomology*, **9**, pp. 275–286, 1984.

[67] Bässler, U., Rohrbacher, J., Karg, G. & Breutel, G., Interruption of searching movements of partly restrained front legs of stick insects: a model situation for the start of a stance phase? *Biological Cybernetics*, **65**, pp. 507–514, 1991.

[68] Bässler, U., The walking- (and searching-) pattern generator of stick insects, a modular system composed of reflex chains and endogenous oscillators. *Biological Cybernetics*, **69**, pp. 305–317, 1993.

[69] Zill, S.N., Load compensatory reactions in insects: swaying and stepping strategies in posture and locomotion. *Biological Neural Networks in Invertebrate Neuroethology and Robotics*, eds. R.D. Beer, R.E. Ritzmann & T. McKenna, Academic Press, pp. 43–68, 1993.

[70] Pflüger, H.J., The control of rocking movements of the phasmid *Carausius morosus*. *British Journal of Comparative Physiology*, **120**, pp. 181–202, 1977.

[71] Kittmann, R., Schmitz, J. & Büschges, A., Premotor interneurons in generation of adaptive leg reflexes and voluntary movements in stick insects. *Journal of Neurobiology*, **31(4)**, pp. 512–531, 1996.

[72] Alexander, R. McN., Three uses for springs in legged locomotion. *The International Journal of Robotics Research*, **9(2)**, pp. 53–61, 1990.

[73] Full, R.J., Integration of individual leg dynamics with whole body movement in arthropod locomotion. *Biological Neural Networks in Invertebrate Neuroethology and Robotics*, eds. R.D. Beer, R.E. Ritzmann & T. McKenna, Academic Press, pp. 3–20, 1993.

[74] Delcomyn, F., Insect locomotion: past, present and future. *Insect Locomotion*, eds. M. Gewecke & G. Wendler, Paul Parey, pp. 1–18, 1985.
[75] Cruse, H., Coactivating influences between neighbouring legs in walking insects. *Journal of Experimental Biology*, **114**, pp. 513–519, 1985.
[76] Pfeiffer, F. & Cruse, H., Bionik des laufens—technische umsetzung biologischen wissens. *Konstruktion*, **46**, pp. 261–266, 1994. [trans. Peter and Birte Pals, November 1995]
[77] Hughes, G.M., The co-ordination of insect movements. II. The effect of limb amputation and the cutting of commisures in the cockroach (*Blatta orienetalis*). *Journal of Experimental Biology*, **34**, pp. 306–333, 1952.

Chapter 10

The palm – a model for success?

A. Windsor-Collins[1], D. Cutler[2], M. Atherton[1] & M. Collins[1]
[1]*School of Engineering and Design, Brunel University, Uxbridge, Middlesex, UK.*
[2]*Royal Botanic Gardens, Kew, UK.*

Abstract

The tree palm provides us with an unusual example of where evolution seems to bring about a reduction in macro structure complexity resulting in a long-lived but not necessarily structurally efficient living system 'design'. The biological background to these durable and successful plants in terms of evolution has been investigated and the results are related to engineering design. The botanical design rationale is essentially survival, not infrequently in harsh weather conditions, and reproduction. More specifically, the mechanisms of the palm functions described offer potential for interpretation in engineering terms. Thus, the structure and features of the palm are discussed in general terms followed by the more detailed thermofluid aspects of the leaf.

1 Introduction

Palms can take many different forms and although they are specialised, their morphology can vary greatly. For ease of reference and purposes of this chapter, the palm shall be divided broadly into three principal areas: the root, the trunk/stem and the crown.

Palms derive from a monophyletic group of plants, meaning that they share a single common ancestor. They constitute a family of about 200 genera and 2600 species [1] and they belong to the family commonly named Palmae, but correctly known as Arecaceae.

The double coconut (*Lodoicea*) produces the heaviest seed, which may be up to 20 kg in weight [2]. The fruit has an outer layer of tough, fleshy, fibrous material which protects it from predators and enables the seed to float on the water (one of the methods of seed dispersal). The seed takes about a year to germinate and about another year to form its first leaf.

Palms can be found in rainforests, in the under-storey region of forests, at mid-canopy level and in the canopy itself as well as in the mangroves, mountain forest, the desert (e.g. *Washingtonia*) and river banks making them the most ecologically diverse family of plants [3]. Palms are adapted to a diverse range of environments. However, palms do not match the diversity of form found in *dicotyledons* (*dicots*). A large percentage of palms follow the same model in that they have simple flexible stems. Two-thirds of this family are associated with rainforests of Central and

South America and Southeast Asia. Some palm genera dominate large areas of forest (*Hyphaene* and *Metroxylon* [1]) while others grow in localised habitats.

The longest stem recorded is about 200 m from a species of *Calamus* (Tomlinson, personal communication) making it twice as long as the tallest dicotyledonous and *coniferous* trees. These palms are familiarly called rattans and are not self-supporting when adult. However, they grow up into the forest canopy by the use of spines (epidermal emergences). *Ceroxylon quinduiense* has the longest, self-supporting, un-branched terrestrial stem in the world and is the national tree of Columbia, commonly known as the Andean Wax Palm. *Ceroxylon quindiuense* can reach heights of up to 60 m [2]. Its trunk is covered in a once commercially important white wax and is known as one of the cold hardiest species. It grows in very large groups as part of the forest and is often left when the forest is cleared because it seems the wood blunts chainsaws almost instantaneously. Although *Ceroxylon quinduiense* has the tallest free-standing trunk in the world, it is not necessarily the most structurally efficient un-branched stem in the world.

Palm leaves vary widely in their shape and size as well as function. *Raphia*, a climbing palm, has the longest leaves which can be up to 25 m in length (Tomlinson, personal communication). *Corypha umbellata* however is a species of fan (palmate) palm that produces the world's largest leaf [4] and can be up to 5 m wide. As palms are usually un-branched, they do not have as many points of attachment for leaves compared with branched trees. Palm leaves tend to be larger than those of other plants. Tree palms often emerge above the forest canopy and spread their leaves ensuring efficient light interception for photosynthesis.

As well as having all of these extreme dimensions, palms are among the least branched terrestrial plants in the world. Their energy goes into extending the trunk and producing fruit, leaves and roots rather than lateral branching growth. However, branching occurs in the inflorescences and roots, and some palms branch basally just below the surface of the soil. Palms have survived very long on the evolutionary tree risking the destruction of the one and only apical bud, a kind of design compromise. Despite their apparently simpler structure, palms have survived as a recognisable group through millions of years.

2 Evolutionary theory and complexity

2.1 The simplicity of monocots

One evolutionary route is for organisms to become more complex over time. However, there seems to be an anomaly with palms as, at first glance, they appear to have either retained a simple morphology or perhaps have evolved a less complex form; yet they are mostly tree-like. They are described as woody *monocotyledons* (*monocots*). They lack the vascular cambium found in most dicots and all *gymnosperms* and have a different method of stem thickening based on primary growth in thickness near the growing tip. Palms have hardened ground tissue and, as with many other types of trees, have developed to become mechanically strong rather than changing their form. Palm mechanics do not depend on their complexity of having an expanding living cylinder of tissues. They retain functional conducting cells throughout their lifetime.

According to Pearson [5], *gymnosperm*s are considered to be the first seed plants to develop on earth, although according to Ennos [6], the seed ferns pre-dated these. The *angiosperms* are by far the most successful and the most recent of the plant groups. The group contains around 80,000 species of trees compared to only 600 species in the other groups. The monocots arose soon after the dicots from which they 'branched' in evolutionary terms. It is thought that the first monocots were aquatic plants with their woody tissue in isolated strands throughout their body. This enabled

them to resist stretching by water currents [6]. Although it is possible that the early *pinnate* and *palmate* leaves formed at the same time on the evolutionary scale, interestingly, pinnate leaves may also have derived from palmate leaves.

Conifers are the most successful gymnosperm*s* according to Ennos [6] and their wood contains only *tracheids*, no *vessels* and minimal *parenchyma*, so conifer wood appears more uniform than the wood of *angiosperms*. Angiosperms have vessels which are normally wider in diameter than tracheids, along with numerous fibres. Angiosperm wood seems to be efficient at conducting water, though this may on occasion be at the expense of safety. Vessels are made up of a linear sequence of short vessel elements and the vessel elements have pores in their end walls. Vessels may reach several metres in length. Consequently, angiosperm wood as a whole transport water very efficiently, but is prone to embolism in dry or freezing weather. In other words angiosperm wood is particularly well suited to the climate in tropical rainforests where it is warm and wet. Palms are among these angiosperms.

Most palm species do not grow as tall as the larger dicot trees. This may reflect mechanical constraints based on the internal structure of palms. Here it seems, secondary growth in thickness gives the dicots the advantage. This manifests itself by the presence of both niche and generalist palms. Perhaps the structure of the palm provides it with extreme physical flexibility. Adaptation to changing environments is one of the key features of evolution in which selection pressure drives change in form and physiology as well as many other aspects. Retaining a simple un-branched form over millions of years is of significance. It may reflect the efficiency of the early model or it may just represent a constraint that is tolerated while other features permit success.

2.2 Neoteny in palms

The meaning of Neoteny means that reproduction occurs while an organism is still in, or maintains many characteristics of, its juvenile stage. An example of this may be in the palm *Arenga caudata* shown in Fig. 1 where it can be seen that the leaves are very large in comparison with the size of the rest of the palm. Takhtajan [7] suggested that neoteny played a significant part in the evolution of angiosperms.

2.3 Survival and other consequences of lack of branching in palms

One of the ways in which the palm differs from dicot trees is with the number of branches it possesses [6]. Palm trees usually have no branches, apart from the root system and inflorescences, although they may be multi-stemmed or have creeping rhizomes which can bud from the base. The palm crown consists of closely packed, overlapping leaves which protect the all-important apical cell-producing meristem of which there is only one for each aerial stem; in this respect palms differ from most other dicot trees. From this single point in non-branched palms, all leaves emerge. Structurally, in effect, the petioles are the branch equivalents (not branches, as petioles are of determinate growth and do not re-branch) as they support often larger than average laminas. A high proportion of the energy produced by the palm during the vegetative stage is used in trunk/stem extension (e.g. rattans) rather than lateral branching, whereas in branching trees a higher proportion of energy is typically used to widen the trunk and grow branches from the trunk. Perhaps this enables some canopy palms to grow relatively quickly into a small space in the canopy compared with branching trees. However, some palms grow very slowly even after the establishment phase. Before the extension upwards occurs, the basal subterranean part of the palm stem expands to maximum girth whilst still underground during the ‘establishment phase’ after which it grows upwards [6]. Palms can remain short even after the establishment phase is complete

Figure 1: The palm *Arenga caudata* showing possible traits of neoteny. Photo taken at Fairchild Tropical Gardens, Florida.

and when there is enough sunlight, they can grow further – not necessarily upwards as in the case of the climbing rattans which can also grow laterally some of the time. In the case of a forest, the growth upwards is more likely to happen when there is a space in the canopy of the forest above allowing the sunlight to filter through. Sometimes when a tropical forest is looked upon aerially, palms can be found to be a few metres taller than the other trees especially in the neo-tropics. This is important as the one and only group of leaves, the crown, needs to have access to sunlight to photosynthesise. Another way for some species of palm to survive is by having short trunks and very large leaves, as those have in the forest understorey. The leaves have an advantage of having large laminas to intercept enough of the weak sunlight needed for photosynthesis. These palms are often multi-stemmed, sharing the same root system, but are usually un-branched. Perhaps this is because the arboreal woody monocot stem is not strong enough to support branches. It may be that the ability to branch has been lost during their evolution or more likely it never gained it throughout, except for the rare exceptions.

2.4 Design constraints in palms

The main design constraint in palms is to protect the apex of the stem as this contains the only shoot meristem producing cells for growth. Each root has its own apical meristem, of course, and their own inflorescences. As soon as the apical meristem ceases to operate, the palm dies. The meristem is protected by tightly sheathing leaves under great hydraulic pressure that completely surrounds it. For long periods of the life of a palm, the apical meristem is either underground, expanding laterally to the maximum girth or at full adult height away from predators obtaining maximum sunlight. As the stem does not branch, it is more likely than other trees to be straight

especially in a forest where they may grow towards the nearest hole in the canopy for sunlight. Being straight is an important structural property enabling ordinarily straight palms to support their own weight better and to withstand external forces.

Taller trees have to have greater hydraulic resistance than short trees and water is under greater cohesive tension in the trunks.

3 Botanical aspects of palms

An overview of the botanical aspects of palms proceeds highlighting interesting cross-relationships with engineering structures. The success of the palm relates to its ability to reproduce and therefore it has to survive structurally until this time. All plants compete for water, nutrients and especially light [6].

3.1 Palm trunk anatomy

Some palm stems have their strength concentrated towards the periphery and some may in effect act as a fairly stiff rope enabling the trunks to bend easily without breakage. Others have their strength distributed throughout the stem section.

3.1.1 Arrangement of vascular bundles

The internal structure of a palm trunk can be likened to longitudinally reinforced concrete. According to Winter [8], palms are the largest terrestrial plants lacking secondary *xylem* and instead there is what initially appears to be a random distribution of vascular bundles throughout the stem some with a higher concentration towards the periphery. The tensile modulus of palm vascular bundles is astonishing at 100 GPa [4] compared to Kevlar which ranges from 83 to 186 GPa (grades 29–149). *Sclerenchyma* (derived from the word 'scleros' in Greek meaning 'hard') is the lignified material that partially or entirely surrounds the slightly helical to near vertical vascular bundles which in effect form tubes. The lower two thirds and the periphery of the stem usually have the highest elastic modulus which decreases by an order of magnitude towards the crown.

In Fig. 2, the axial part of the vascular bundles is shown as vertical but in reality, palm vascular bundles are approximately helical. Although the branches into the leaf trace pipes are lost (dashed lines), the axial part still functions hydraulically and structurally. In full sunlight, it is better for a tree to have a multilayer leaf arrangement rather than a monolamellate one, so that more light can filter through to the leaves below and there is more likelihood of light at higher angles of incidence. According to Ennos [6], many leaves photosynthesise at their highest rate even at 20% full sunlight, so leaves which are partly shaded can still be 100% effective.

3.1.2 Self-defence mechanism of the trunk

As palms have no peripheral vascular cambium as in dicots, but have scattered vascular bundles throughout the stem, they can remain alive even after the outer layers of the stem are damaged to a great extent. Translocation is the process of transporting dissolved nutrients within a plant and takes place within the *phloem*. Palm phloem *translocates* for many years unlike that of many other plants which do so only for a few seasons. Indeed, the palm phloem has to remain active in many vascular bundles for the life of the palm [10], as palms do not grow new phloem tubes like most dicots. The oldest parts of the phloem are situated at the base of the tree. Unlike dicots, palms do not make phloem and xylem each year to survive; this is done only in their apical meristem. The construction of palms is 'cheaper' than that of dicots as they use far fewer resources for

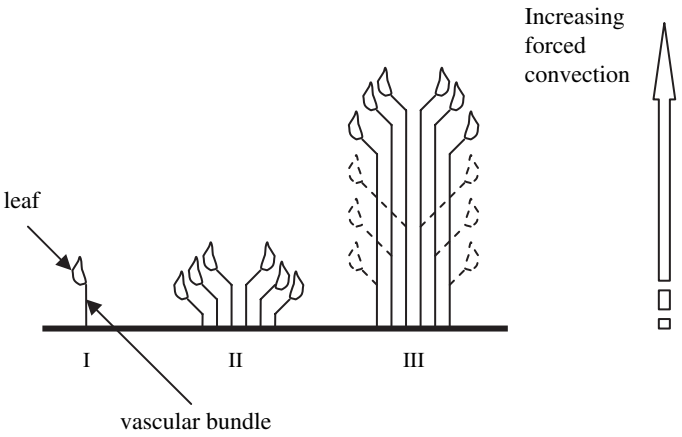


Figure 2: Pipe model of monocot tree construction (after Shinozaki *et al.* [9]). (I) Unit pipe. (II) Mature crown with fixed number of pipes. (III) Older leaves die but number of pipes is constant.

mechanical purposes. Trees are able to live a long time relative to herbaceous plants because the wood is a strong, long-lasting material [6]. The stem model is robust compared to that of the leaf as it does not vary as much in size or form.

3.2 Palm blade anatomy

3.2.1 Three main designs of lamina

The three main forms of palm blade are *pinnate*, *costapalmate* and *palmate*, which are shown in Figs 3–5. The blade is connected to the petiole which in turn is connected to the stem. All palm leaves are plicate, i.e. have 'accordion folds', which are formed by the differential growth of the surface of the leaf initial. Segmentation is then superimposed on the accordion folds of the leaf. Segment placation is usually of two kinds: reduplicate (edges folded down) and induplicate (edges folded up, holding water 'in').

Pinnate blades have separate leaflets attached to an extension of the petiole called the rachis. Often the structure is likened to that of a feather. They can survive exposure to winds more than the other blade types because of the independence of the leaflets, their flexibility and consequent ability to be aligned parallel to the direction of the prevailing wind. The tallest palms usually have pinnate leaves and this may be a result of the fact that wind speed generally increases with height from the ground. So in the case of tall palms, the leaves provide little of the overall wind resistance.

Palmate blades do not have a rachis, but are largely corrugated in structure and directly connected to the petiole. They can be found in the forest under-storey where the sun is weakly filtered through the canopy and where the wind speed is low, although quite a few semi-arid and adapted fan palms occur in open habitats. The leaflet equivalents are mainly fused to support the large lamina although at the distal ends, they are often separated to reduce the turbulent flow of the wind over them. Palms with palmate leaves are usually found in dark, wet areas and are often very large. More exposure to the weak light is required for photosynthesis, so there is a high density of photosynthetic pigments (e.g. chlorophyll) in many under-storey fan leaves

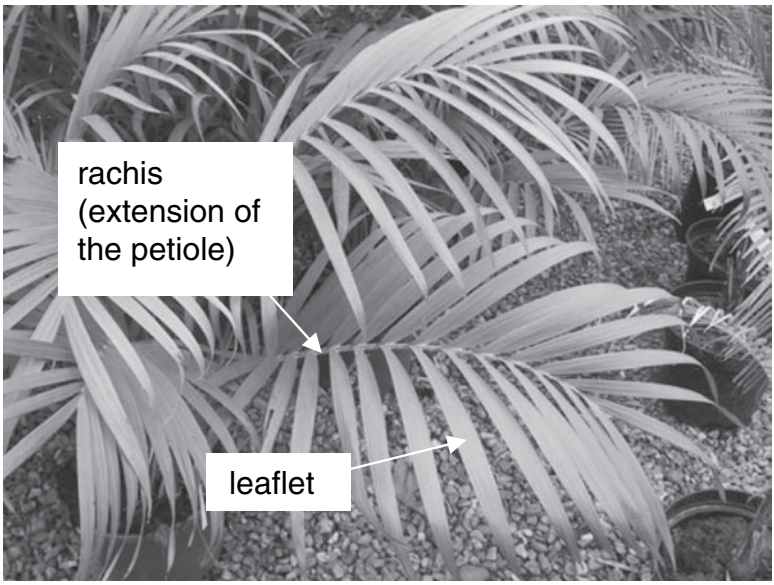


Figure 3: Pinnate leaves from the palm *Dypsis lutescens*. Photo taken at The Palm Centre, Ham, Richmond.

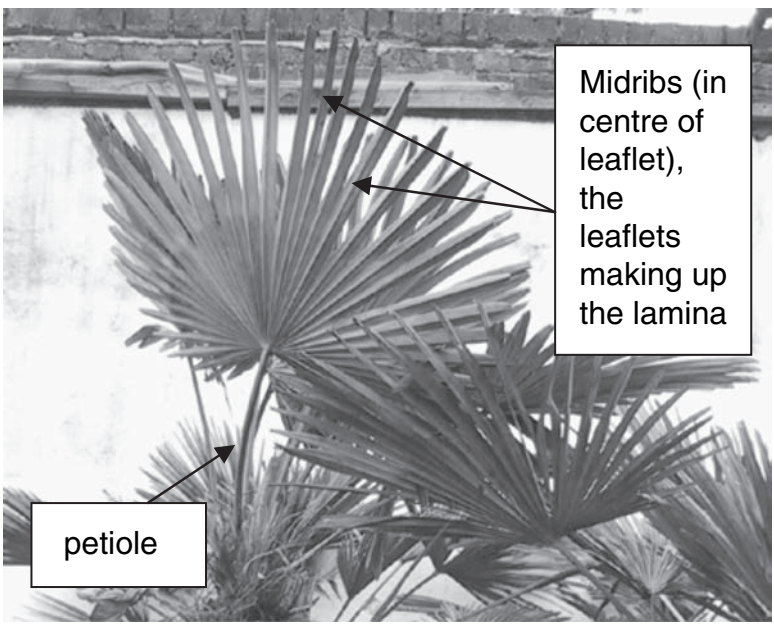


Figure 4: Palmate leaves from the palm *Trachycarpus wagnerianus* (Miniature Chusan wax palm). Photo taken at The Palm Centre, Ham, Richmond.

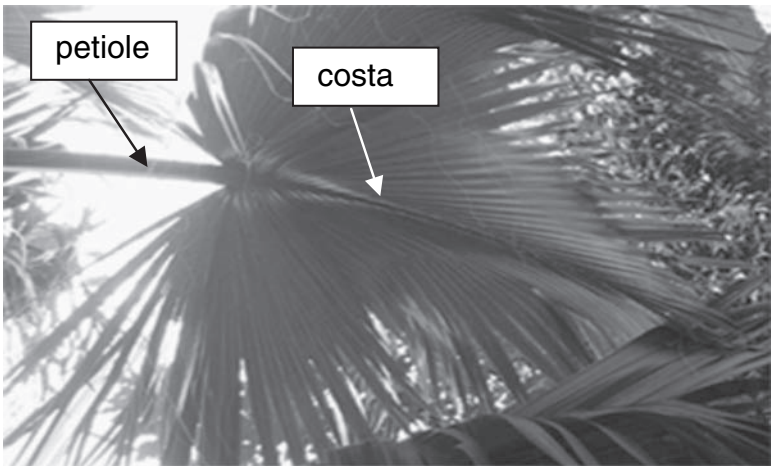


Figure 5: Costapalmate palm leaf showing the petiole extending into the blade from the palm *Sabal bermudana*. Photo taken at Royal Botanic Gardens, Kew.

which look relatively dark (compared with palms whose habitat is in the sun) because of their higher than normal pigment concentrations. The presence of wax on thick leaves helps prevent UV damage to chloroplasts near the upper surface by the reflection of light and heat.

The costapalmate type leaf shape can be described as being of that between pinnate and palmate. They display corrugated blades incorporating a rachis which can vary in length according to genus type. These leaves are thought to be the largest of the three types and so this design can support the heaviest load although this has not been verified.

3.2.2 Leaf age and resistance to weathering

The larger palm leaves have to be very tough to withstand the external forces of wind, snow (for some species) and water and this also reflects their age as they are relatively very long lived compared with leaves from deciduous trees. Palms generally are not frost hardy and some have leaves that die while their bases persist and lag the stems. In most palms however, these leaves will die (senesce) and then become detached from the trunk. Different parts of a palm have their own frost resistance levels. The tips of leaves are the most prone to frost damage and the growth centre is the most frost-resistant. The apical bud can also survive temperatures far below the values which kill leaves that are fully developed.

3.2.3 Effects of leaf shape and structure

The availability of water and nutrients determines mainly the volume of the leaves as these provide the material to build the cells for growth. *Forced convection* (wind currents) on the leaf affects *transpiration* and cooling rates as well as providing physical stress which when applied (in non-destructive ways), induce a strengthening response in plants. Much research on the plant response to mechanical and *forced convection* has identified mechanosensitive ion channels which regulate the amount of calcium ions entering the cell as a result of stress. Calcium ions probably regulate the first stage in the signal transduction pathway from the first perception of the mechanical stress to the activation of genes which synthesise proteins in response to the stress [11].

This wind stress is influenced by the shape of the leaf and may be partly self-induced by the forces produced by the channelling of air flows especially in palmate palm leaves.

Figure 6: The ruffled fan palm (*Licuala grandis*). Photo taken at The Palm Centre, Ham, Richmond.

The surface area, angle and scattering properties of leaf surfaces together determine how much light flux is intercepted by the leaf. Figure 5 is an example of a palm leaf with a large surface area for maximum exposure to sun rays as discussed in more detail in Section 3.2.1. The habitat of *Licuala grandis* (Fig. 6) in the under-storey of rainforests across southern and eastern Asia [2] is one of low light levels so it needs to obtain as much of this necessary resource as possible for survival.

3.2.4 Leaf wax

When present, leaf wax is continuously produced up to a certain threshold and protects the leaf. The form of wax crystals appears to have evolved in relation to the degree of wind and insolation exposure by the species. Plants subjected to high insolation may have thick wax on their leaf surfaces particularly on the *adaxial* surface. This is part of the mechanism that results in the reflection of both heat and light from the surface of the leaf.

The texture of wax can range from being rough to smooth. The rougher the wax, the more heat and light appears to be reflected. Visually the palms appear to be more glaucous/whitish in colour when this is the case. The structure of epicuticular wax may be influenced by the conditions of the environment [12].

3.2.5 Predation and leaf form

The amount of predation on a leaf is partly determined by the surface texture of the leaf and also by the internal structure in the form of rachides: sharp silica bodies which tear organisms that come into contact with them or try to eat them.

Some palm spines are specialized leaflets that are usually tough and sharp. However, most spines are fibrous or epidermal emergences. Their length varies from being very short up to 20 cm and their width ranges from being very thin to about 1 cm. Amongst other places on palms they

can be found at the proximal end of the midrib where they protect the central tender parts of the palm [13].

3.3 Palm petiole anatomy

3.3.1 Petiole shape and strength

Petioles are flexible and strong and can both support the weight of the leaf and withstand forced convection from the wind. In general, the upper (adaxial) surface of the petiole is often concave and the lower (*abaxial*) surface is convex. This shape fits closely to the shape of the trunk when the petioles are parallel to it within the crown shaft. There are a few variations with shape on the adaxial surface that would be interesting to model and compare with engineered cantilevers as the petioles appear to provide structural support as well as withstand torsion forces. Some petioles have a central ridge running along the length of the adaxial surface of the petiole which seems to have developed after partial separation from the trunk. The length and cross sectional width of the petiole also varies from genus to genus. During its development, the petiole is subject to great hydrostatic pressure within the crown shaft which completely surrounds it. When it emerges it splits either on the opposite side of the trunk to the leaf or on the same side of the leaf initially forming a diamond shaped opening as in the *Hyphaene* type leaf base shown in Fig. 7.

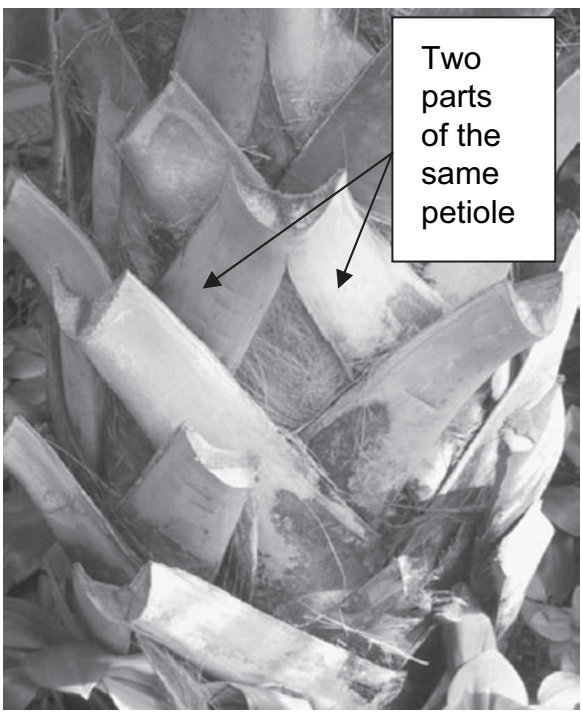


Figure 7: Example of a palm tree (*Sabal* palm) with a hyphaene type leaf base. The petiole splits more or less equally in two allowing for the expansion of the growing part of the trunk while retaining strength. Photo taken at Royal Botanic Gardens, Kew.

3.3.2 Internal arrangement of the vascular bundles within the petiole

The vascular bundles are sometimes concentrated towards the external boundaries of the petiole but are not necessarily arranged concentrically. This type of reinforcement counteracts the external twisting forces in the cantilever, namely the petiole, at the same time supporting a large weight.

3.4 Palm root anatomy

3.4.1 Palm root stresses

The main stresses on the roots are not bending stresses as in stems, but tensile stresses caused by tension in the roots and the effects of leverage of the stems. Roots can be likened to narrow ropes in the sense that the strength is in the centre of a root, in contrast to the strength in some stems which can be concentrated towards the periphery. Roots are very brittle and can be snapped very easily. Palms have many thousands of fine roots which are embedded in the soil, which acts as a composite material. Relatively few palms blow over in storms which, although partly to do with the flexibility of the stems, may be due to the effective root system as the roots fail by being pulled out longitudinally rather than broken.

3.4.2 Palm root regeneration and transplantation

Palms can regenerate adventitious roots readily and so are easy to transplant. Palm roots are all primary and do not have secondary thickening. Transplanting is more successful if some of the outer leaves are removed to decrease transpiration and the stem is supported while the new roots develop. When palms are under more bending and levering stress, they grow more roots in response. As well as this, palm roots (unlike most palm stems) can branch, a useful facility to have, enabling them to respond to stress in localised areas relatively quickly.

3.4.3 Root systems

After the primary root emerges from the base of the palm stem many adventitious roots are produced directly from the bottom of the stem. Often plants respond to environmental stresses by developing structures providing equal and opposite forces for stress compensation. Palms respond to this stress by producing more roots. Stilt roots or prop roots are sometimes formed, e.g. in *Socratea exorrhixa* [2]. These arise above ground and grow into the ground. However, they are not necessarily a stress response, but maybe a strategy to escape the protracted establishment phase.

3.4.4 Pneumatophores

Large air canals are common in the root cortex of palms which grow in waterlogged, anaerobic soils. The air comes from the *pneumatophores,* specialised roots, which grow up from the soil or water surface into the atmosphere.

4 Engineering aspects of palms

4.1 Structural mechanics of palms

4.1.1 Young's modulus and the role of cellulose

Two of the basic engineering properties commonly used to describe a material in engineering terms are Young's modulus (E) and tensile strength.

E is an indication of the elastic flexibility or stiffness of a material and is the ratio of stress to strain of the material (at strains of $\pm 1\%$ for most materials). The density of the substance is also important in terms of its contribution to self weight, which is another design constraint. The tensile strength of a material is its resistance to failure under tension and the higher this is, the more brittle the material. However, these properties cannot be simultaneously, individually maximised.

When dealing with structures that are in some way airborne, it is more useful to compare the *specific* properties of different materials than the absolute values. So, the specific Young's modulus shown in Table 1 is the value of Young's modulus divided by the specific gravity or density of the material.

In palms with straight, vertical trunks, the weight of the trunk itself above ground is the principle component of stress, another being the external forces of wind. The trunk of the climbing palm *Calamus* is probably lighter per unit length than that of a date palm since it is supported by the vegetation through which it scrambles. The weight loading is shared with other vegetation.

The data in Table 1 give the values for tensile and compressive strengths for some common materials, and these are compared graphically in Figs 8 and 9.

Table 1: Table showing values from Gordon [14] for Young's modulus (*E*) and the density of various materials together with the ratio of *E* to specific gravity (apart from other references indicated in the last column of the table).

Material	Young's modulus (*E*) (Gpa)	Tensile strength (*T*) (Mpa)	Density (kg/m^3)	Specific gravity (SG)	*E*/SG (Gpa)	Reference
Coconut palm trunk wood (average)	9		600	0.6		[15]
Spruce (parallel to the grain)	13	100	500	0.5	26	
Spruce (across the grain)	14 (approx.)	3	500	0.5	28	
Date palm leaf midrib	14	90	660	0.66	21	[16]
Concrete	17	4	2400	2.4	7.1	Density [17]
Bone	21	140	1900	1.9	11.1	[18] (apart from Density [19])
Aluminium (cast, pure)	73	70	2700	2.7	27	
Cellulose (single fibre)	100	1000	1500	1.5	66.6	[18]
Kevlar 49 (aramid fibre)	130	3600	1440	1.44	90	
Steel (low tensile)	210 (just 'steel')	400	7800	7.8	26.9	
Steel (high tensile)	210 (just 'steel')	1500	7800	7.8	26.9	

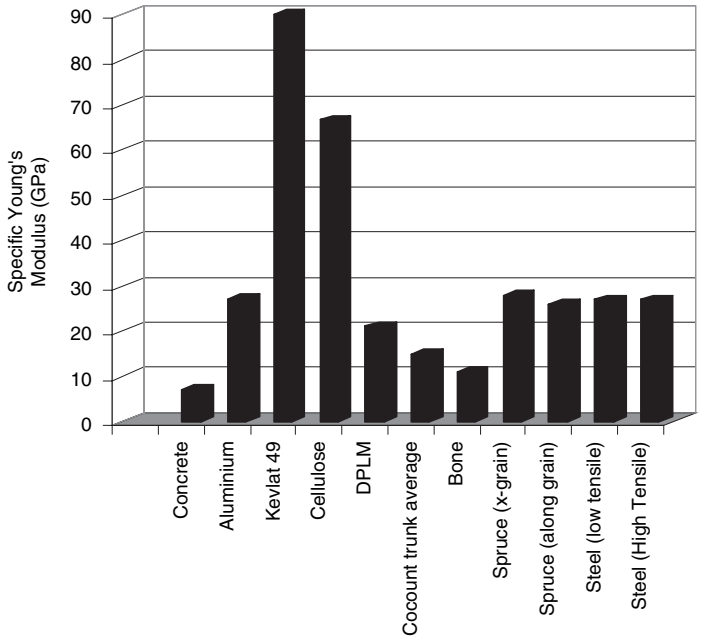


Figure 8: Graph showing the specific Young's modulus per unit of specific gravity (E/SG) for a variety of materials.

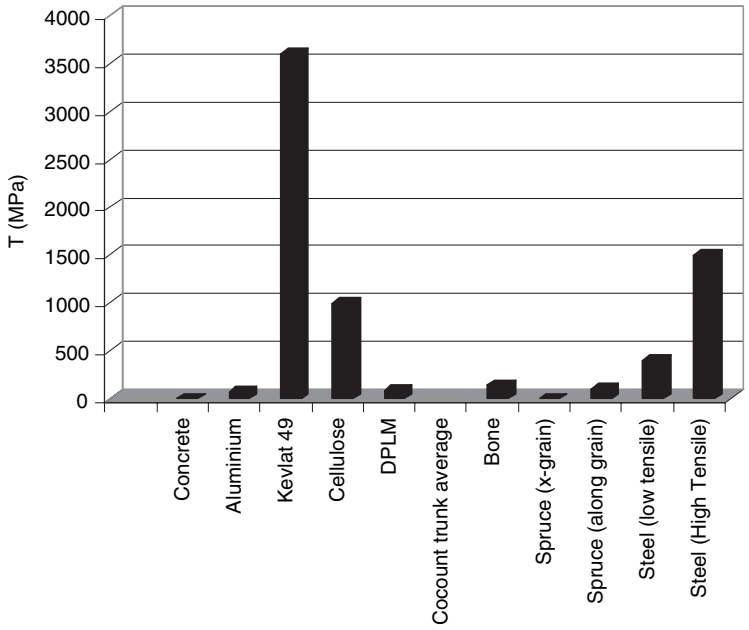


Figure 9: Graph showing the tensile strength for a variety of materials. The tensile strength for the coconut trunk could not be found. The graph shows that the tensile strength of cellulose is similar to that of medium tensile strength steel.

Figure 8 shows the stiffness per unit of specific gravity of the various materials. Spruce, steel and aluminium have approximately the same values and the value for DPLM (date palm leaf midrib) and coconut trunk is a little less. Kevlar and cellulose are the most stiff of the materials analysed.

The very high stiffness of cellulose is interesting in that it is constituent in both spruce and palm (and all wood), but the arrangement and distribution of the cellulose can be just as significant as the properties of the cellulose alone. The compound stiffness of the palm and the spruce is less than that of cellulose. This is because in cellulose, there are weaker materials in addition (law of mixtures). Perhaps other features such as structure account for the extraordinary bending, damage-tolerant, wind-resistant characteristics of palms. The bending moment is maximum at the base of a straight vertical trunk and decreases to zero at the apex. The loading stress is maximum also at the base of a tapered trunk, although the bending moment along the length of the trunk will be different from that of a straight trunk as the top is more flexible. When a straight trunk is tapered only at the top a different bending stress profile exists.

4.1.2 Palm allometry

Allometry is the study of the relative growth of a part of an organism in relation to the growth of the whole. The way in which the tapered self-supporting trunk (axis) height of a mature tree varies in relation to the diameter of the trunk is described by eqn (1) and is calculated from empirical data on the principle of elastic self-similarity.

$$d = kh^n, \tag{1}$$

where $n = 2/3$, h = total height of tree, k = constant and d = diameter of the trunk base. When $n = 1$, then the diameter of the trunk is uniform at different heights [20].

Height and diameter measurements were made from many individual trees in the lowland tropical rain forest in Costa Rica [21]. These trees, in general, have a four-fold safety factor, i.e. they could grow to four times their own height before reaching their buckling limit and collapsing under their own weight [20].

Depending on the application, a typical engineered design would require a safety factor of between one and two, much less than the apparent value in normal tree design. For straight trunks, the Euler strut formula [4] can be applied to calculate the critical load, which is the smallest axial load sufficient to keep the column in a slightly bent form.

$$P_{\mathrm{cr}} = \pi^2 El/4l^2, \tag{2}$$

where P_{cr} is the critical load, E is the Young's modulus and l is the length of the column, in this case the trunk.

However, palm trunks are not perfectly straight, although they tend to taper much less than the majority of conifer and dicot tree species. The stem is not isotropic in terms of Young's modulus, although it is much more isotropic than conifer and dicot species.

Most tree palms approximate to eqn (1) although the stem diameter is usually randomly uneven. Assuming a constant safety factor (k) of 0.1 in eqn (1), it can be seen in Fig. 10 that for a given tree height, the trunk diameter is much thicker for a straight trunk than in a tapered trunk and therefore uses more wood for the same height. It can be said that for a given average diameter, tapered trunks are efficient with their use of material and that palms are structurally inefficient in this respect. However, angiosperm trees and gymnosperms normally develop trunks which are overall, thicker than those of palms in mature trees and there is a cross-over point where they become far less efficient.

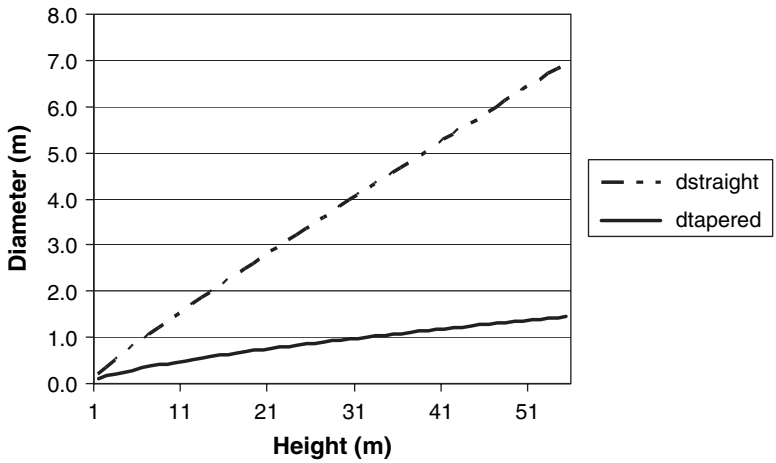


Figure 10: Graph showing how theoretical height (h) and diameter (d) vary for a tapered trunk ($d = kh^{2/3}$) and a straight trunk ($d = kh$) assuming $k = 0.1$. Formula from [20].

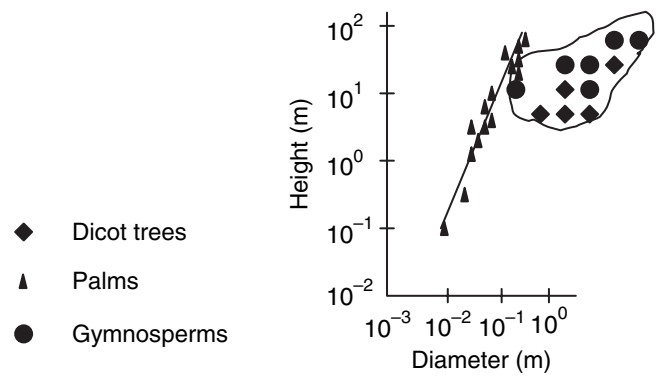


Figure 11: Graph showing tree height versus trunk diameter on a log–log scale [22].

As mentioned previously some palms have *adventitious roots* which support the stem, but these are not as large as *buttress* roots. These join the trunk like angle brackets to the lateral sinker roots [6]. They are overbuilt and have a very large margin of safety compared to other trees. It is not known how close tree palms get to the buckling limit but in the palm *Socratea exorrhiza* (tribe Iriarteae) the trunk exceeds the buckling limit, the reason for which is unknown [3]. According to Gibbons [2], older trees of this type are supported by stilt roots.

The slenderness ratio is described as the ratio of tree height to trunk diameter. Figure 11 is a log–log plot of mean tree height vs. mean diameter for several dicot, coniferous and monocot mature trees which shows a linear relationship and the gradient represents the 'slenderness ratio'.

The groupings are distinct and interestingly palms are shown to be anomalous. The plot shows that for unit diameter, palms grow taller per unit diameter than the other trees shown on the graph (with annular internal structure). This is interesting as the paper by Winter [8] suggests that dicot trees sway more at canopy level than young, tall, Mexican fan palms when subjected to lateral forced convection. This, he says, is because the young palm is 'overbuilt' for reasons explained in

Section 4.1.3. Many palms especially of the pinnate and wide girth variety have no need to sway in the wind as the leaf material is very tough and flexible enabling it to move in the slipstream of the wind causing least resistance. Tree palms in a forest habitat need to be very strong and remain above the canopies of the other trees (emergent palms) as they only have one canopy to process the sunlight by the exposure of their leaves. Of the tree palms, these are nearly always the pinnate leaved varieties perhaps as a result of their ability to withstand more forced convection as detailed above. The extent to which the palm does this also depends on the smoothness of the leaf surfaces together with the shape and smoothness of the trunk.

Generally speaking, when designing a wood strut support (of dicot wood), the slenderness ratio must not exceed 50. The slenderness ratio of a young dicot tree is often 100, although this is reduced with time as it becomes more squat. For an Andean wax palm this can be 150 in a mature specimen. This may mean either that mature trees have no built-in safety factor or that their entire strength is greater than that of a dicot wood strut support. Such struts would usually be rectangular in cross section and consist of a portion of a branch or a trunk rather than the concentric structures of an entire tree trunk cross section. This arrangement may be a contributing factor to its strength as well as the fact that trees are normally tapered, although this is less so in the palms. When designing a strut, the slenderness ratio is calculated for each of the three strut axes and the largest value is used as the critical one to determine at what ratio buckling starts.

The slenderness ratio and Young's modulus values for a few structures approximating to cylindrical prisms have been estimated and are shown in Table 2. It can be seen that the Andean wax palm has a greater slenderness ratio than all of the other trees displayed, but less than for grasses including bamboo, fishing rods and scaffolding. Reed grasses appear most effective at supporting a high load per unit diameter. For hollow structures, two times the radius of gyration was used in the slenderness ratio calculations as it is a more comparable parameter than the diameter when relating it to the diameter of homogeneous structures.

4.1.3 The palm trunk

According to Winter [8], the Young's elastic modulus of the palm stem periphery and the lower two thirds of stem as a whole (30 GPa) is an order of magnitude higher than that of the crown and the rest of the stem (3 GPa). This has interesting consequences. In an experiment where the *Washingtonia* palm wood was replaced by homogeneous hardwood (oak) and then softwood (Douglas fir) whilst retaining the same external trunk dimensions, the crown in the latter two instances was found to move more under wind stress relative to palm wood [8].

Palms differ from dicot trees and *coniferous* trees in that they have a relatively uniform stem structure. In order for palms to reach such great heights, in effect the young palm is overbuilt mechanically and hydraulically to provide for the future requirements of support and fluid transport. The palm remains stiff during elongation by lignifying and adding cellulose to the cell walls without adding to the girth. This results in an increase in the density and the stiffness of the lower two-thirds and the periphery of the stem [8]. Palm trees however can bend right over in strong winds and can withstand even the strongest of hurricanes [6].

One of the ways to prevent cracks in a brittle material is only to use them in compression. For example in pre-stressed concrete the brittle concrete is held in compression by high tensile steel wires within. Another example is with toughened glass where the outer layers are put into compression at the same time as the inner part is in tension. This latter structure does not appear to occur in biological materials which all depend on reducing the stress concentrations at the tip of cracks as with some man-made materials.

The vascular fibres situated in the periphery of the palm stem, together with those of the leaf bases operate together like a series of intertwining guy wires which transmit compressive stress

Table 2: Table showing approximate values for Young's modulus and slenderness ratio for various materials in order of increasing slenderness ratio.

	Slenderness ratio	Young's modulus (E)
Roman column (Rome, Pantheon – limestone)	12	32
Giant sequoia	22	Not found
Bone (human femur)	25	20
Red oak	27	11
Western red cedar	29	Not found
Douglas fir	34	9
Californian redwood	56	Not found
Socratea exorrhiza	133	Not found
Andean wax palm	150	Not found
Scaffold tube (structural steel ASTM A36)	193	200
Fly fishing rod graphite	302	220
Bamboo	400	Not found
Reed canary grass	949	Not found

into tensile stress throughout the whole trunk. This causes any stress to be dissipated throughout a large region [4].

Cellulose, which is the main constituent of wood, is very tough and constitutes chains of glucose molecules. They take the loads and provide the strength in the cell walls [14]. Cellulose molecules are the same in all plants even though the shapes and functions of plants vary. Initially, the simple sugar glucose is synthesised in leaves from the carbon dioxide in the atmosphere with water by the sunlight reacting with the catalyst chlorophyll. On a glucose molecule there are five hydroxyl groups which easily combine with water forming a solution. In the plant, the glucose molecules form a dilute solution with the sap water until they reach the growing cells. There, the glucose molecules join together by condensation and expel the water into the sap. The cellulose chains in the cell wall are long and nearly parallel to the fibre or cell wall which is in the direction of the applied stress. The chains form a steep helix ranging from 6° to 30° from the vertical whether it is clockwise or anti-clockwise depends on the plant and is curiously consistent throughout the plant. When composites are engineered however, this is not thought optimum. It is interesting to note that the cellulose is laid down in cells already containing stresses and strains which it has to bear. The cellulose chains do not branch and are thread-like [14].

4.1.4 The palm petiole

The flexural stiffness of the petiole is scaled to the size and weight of the leaf which it supports by virtue of its cell structure rather than its cross-sectional area. Large leaved petioles tend to contain a greater number of thick walled cells near to the petiole/air surface boundary compared to small leaved petioles. These outer cells in both cases enclose a core of thin walled *turgid* cells. Thus the flexibility of the petiole depends both on the stiffness (E) and the internal and external

geometry of the composite material. Leaf weight is another design constraint in the stiffness of petioles.

The external geometry of petioles varies in morphology and internal structure between palm types and is an indication of their structural function and flexibility. With time in some species, the adaxial surface appears to grow more flat and in some cases convex to nearly circular.

The leaves of the palms supported by the *Hyphaene* type petioles are thought to have the largest area compared to those of the other palms and the petioles are correspondingly larger.

4.1.5 Structure of the palmate leaf

The V-shaped grooves making up the palmate palm leaf give it much less flexibility than if the blade was flat for the same stiffness (E). This geometrical arrangement of corrugation provides greater resistance to snapping and enables the petiole to support a much greater area of supported leaf than a supported flat leaf of the same area. The size and weight of these blades are apparent in Fig. 5.

The ribs which are located at the vertices of the corrugations on a palmate blade can be internal or protrude from the epidermis of the blade to varying extents. Sometimes the ribs are raised on both or one of the abaxial or adaxial surfaces [20].

4.2 Fluid mechanics and heat transfer in palms

The two main fluid transport processes in palms are mass transfer, also called *evapotranspiration*, and heat transfer, which are physically analogous.

4.2.1 Mass transfer in palms

4.2.1.1 Capillary action Water and nutrients are transported from the roots to all other parts of the palm via the *xylem* 'tubes'. The xylem in monocot trees must be robust as most of it operates for the life of the tree unlike a dicot tree. Surface tension of the solute, adhesion to the xylem walls and the cohesive forces between the molecules of the solute and the root pressure are the four forces that make up capillary action and the narrower the tube the higher the solute will rise all other things being equal. The solute is brought into the roots through the process of osmotic pressure and effectively 'pulled out' in tension usually through the leaves by the process of transpiration. So for a continuous section of xylem, the column of solute is compressed at the bottom and under tension further up the stem. The diameter of the vessels is related to this. The work of Tomlinson *et al.* [23] states that in the stem the protoxylem (tracheids) and metaxylem (vessels) in the rattan palm, *Calamus* are not contiguous like in other palms which suggests that water can only move from the metaxylem to the protoxylem and so into the leaf across a hydraulic resistance which may minimise cavitation of the vessels and may be associated with an unknown mechanism which refills embolised vessels. The ends of the vessels carrying the solute do not act as valves and cavitations can occur but these may be re-charged, the mechanism by which varies with palm species. According to Tomlinson and Zimmermann [24], *Desmoncus* has a vascular system that is more continuous than Calamus but the arrangement of the vessels causes high hydraulic resistance in the axial xylem. They both have one large diameter metaxylem vessel in each central axial bundle without any vascular connections with *protoxylem* and metaxylem tracheary elements. Like tree palms, the stem vascular bundles in *Desmoncus* are continuous and branch out to each leaf. Associated with this are many bridging connections. According to Tomlinson *et al.* [24], the climbing stems of Calamus can reach lengths of well over 100 m,

yet their vascular tissue is entirely primary. This implies that the vascular system of the stem is efficient and resistant to hydraulic disruption.

4.2.1.2 Viscosity and fluid friction The causes of fluid friction are similar to those responsible for friction between solid surfaces and it depends on the nature of the fluid and the nature of the surface over which the fluid is flowing. In the case of the palmate palm leaf, the surface is a mobile corrugated-like surface and the fluid is air (which may be hot or cold and therefore vary in viscosity). Friction is also affected by the relative velocity of the air on the palm leaf.

4.2.1.3 Pressure, energy and the Bernoulli effect For horizontal fluid flow, an increase in the velocity will result in a decrease in the static pressure according to Bernoulli's law [25], which is how lift is produced by an aerofoil. Therefore the effectual convex, adaxial surface of the palmate leaf, may well act as an airfoil with the lift contributing to the support of the often very heavy and cumbersome palmate leaves. The effect of gravity acting on horizontal palmate leaves together with the often torn and thin edges of the leaves creates a convex upper surface.

4.2.1.4 Pre-stressing, hydraulics and Pascal's principle Tree palm trunks contain high pressures and are pre-stressed. Pascal's principle states that when there is an increase in pressure at any point in a confined fluid, there is an equal increase at every other point in the fluid. However, it is not known exactly where in the palm trunk these high pressures are distributed. Because of osmosis and the many branches connecting the fluid tubes, the plant 'system' contains many 'leaks' and there is a variable throughput of fluid depending on the climatic conditions.

Buckling in trees is prevented by pre-stressed trunks and branches. As a result, wood is only about half as strong when compressed as when stretched. In dicot trees new wood cells are laid down on the outside of the trunk in a fully hydrated state and as they mature, the cells dry and shrink or shorten. As these cells are attached to the wood inside, they are held in tension. This results in the outer part of the tree being held in tension and the inner portion of the tree being held in compression. Therefore, when the trunk is bent over by the wind, the wood cells on the concave surface are not compressed but some of the pre-tension is released [6].

Not so much is known about pre-stressing in palms, there are dead cells in the trunk but these would normally be hydrated.

4.2.1.5 Retention and loss of water in plants Plants regulate water loss through their stomata. The pore or stoma is controlled by two guard cells. Changes in turgor of the guard cells regulate the aperture [6]. Often there is water-resistant wax on the leaf *cuticle*, but this does not cover the stoma. All other things being equal, *transpiration* is greater when the *boundary layer* is thick (in turbulent air) compared to when it is thin as in laminar air flow as there is more circulation of the air. This enables more water to be transported away per unit time and the structure of the leaf wax as well as that of the epidermal cells and overall morphology of the leaf would affect this. If the insolation produces temperatures over the melting point of the wax, the nature of the wax will change. This seems to be rare.

Plants growing in conditions of high humidity with little wind have a thin boundary layer and the wax tends to be smooth. Palms growing in these conditions often have palmate leaves and can be found in an environment in the understorey of forests where the insolation and forced convection (strong winds) is low.

4.2.1.6 Uses of transpiration Of the transpired water passing through a plant only 1% is used for growth. The process of *transpiration* transports water from the soil into the roots and carries

them to the leaves, stem and branches of the plant via the xylem tissue [6]. Evapotranspiration prevents the tissues from becoming overheated.

The water potential of plants may be quantified as the potential energy of water per unit mass of water in the system. It is the sum of four component potentials; gravitational, matrix, solute and pressure.

Gravitational potential is always present but is usually insignificant in short plants in comparison with the other three potentials. The magnitude of this depends on the position of the water in the gravitational field. It can be significant in tall trees such as tall palms.

Matrix potential is the force with which water is held in plant and soil constituents by forces of adsorption and capillarity. It can only be removed by force and so has a negative value.

Solute potential (osmotic potential) describes the potential energy of water as influenced by solute concentration. Solutes lower the potential energy of water and result in a solution with a negative value.

Pressure potential (*turgor* pressure) is the force caused by hydrostatic pressure and usually has a positive value. It is generally of minor importance in soils but of primary importance in plant cells.

The rate of evapotranspiration is controlled by energy availability (as it takes about 600 calories of thermal energy to change 1 g of liquid water into vapour), the humidity gradient, the wind speed immediately above the surface and the water availability within the plant and all of these controlling factors are interrelated. If the wind speed increases significantly, more water is taken up by the roots if available, but if not, the stomata on the leaves would close to maintain water levels within the plant mainly for turgidity support.

4.2.1.7 Relationship between mass transfer and the palmate palm leaf The large scale mass transfer is achieved in palmate palm leaves via the water conducting elements called the xylem vessels. These occur together with the nutrient carrying phloem tubes as vascular bundles within a hardened protective bundle sheath composed of sclerenchyma cells. The widest diameter vessels are contained within the ribs.

The ribs at the apices of the corrugations tend to be covered with thinner wax compared to that on the main body of the leaf on palmate palm leaves, so heat may be released more readily from these parts (Fig. 12). As wax is immiscible with water (rain or mist), it helps prevent water rot developing in the main body of the leaf. The *Copernicia prunifera* palm, known mostly for its famous wax has several types of wax on its broad palmate leaves. These types of wax may be found in different locations on the leaf and petiole and is used to make candles as it has a high melting point.

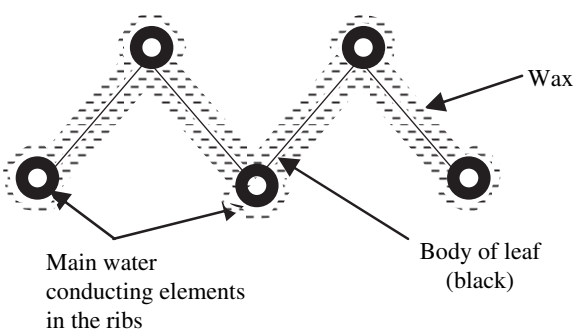


Figure 12: Schematic diagram of the cross section of a palmate palm leaf showing the angular structure which could produce areas of light and shade in directional sunlight.

4.2.2 Heat transfer in palms

Classically, there are three modes of heat transfer, namely radiation, conduction and convection. It is unusual for only one mode of heat transfer to be present in a given situation and often all three modes are present. In general, for vegetation, we have solar radiation and wind-driven convection. For a given leaf area there is also conduction from the surrounding structure.

4.2.2.1 Effects of solar radiation on palms All of the wavelengths in the electromagnetic spectrum apart from gamma and most X-rays are transmitted through the atmosphere to plants. These latter types are the shortest of the wavelengths absorbed by the atmosphere. The leaf cuticle is translucent so that photosynthetic active radiation (PAR) can be passed through to the chlorophyll beneath.

PAR is defined in terms of photon (quantum) flux, specifically, the number of moles or photons in the radiant energy between wavelengths of 400 nm and 700 nm. These wavelengths occur within the visible light band and partially in the infra red and ultra violet bands.

UV light is filtered out by some plant species via sunscreens and as a result, plant cells are less likely to be destroyed by the UV light so some plants can filter detrimental wavelengths.

A body which is light in colour (as seen by the human eye) reflects light and heat more readily than one of a dark colour. A light-coloured body therefore is more 'heat stable' and does not as readily change temperature to match that of the incident temperature. Dark green leaves, which have a high density of green coloured chlorophyll, tend to be found in shady environments but at canopy level in an area of high light levels, leaves are usually lighter in colour possibly as a result of the chloroplasts being killed or the chlorenchyma layers being located deep within the leaf.

4.2.2.2 Conduction in palms As a general rule, plants best survive when their temperature is evenly distributed and the critical levels for protein structure are not exceeded. This is achieved by the conduction of water through the veins and initially is absorbed by the roots via osmosis and drawn up the stem by capillarity, adhesive but mainly cohesive forces between the water molecules [6]. As well as being transported into the cells by osmosis, water conducts out to the leaves via the *stomata* where it is transpired and released. Heat conduction means that heat is transferred from warmer to cooler areas, the transport process being diffusive. In terms of plants, water is transferred from the roots and soil to areas which are hot such as the leaf immersed in sunlight.

4.2.2.3 Convection in palms *Natural convection* occurs when there are density differences in a gas or liquid which can flow [26]. When solar radiation is incident upon a palmate palm leaf, the angle of the corrugations which make up the leaf cause light and shady surfaces especially if the insolation is unidirectional. Forced convection is where the fluid movement is caused by external means usually a wind. Large leaves need to be able to spill the wind to prevent any breakage of the leaf or petiole and there are several mechanisms to facilitate this. As mentioned above, the petiole is grooved which allows the leaf and petiole to twist whilst retaining cantilevered support. The leaves are often tough and flexible (in the case of pinnate leaves) and can resist much of the wind action by positioning in line with the wind direction thus taking the line of least resistance. Palmate leaves by comparison are much more inflexible and tend to spill the wind at the terminal or distal ends of their leaves. Often these leaf margins are split and have flexible ends which become aligned with the wind perhaps stabilising the leaf. Seaweeds are good examples of organisms with leaves that effectively allow wave spill, their weight being supported mainly by the seawater, allowing the leaves to grow longer than they would in air. Forced convection effects *transpiration* and therefore cooling since it supplies unsaturated air below the stoma normally found on the

underside of the palm leaves. This allows water to be evaporated into the air more readily and as a result more water is pulled into the plant by cohesion of the water molecules via the roots. Even the slight doming of epidermal cells would enhance convection because of the increased roughness via increased air turbulence. *Micro-papillae* and striations and their correlation with plants exposed to heat and wind stress would also modify the boundary layer to some extent.

5 Conclusions

In this review, we have studied the processes by which tree palms survive. In so doing, a series of engineering-related projects may be identified. More specifically, the description of the extreme values of the features of some palm leaves and trunks, from both a microscopic and macroscopic perspective has instigated the idea of testing many of the underlying concepts. Although the tree palm provides us with an unusual example of where evolution seems to bring about a reduction in macro structure complexity, it is probably the case that some tree palms have not changed much through evolution and have not in fact evolved to become 'simpler'. The fact that palms have existed for so long is proof that they have been successful in terms of the continuation of their gene pool. A by-product of evolution is often the presence of excess material that is not critical to the survival of the organism from which it is a part. The palm operates with the controlling design criterion of 'survival' of the species. It should be a robust design in the early stages of development, in the establishment phase and in maturity over the long period of a lifetime and also for reproduction.

6 Glossary

abaxial surface	The lower surface of a bifacial leaf
adaxial surface	The surface of a leaf or bud adjacent to the stem
adventitious root	Root forming on stem, root or leaf
boundary layer	The layer of fluid in the immediate vicinity of a bounding surface
buttress	Flared base of certain tree trunks
conifer	A cone-bearing tree belonging to the largest division of the gymnosperms
costapalmate	Shaped like the palm of a hand with part of the leaf stalk extending into the leaf blade
cuticle	A layer of fatty material (cutin) covering and partially impregnating the walls of the epidermal cells of the stem and leaf
dicotyledon	One of the two groups comprising the flowering plants; the dicotyledonous embryo has two cotyledons (seed leaves)
evapotranspiration	Loss of water by evaporation from the transpiration from plants
forced convection	Where the motion of the fluid is imposed externally
gymnosperm	Seed plants in which the ovules are not enclosed within an ovary
insolation	Exposure to the rays of the sun
meristem	A tissue primarily concerned with growth and cell division in an organised manner
metaxylem	The last-formed region of the primary xylem which matures after the organ has ceased to elongate

monocotyledon	One of the two groups comprising the flowering plants; the monocotyledonous embryo has a single cotyledon (seed leaf)
natural convection	Where the motion of the fluid is caused internally by density differences
neoteny	Retention of juvenile characteristics in the adults of a species or the attainment of sexual maturity by an organism still in its juvenile stage
pinnate	Leaf blades on one leaf stem arranged like the veins of a feather
palmate	Leaf blades on one leaf stem arranged like fingers on the palm of a hand
parenchyma	Unspecialised, highly vacuolated cells typically with only a primary wall of uniform thickness; it occurs as extensive regions of tissue in the pith, cortex and mesophyll of the plant body; secondary thickening of walls may occur
phloem	The main food transporting tissue of vascular plants; consisting in palms of the conducting sieve elements, companion cells, various types of parenchyma and sclerenchyma
pneumatophore	A negatively geotropic root projecting from the substratum; produced by trees living in swamp conditions and serving for aeration of the underground root system
sclerenchyma	A supporting tissue whose cells are commonly dead at maturity and posses thick, lignified secondary walls, as in fibres and sclereids
stoma (pl. stomata)	A complex consisting of a pore in the shoot or leaf epidermis which is surrounded by two specialised guard cells changes in their turgidity causes the opening and closing of the stomatal pore and thus controls gaseous exchange with the external atmosphere
tracheid	An elongated imperforate tracheary element with various patterns of secondary wall deposition
tracheary element	A collective term for the vessels and tracheids of the xylem
translocation	The movement of sugars and other organic substances throughout the vascular plant body via the sieve elements of the phloem
transpiration	The movement of water from the root to the shoot in the tracheary elements of the xylem and the subsequent loss of the water vapour from the leaf surface via the stomata or leaf surface generally
turgor	The state of fullness of a cell or blood vessel or capillary resulting from pressure of the contents against the wall or membrane
vessel	A series of vessel elements joined end to end by their perforated end walls
xylem	A complex tissue composed of the water conducting tracheary elements, parenchyma and sclerenchyma

References

[1] Uhl, N.W. & Dransfield, J., *Genera Palmarum: A Classification of Palms Based on the Work of Harold E. Moore, Jr.* The L.H. Bailey Hortorium and the International Plam Society. Allen Press: Lawrence, KS, 1987.
[2] Gibbons, M., *A Pocket Guide to Palms*, Chartwell Books, Inc.: New Jersey, 2003.
[3] Tomlinson, B., Lecture at Miami University, Florida, 27 February 2004.
[4] Niklas, K.J., *Plant Biomechanics: An Engineering Approach to Plant Form & Function*, 1st edn, The University of Chicago Press: Chicago, 1992.
[5] Pearson, L.C., *The Diversity and Evolution of Plants*, CRC Press, Inc.: Florida, 1995.

[6] Ennos, R., *Trees*, The Natural History Museum: London, 2001.
[7] Takhtajan, A., Neoteny and the origin of flowering plants. *Origin and Early Evolution of Angiosperms*, ed. C.B. Beck, Columbia University Press: New York, pp. 207–219, 1976.
[8] Winter, D.F., On the stem curve of a tall palm in a strong wind. *SIAM Review*, **35(4)**, pp. 567–579, 1993.
[9] Shinozaki, K., Yoda, K., Hozumi, K. & Tira, T., A quantitative analysis of plant form – the pipe model theory. I. Basic analyses. *Japanese Journal of Ecology*, **14**, pp. 97–105, 1964.
[10] Fahn, A., *Plant Anatomy*. 4th edn, Pergamon Press, 1990.
[11] Elliott, K.A. & Shirsat, A.H., Extensions and the plant response to tensile stress. Society for Experimental Biology Conference. *Journal of Experimental Botany*, **49**, p. 16, 1988.
[12] Fahn, A. & Cutler, D.F., *Encyclopedia of Plant Anatomy – Xerophytes*, Gebrüder Borntraeger: Berlin, 1992.
[13] Barreveld, W.H., *Date Palm Products*, FAO Agricultural Services Bulletin No. 101, Food and Agriculture Organization of the United Nations: Rome, 1993.
[14] Gordon, J.E., *The New Science of Strong Materials*, 2nd edn, Princeton University Press, 1988.
[15] Frühwald, A., Peek, R. & Schulte, M., Utilization of coconut timber from North Sulawesi, Indonesia. Technical co-operation between the Federal Republic of Germany and the Republic of Indonesia, Hamburg, February 1992.
[16] El-Mously, H.A., Zamzam, M.A. & Ibrahim, N.H., Poster 310: Mechanical Properties of Date Palm Leaves' Midrib (DPLM) in Relation to its Utilization as a Substitute for Solid Wood, P5.04-00. Production and Utilization of Bamboo and Related Species, Hamed El-Mously, Center for Development of Small Scale Industries and Local Technologies, Abbasiah, Abdou Basha, Cairo, Egypt.
[17] Dorf, R., *Engineering Handbook*, CRC Press: New York, 1996.
[18] Burgess, S.C. & Pasini, D., *The Structural Efficiency of Trees*, The University of Bristol, 2003.
[19] Cameron, J.R., Skofronick, J.G. & Grant, R.M., *Physics of the Body*, 2nd edn, Medical Physics Publishing: Madison, WI, 1999.
[20] Tomlinson, P.B., *The Structural Biology of Palms*, 1st edn, Oxford Science Publications: Oxford, 1990.
[21] Rich, P.M., Helenurm, K., Kearns, D., Morse, S.R., Palmer, M.N. & Short, L., Height and stem diameter relationships for dicotyledonous trees and arborescent palms of Costa Rican tropical wet forest. *Bulletin of the Torrey Botanical Club*, **113**, pp. 241–246, 1986.
[22] Niklas, K.J., *Plant Allometry – The Scaling of Form and Process*. The University of Chicago Press: Chicago and London, 1994.
[23] Tomlinson, P.B., Fisher, J.B., Spangler, R.E. & Richer, R.A., Stem Vascular Architecture in the Rattan Palm Calamus (Arecaceae-Calamoideae-Calaminae). *American Journal of Botany*, **88(5)**, pp. 797–809, 2001.
[24] Tomlinson, P.B. & Zimmermann, M.H., Stem vascular architecture in the American climbing palm *Desmoncus* (Arecaceae-Arecoideae-Bactridinae). *Botanical Journal of the Linnean Society*, **142(3)**, pp. 243–254, 2003.
[25] Watson, K.L., *Foundation Science for Engineers*, 2nd edn, Palgrave: New York, 1998.
[26] *Dictionary of Engineering*, 2nd edn, McGraw-Hill Companies, Inc.: New York, 2003.

Chapter 11

The human world seen as living systems

J. Field & E. Conn
Fellows of the Royal Society for the Encouragement of Arts, Manufacturing and Commerce (RSA), London, UK.

Abstract

The subject is introduced by reference to work begun for the RSA's inquiry into Tomorrow's Company. Ways of thinking are presented that enable a holistic view of human social organisations to provide insights helpful in the understanding of an organisation's development. Seeing human institutions as living systems permits many analogies to be made with systems found in nature, and lessons to be learnt from them. After starting with the company the authors go on to look at the wider human society in this way, at the same time considering the influences of the human factors not present in otherwise analogous biological systems. Aspects of the human world that are treated include social change, democracy and justice, globalisation and local communities, as well as companies.

1 Introduction

It is not only in engineering design that lessons may be taken from nature; many useful insights may also be obtained through viewing human social, political, commercial and financial organisations as living systems analogous to those found in nature. In this chapter we shall look at examples of how this can be done, by bringing together ideas that a small group of RSA fellows and friends (see Acknowledgement) have been exploring at meetings and networking discussions during the past 10 years. But first, a word about the RSA whose inquiry into 'Tomorrow's Company' prompted the convening of the group in 1994.

2 The RSA

2.1 History

In 1754, William Shipley, drawing master and philanthropist in Northampton in the English Midlands, was inspired to create what is known today as the RSA. The RSA is located in the heart of

London in the Georgian house at 8 John Adam Street, designed especially for the Society by Robert Adam in the early 1770s. There are now over 20,000 RSA fellows worldwide – professionals from all walks of life – who have a keen interest, and a track record, in finding creative solutions to the problems of the contemporary world. It all began by offering prizes for inventions and designs to solve current problems. For a quarter of a millennium since then the RSA has played a stimulating and innovative role in a wide range of social and economic affairs. Over that time it has initiated numerous projects which left legacies that are part of the British national fabric today, including organised national examinations and design awards, the 'blue plaques' scheme in London commemorating homes of famous people, the Great Exhibition in 1851, and the Festival of Britain to mark its centenary in 1951. One hundred and fifty years ago it began the lecture series which still continues today, stimulating debate about the underlying issues of the day as well as their practical solutions. Over the years, it has attracted into membership, as Fellows, many famous and influential people – Karl Marx and Charles Dickens among them.

Shipley was motivated by the grim state of the poor in England when the Industrial Revolution was taking off. Decades later, Charles Dickens, in his novels, captured the misery and destitution of the times when only a small minority of people became rich on the fruits of the Industrial Revolution and the Empire. Now, 250 years later, when much, though not all, of that destitution has disappeared for the 1 billion people in the industrialised countries, many of the other 5 billion people across the world endure similar conditions as the poor of London did when the Society was founded. The threefold focus of the RSA in arts, manufactures and commerce, equipped it well to understand the links and connections between them, and the vital role of business and trade in improving life conditions. By the 1990s there was widespread disquiet about the role of business in contributing to environmental destruction and poor social conditions, both domestically and globally. In this climate, RSA fellows, led by Professor Charles Handy, began to ask 'What is a company for?', a debate which gave rise to a new RSA inquiry on 'Tomorrow's Company' in 1993.

2.2 The Tomorrow's Company inquiry

Concerned at the relative decline of British business, 25 UK business leaders led the inquiry. Alongside the work carried out by the business leaders, RSA fellows set up several groups to look at separate particular aspects of the problems. RSA fellow Eileen Conn saw the need to have an additional group to focus not on the parts but on the whole integrated system of a company, and convened a group to look at the 'company as a living system'.

This provided an integrating perspective for the many different aspects of corporate structure and behaviour being studied in the inquiry. There was a particular need to show how Tomorrow's Company would relate to the RSA's vision of 'a civilised society based on a sustainable economy'. The role of the company in such a context, as indeed the context itself, is difficult for conventional thinking, based on mechanistic models, to handle adequately. The belief was that then recent developments in understanding the way complex living systems operate and evolve might offer helpful analogies and insights.

Following the conclusion of the Tomorrow's Company inquiry, the RSA supported the creation of the Centre for Tomorrow's Company to carry forward this work and separately the Living Systems Group decided to continue, and extended its focus to society as a whole. In its discussions it has considered a wide range of issues ranging from the globalisation of industry and the economy to local communities. The special characteristic of these discussions has been their focus on using a living systems perspective to examine the issues and think about potential solutions. During the 10 years over which these discussions have taken place there has been continued development,

both in the group itself and in academia, in this perspective and what it means to take a living systems approach.

3 The living systems approach

3.1 Ways of thinking

The approach is, in practice, a way of thinking. Much of our thinking uses images and metaphors. It is, in large part, the way language itself has evolved. We are rarely aware of how much our language and thinking depend on metaphors and analogies and how significant these are in determining our view of things. They can either open up new horizons or narrow our field of vision. Some of the images prevalent today are of recent origin, flowing from the development of science and technology in the last few centuries. Based mainly on the concept of the machine, they limit to that extent our ways of looking at the world.

Mechanistic ideas have their place, and will continue to be useful. But using a living systems perspective shifts our images deliberately into those which derive from the expanding understanding of how living systems live, operate and evolve, and opens up new possibilities to meet the challenges of the modern age. For simple illustrations of some of the implications of using machine and living images in our thinking consider taking apart a clock and a plant. Having done so we can examine all the parts of each in detail, but the clock parts will remain the same and can be put together again to recreate the clock, whilst the plant parts will start to change and decay immediately and cannot be put together again. If we set a machine in motion, and release a bird into the air, we can calculate and predict the movements of the machine with accuracy, but not that of the bird. This shows the difference between linear mechanics and non-linear dynamics, the importance of the latter now being understood in many disciplines [1]. John Urry, professor of sociology at Lancaster University, considers that the theory of non-linear complex systems has become a successful problem solving approach in natural sciences, and it is now being used in social science thinking [2]. Human social and organisational systems are in many respects much more like non-linear living systems than like linear machines, but they are mostly designed and managed as if they were machines.

3.2 The holistic approach

Sometimes the non-mechanistic way of thinking is referred to as holistic and is concerned with the dynamics of holistic systems. But the main focus here is to identify the parts of a wider system and reconnect them; in organisations this is often called ‘joined-up’ thinking. While this is essential if we are to understand the nature of living systems there is much more to it than that. We also need to pay attention to the dynamics between the interior of the system and its environment. This is particularly important for human systems where the inner emotional, psychological and spiritual life of the individuals significantly determine their external behaviour, and thus the behaviour of the groups of which they are a part. So even with apparently ‘joined-up’ thinking, social, economic and organisational policies can fail to deal with these dimensions and consequently produce many unintended effects without realising their aims. The essence of holism is that systems are made up of other systems, with systems of different scales embedded within larger systems all the way from the smallest to the largest on a cosmic scale. At every level of this scale there is an interior life of the system comprised of the collective life of the smaller individual systems embedded within it.

These holistic systems embedded within other holistic systems, sometimes called 'holons', are also known as complex systems. The recently emerging new sciences studying the dynamics of complex systems are very fruitful in increasing understanding of their behaviour. Because it is non-linear, the behaviour of the complex system emerges over time, which is why it is not predictable in the way a machine is. A further distinction is drawn between complex living systems and other complex systems to refer to complex adaptive systems or complex evolving systems. It is only systems which are living and have an interior life which can adapt behaviour through experience. The science of complex systems, both living and non-living, is producing greater understanding of these dynamics, and gradually producing new imagery to use in thinking about human systems.

3.3 Living systems

Taking a living systems perspective is to use these new and emerging insights in looking at the operation and evolution of human social systems. Treating the human social system of relationships as no more than a machine misses the essential nature of the system. Mathematical models of the economy, or any other social system, which include only that which can be quantified, have their place, but our ideas of social systems must expand to include all key relationships in society and their relationship with the environment.

Thus, in our approach, living systems are seen as being almost infinite in their variety and complexity. Any one system will have a place in a hierarchy of systems at different scales, and experience complex relationships both internal and external to itself. The quality of such relationships and the system's interaction with its environment are important in determining its health and survival. The study of its relationships and of its adaptive capacity offers a useful means of assessing the health and evolutionary prospects of a living system, where analyses using quantitative models are inadequate and likely to be misleading by themselves.

Observation of natural systems can provide clues as to how our human organisational systems might evolve. As we now move on to look at companies we shall make frequent use of analogies with biological systems to provide us with such insights. However, as we take this approach, it is important to keep in mind that no analogy is exact and to recognise where the differences lie. It is also important to appreciate that we are dealing in analogies and not advancing natural laws of, say, biology as being applicable to human society, as some political theorists have done inappropriately and with disastrous results.

4 Companies

This section on companies is based on the paper produced for the Tomorrow's Company inquiry by the living systems group of the RSA inquiry team. It illustrates the wide-ranging use which the living systems perspective can have in thinking about companies.

4.1 Evolution and adaptation

Systems and their subsystems develop as a result of an evolutionary process in the general sense. In nature we see both individual organisms and the ecosystems which they inhabit, and which generate and sustain them, developing in this way. It is the capacity to adapt that ensures the long-term sustainability of natural systems. In research concerning these issues it has been shown that the capacity of a living system to adapt and change resides in micro-diversity of the populations

that inhabit it. There is a short-term cost of maintaining micro-diversity, since at any given time it implies that there are many sub-optimal individuals, judged from any particular criterion of efficiency. However, in the longer term, evolution will select for populations with the resulting adaptive capacity. Studies show that although competition is instrumental in driving evolutionary process, what emerge are co-operative groups of synergetic activities, whose activities favour each other. Uncertainty about the future means that no particular evolutionary scheme will inevitably be successful over time, and that a portfolio of parallel 'strategies', arising from rich and mixed cultures, is needed for success [3, 4].

The lesson for the management of companies in this analogy is that for their existence to be sustained it is necessary for companies to be made adaptable, through the encouragement of diversity in their constituent systems, the fostering of the internal relationships that enable particular constituents to fulfil themselves at the appropriate time and the fostering of external relationships that keep the companies in harmony with their environments.

On the other hand there is an implication that continuous existence is not the destiny of all systems at all scales. Many go through a life cycle of birth, life and death with some continuing the process with rebirth.

4.2 The cycle of life

Observations of nature and the history of human civilisations show this cycle of life to be the norm rather than the exception. Indeed where evolution can be seen as providing a process of continuous enhancement of life we may only be looking at a short segment in a large cycle. Life on earth itself may be destined for such a cycle, but such a large one in relation to our individual life spans as to be beyond any consideration other than by the purely academic.

Global sustainability does not require that all constituent parts be sustained. Nations may rise and fall and companies may come and go, but, as we have already noted, adaptability provides the option of a continuous, evolving existence if not the conservation of a particular character. Some companies may opt for a limited existence to fulfil a particular role for as long as the need exists. Provided that this is fully understood in its internal and external relationships and their actions do not compromise the sustainability of the wider systems, such companies can be a healthy feature of a complex socio-economic system.

However, many, probably most, companies are established with the intention of continuity and the fact that many do not survive for more than a limited period, perhaps spanning a generation or two, is due to a lack of evolutionary adaptation rather than of intent.

4.3 The sustainable company

The first feature to note when viewing a 'perpetual' company is that it has an existence as a legal person and characteristics of substance that are separate from, and are intended to outlive, any of its individual human members, such as shareholders, partners or employees. Many shareholders and employees remain in those roles for only a very short period of time whilst others stay for a lifetime or a working lifetime. With the many forms of shareholding and of employment that exist, it is difficult to identify the membership composition of a company with any precision. An analogy can be made with a human being who can be formally identified by identification documents or by thumbprints, has intellectual and physical characteristics of substance that are always developing, but whose body cells continually change. However, unlike mortal beings, a company has no inherent limit to life.

The only manifestation of a company that is constant, until formally altered, and precisely identifiable, is its legal entity. The substance of a company, however, lies in the complex and living system of resources, intellectual capabilities, skills, brands, relationships, markets, reputation, source of livelihood and fulfilment and other characteristics that change over time. When a company dies, as when a person dies, the loss to society lies in its substance rather than in its legal entity. The analogy can be extended if we wish to refer to the soul of a company to describe that indefinable characteristic that has a spiritual quality and adds to the substance.

When we wish to sustain a company, we do so for its substance, which needs to constantly evolve in response to changes in the environment, rather than for its legal identity, which need not change, or the individuals who come and go.

The purposes of companies and of their continuity can be open to different understandings when seen from different viewpoints. Shareholders may see the purpose as being to enhance the monetary value of their shareholding and their idea of adaptability might be more akin to that of a surgeon, always ready to cut out or to graft on, than to that of a physician tending to the health of the body as it is. On the other hand those working in or for the company may see it as maximising their fulfilment and earning capacity in the use and development of their skills and know-how within the company. Their view of adaptability is more likely to be that of the physician. The wider community in which a company operates is concerned with the quality and value of the contribution to the community and the effect of the company activities on the environment.

There is nothing incompatible in all these views taken together, but the balance and adaptability required for sustainable evolution will only be achieved through healthy and mutually supporting relationships.

4.4 Relationships

In discussing relationships involving companies the company is thought of as a living system having a legal identity and a substance. Shareholders, employees and outsiders have their relationships with the company but no one group is the company. Their relationships are based on reciprocal rights and duties with it and, from this viewpoint, it would seem more appropriate to think in terms of ownership of rights rather than of ownership of the company. Acceptance of such a line of thinking could greatly assist in the understanding of and development of the relationships needed for a sustainable company.

An important theme thus emerging from considering the company as a living system is that of the quality of relationships. All relationships are important whether they be internal, between the constituent parts of a company, or external, between the company and its environment, the socio-economic system in which it operates and other organisations within that system. The RSA Tomorrow's Company inquiry embraced the concept of the inclusive company; the living systems analogy can greatly help the understanding of the diversity of the relationships involved, and of the changes in our approach to them that are required. Let us now look briefly at three categories of these relationships.

4.4.1 Company–competitors–collaborators

Those who believe that a free market economy is the best prescription for a healthy economy and a better world for all see a high degree of competition between the constituents of the economy as desirable. On the other hand research in the field of ecosystems, to which we have already referred, has shown that evolution tends to favour co-operation although competition plays a part. Examples of co-operative relationships can be found in what has been termed industry clusters, defined by Michael Porter [5] as 'geographic concentrations of interconnected companies, specialised

suppliers, service providers and associated institutions in a particular field, that are present in a nation or region.' New clusters may arise from one or two innovative companies that stimulate the growth of many others. Porter tells us that Medtronic played this role in helping to create the Minneapolis medical device cluster. The best-known example of a cluster is perhaps the high technology cluster of Silicon Valley. It is claimed that healthy, outward oriented industry clusters are a critical prerequisite for a healthy economy. Clusters are dynamic; over time, existing clusters will transform and new clusters will develop from a region's talent and technology base.

4.4.2 Company–environment

By referring again to the current understanding of ecosystems, it is seen that for a successful and continuing existence an organism needs to be part of the ecosystem of its environment and thus to fulfil its role in that system. When conditions change, the organism with adaptive characteristics will then be able to evolve with the ecosystem of which it is part. Growth beyond the capacity of the environment to sustain and be sustained leads to collapse. A rapidly breeding locust does not of course have the intellectual capacity to understand this, but humans do and our relationship with the environment depends on our use of it. In fact, to extend the analogy, plagues of locusts represent temporary explosions of population probably in response to chaotic variations in long-term stable ecosystems, and human activity aimed at maximising production can cause destabilising variations in an ecosystem. For human intelligence to be used as a stabilising rather than destabilising factor for the environment on which we depend, the motivations within the socio-economic system need to be turned in favour of the longer term and wider view.

4.4.3 Company–people–society

These relationships are between systems at different scales. People and companies are parts of society and people are also constituents of companies. In nature, healthy living systems reflect and are reflected in the systems at different scales of which they are part or which are part of them. For example, a whole tree, a branch of that tree with its attendant foliage and each leaf will be in harmony and have forms reflected in each other. In human systems stress resulting in the eventual collapse of the system will occur when there is no harmony between the values, beliefs and resulting behaviour of people as individuals, the organisations to which they belong and society in general. The requirement is for vertical harmony rather than lateral uniformity. There is a rich diversity in the human race, and micro-diversity has already been seen as a virtue in living systems, which should be reflected in society and its organisations. The relationships between people, companies and society should be such as to ensure this combination of harmony and diversity.

4.5 Companies in the wider world

In this section we have focused on companies as living systems in a way that helps us to understand better the behaviour of these particular forms of human organisation. It has given us a holistic view of companies so that in considering their future development and place in society we may take due account of the complex and intangible factors that are all important influences. But in taking this approach it has not been possible, and would not have been useful if it had been, to consider the company in isolation from society as a whole. It is presented as an example of the use of an approach rather than as any definitive statement. Many problems of human organisations, whether concerning local communities or the global society, can be looked at in a similar way to provide a holistic understanding of how they function. But when this is done there are no

well-defined boundaries, only a focus on a particular part or a particular aspect of a large whole. In the following sections we shall broaden the approach to relate to society as a whole.

5 Changing society in the modern world

Throughout history there have always been thinkers and writers with ideas for changing the social order of humanity for the better, whether nationally or on a global scale. The need for change has been seen in the failures of existing systems and processes to serve the needs of members of society and to respond to new threats and opportunities that face us. Religions and philosophies offer different perspectives on the meaning of life, the mystery that is always with us, and although their moral teachings may be aimed at the individual, many seek to apply them through the social structure. Despite a diversity of faiths and beliefs in the world there is probably greater agreement on right and wrong than a simple observation of the rich diversity of cultures might suggest. But it remains that the qualities of a society ultimately depend on the qualities that are developed in the individuals that make up that society and particularly in those that lead it. Whatever our criteria for the ideal may be, a society ideally structured in theory (e.g. a just system) will always disappoint in practice to the extent that those managing it and operating in it fall short of the ideal (e.g. are unjust). On the other hand, if the culture were to come closer to the ideal (e.g. a culture of justice) it is more likely that an improved (more just) social system would emerge and be sustainable. Cultural change and systemic change need to go hand in hand, one being both helpful to and necessary for the other. Thus a social system cannot be engineered successfully. Rather it needs the nurturing of cultural conditions so that social systems which enable human society to meet human aspirations can emerge naturally.

When seeking to change society for the better, enthusiasm for our own particular scheme all too often stops us from seeing the need to ask some all important questions: How do we know that our scheme is the right one for present conditions allowing for the human frailties of those who will operate it? Is it adaptable to external changes? How does it fit in with those of other groups? The truth is that no one plan for the social order is wholly right or will work as intended given the organic living nature of the human society. What is more, movements for social change can very easily become power bases for the power hungry. The 20th century was the century of social ideologies, of blueprints for society that were to be imposed by force without question of their efficacy, and by individuals possessing the usual human frailties, including lust for power. This resulted in the tragedies of that century and the mood now is to find another way, although vestiges of the old approach can still be seen in the policies of some political groups. Not only ideologies but religions too have been hijacked by the power seekers, both in the past and today.

Taking society as a living system, lasting and sustainable change can be brought about only through the participation and consent of its members. Leaders can provide vision, motivation and articulation of the norms of behaviour, they can inform and facilitate, but they cannot direct what the collective free will would not of itself create. Nature evolves through the application of natural laws rather than by human direction. Human intelligence provides a facility for assessing the consequences of actions and of communicating information, but the lessons of history and the view of society as a living system lead us away from seeking authoritarian direction that can so easily turn to misdirection and create injustice.

The fall of Soviet communism has liberated thinkers from the categorising of left or right and this has resulted in much new and positive thinking. The world is not short of ideas for its improvement. The big question is: how do we ensure that the right idea comes to the fore at the appropriate time and place? There was an American management consultant who maintained that

whenever he was called in to solve a structural problem in a large organisation, there was always someone somewhere in the organisation who had the answer. His approach was to seek out that individual and then collect his fee. It is for the process of selection that we are drawn to look at nature in all humility. However, natural systems are without the intellectual capacity to reason and calculate. As we saw with companies, we have that capacity and how we use it, together with our feelings, emotions and spirit, is the human factor that needs to be added to the biological when looking at human society and institutions as living systems.

6 The human factor

When we looked at companies as living systems we made use of biological analogies. We saw human beings as living systems that are a part of companies and we saw companies as being a part of the wider human society. It was in this way that human characteristics were brought into our perspective. But now, when we want to look at that wider society and other elements of it, for which objectives such as social justice (not found in purely biological systems) are important, we need an understanding of the human factor that makes us different from other species.

Humans may not be the only species to have an intellect of some sort but the power of the human intellect sets *Homo sapiens* apart from all other species. With our intellect comes consciousness, language and culture, as well as the power to reason and to calculate. We seek the meaning of our existence and with our consciousness we have an awareness of the needs of others as well as of ourselves. How we act on that awareness depends on many factors: the culture developed in a society, emotions such as those of love and hate in individuals, and altruistic motivations for which some have found biological explanations (in sociobiology) and others theological explanations or a combination of the two [6].

It seems to be a natural human characteristic to have some basic bi-polar needs such as:

1. to love and to be loved
2. to understand and to be understood
3. to be unique and to be part of a larger whole
4. to see justice done and to be justly treated.

Whereas some would see these needs as a necessary part of the human condition as biological animals, there being a biological basis for a human spiritual nature, they can also be seen to support the views that humans are more than biological animals, and have a spiritual nature beyond biology. But, whether these characteristics are part of human nature biologically or part of a spiritual nature beyond biology, they give rise to a search for meanings to life, and an understanding of how that should translate into our relationships one to another. However it may come about, these human needs are an important factor in human society seen as a living system.

7 Democracy and justice

7.1 Democracy

The concept of democracy itself is not generally found in natural systems, neither is it to be found in the teachings of religious faiths whose moral teachings concentrate on the behaviour of individuals within a system rather than on the system itself. Yet there seems to be a fair consensus amongst people of goodwill around the world, thinking as individuals, that democracy is an ideal aim for the governance of the human community or, as Winston Churchill once remarked, the

worst form of government except all those other forms that have been tried from time to time. This last point suggests that, although for biological evolution the Lamarckian theory of learning passed from generation to generation has been generally discarded in favour of Darwin, it may still be relevant to the evolution of human social systems on account of our intellectual capacity.

7.2 Gaian democracies

In a recently published book [7], John Jopling and Roy Madron see a crisis in democracy and the need to articulate something that would be relevant in the 21st century. Their message is that rather than experiencing 'true' democracy we are at present being governed by a global 'monetocracy' which is unjust and unsustainable and needs to be replaced by a system that is in tune with the natural systems that have made possible sustained life on our planet.

Their view is that the monetocracy is seen to have the purpose of money growth in order to sustain the debt-money system. In following this purpose we are destroying the system on which we depend and to correct this we should, for human systems, look to the Gaian theory of the planet's physical, chemical and biological systems being self-organising and interactive. We should be looking at systems of government that interact symbiotically with Gaian systems.

Jopling and Madron argue that societies must see themselves as part of the Gaian system and the message concludes that we must make a transition from monetocracy to Gaian democracies. One key concept of such democracies is that of soft or purposeful systems of human society whose leaders, termed 'liberating leaders', operate a network governance through dialogue.

The above is a brief summary of the thinking behind Gaian democracies. It embraces the ideas of leaders facilitating rather than directing and of harmony with the natural world and the concepts of justice which stem from the human needs, discussed in Section 6, and of democracy.

7.3 Justice

Another book that deals with justice whilst at the same time views society as a living system is *Seven Steps to Justice* by Rodney Shakespeare and Peter Challen [8]. The book aims to look at justice as a whole although it focuses on monetary justice, an aspect often overlooked. Among the steps proposed are the introduction of healthier forms of money than interest-laden debt money and the distribution of income through the application of binary economics. This denies the Adam Smith view that labour is the source of all wealth, and seeks a partnership between labour and productive capital for a just, functional distribution of income. A particular interest for us about these proposals lies in the method of implementation. This would be to introduce healthier money and healthier distribution systems alongside the old, so that the overall system can select the better way. This of course requires there to be a selection process in the system that includes justice as a criterion as well as survival and enhancement. It is ensuring the presence of such a process in the system, rather than the detail of the ideas to be introduced, that is seen as all-important in the living systems approach. Shakespeare and Challen also believe the introduction of justice, in a way that can be supported by all the main religions in a region, to be the key to resolving conflicts in areas such as Kashmir and Israel. And this has to be a justice that includes the minorities.

8 Globalisation

Recent years have seen amazing advances in communications technology enabling the global spread of multitudes of ideas, people from different cultures, modern industry and environmental

destruction. But what is globalisation? The answer depends on where we come from. Here are three views:

- The stirring new awareness of global interdependence
- The economic manipulation of trade and credit to increase the power and material wealth of a minority at the cost of a vast and impoverished majority

and the eighteenth precept of Buddhism:

- 'I will consider myself forever on a journey and will see the whole earth as my true home.'

The second one of these is the negative view. Is it the cynic's view or the realist's view? Maybe some of both, but whichever it is those who hold it will either seek to destroy globalisation or, if they feel that it can be changed, want to do so. They would like to see more regulation of activities at the global scale and less regulation at the local level, the purpose being to inhibit activities that do global harm and result in the exploitation by global operators of local communities in which the operators have no interest. They wish to encourage activities which give benefit in the regions in which they are carried out.

In truth, striking the right balance between global interdependence and local autonomy is the kind of balance and harmony that any living system needs to achieve if it is to survive and develop. The global ecological and natural system itself constantly needed to adjust that balance before the world became dominated by human beings. Darwin was able to observe the very different courses that evolution took in parts of the world that were remote from each other. And yet they were all on the same planet and subject, perhaps in different ways, to the same global forces, of air and sea currents and climate change for example. Before human beings built ships or aircraft, birds and fish migrated long distances for the benefit of their own species and at the same time contributing to the ecologies of the regions that they visited. A balance of competition and co-operation and the maintenance of diversity provided the adaptability that the living system of the planet required. A migrant that destroyed local habitats and impoverished the local inhabitants would not itself survive, but those which benefited the local ecology as well as themselves would.

We shall now look at some thinking which embraces the living systems approach in a way that can help to support the positive and encourage change to remove the negative.

8.1 Global issues

In his book *Global Forces* Bruce Nixon [9], setting out with the basic belief that most people are good people, identified 10 important issues now on the global agenda. Starting with 'The ecological crisis and sustainability' and 'The increasing gap between rich and poor', issues that by their nature are on a global scale, the list continues with concerns and aspirations that have recently taken on a global dimension, due to the advances in communication technology that make people in all corners of the world instantly aware of happenings and thinking everywhere else. The concerns and the aspirations of others can no longer be ignored by those in a position to make a difference, through ignorance or because they are distant and out of sight.

In considering these issues we must recognise that everything has an upside and a downside. We are a mixture of good and evil, light and dark (rather than either one or the other) and there is a lot of good news to be found. The book recognises that the emerging thinking about organisations is moving towards viewing them as living systems and an acceptance of chaos as a part of transformation and creation. Understanding this is more likely to lead to positive outcomes. The message of the book is that 'we must get the whole system into one room' and to be a whole person or organisation we must unite body, mind, heart and spirit.

Nixon quotes Nelson Mandela: '. . . what is required is that the mutual respect that underlies the mere possibility of negotiation should always inform the way we relate to one another as representatives of different nations and different sectors of the world community. Such a change – for it would be a change! – would be part of building the new post colonial global order on the international system established some 50 years ago to ensure that the world never again experiences the destructive violence of economic crisis and world war. It would be part of democratising the world in which we live. It is a necessary condition for world peace and development' [10]. The quality of relationships between diverse elements is something that we have already observed to be of fundamental importance to the health of a living system and its ability to adapt.

8.2 Simultaneous policy

Nelson Mandela mentions negotiation. Another book that picks up on that theme, in particular how methods used in negotiation can be employed in removing the negative aspects of globalisation, is *Simultaneous Policy* by John Bunzl [11]. The objective of simultaneous policy is to turn destructive competition into constructive co-operation. The proposals are based on the observation that many necessary global measures, relating for example to the economy or the environment, are not taken by governments, when they might wish to take them, because it would put their countries at a global competitive disadvantage if others were not to do the same. The idea is for an international organisation to be established that would obtain commitment to such policies, conditional on them being implemented by all. When full commitment had been achieved all states would implement the policies simultaneously.

Bunzl also picks up post-Darwinian theories of evolution (that do not deny Darwin but see Darwin as only part of the story) which observe that natural systems when evolving to larger scales can do so only by co-operating at that level rather than competing. Bunzl sees the analogy with human systems as pointing towards the need to develop co-operative systems at global level for just those global decisions which the current competitive global economy makes impossible. Simultaneous policy is a way of trying to facilitate co-operation at a global level.

8.3 Charter 99

A different, though complementary, approach to world decision-making is that of Charter 99 [12], produced by a coalition of organisations committed to pursuing global democracy and addressed to the governments of the world. The premise is that world government already exists working through many different institutions such as the G8, The Bank of International Settlements, World Trade Organisation etc. The challenge is to make them democratically accountable in some way. Although many would like to see accountability through a democratically elected world body of some kind there are dangers in this and the promoters of the Charter prefer to move the process through intergovernmental conferences. Making change through involving existing systems, whatever their shortcomings, has the advantage of evolution through the fostering of microdiversity and the establishment of relationships, and avoids the perils of designing a blueprint which destroys the existing before it has been tested in time. Like adaptations in natural living systems it is slower but continuously maintains the system's adaptability for the future as experience and changes in environment may demand.

9 Local communities

The living systems approach can be applied to local community organisations much as we have seen it applied to companies. The organisations are within a wider society that provides the internal values, and the individuals within the organisations are also part of the wider society in their own right. The same requirement for facilitating relationships and compatible values held at the different levels of the living system apply to these organisations as they do to companies. This is illustrated below by reference to cases.

9.1 Police and Community Consultative Groups (PCCGs)

In the early 1980s in some inner city areas in Britain, there were disorders as a result of disaffection between young people, mainly of black and ethnic minority origin, and the local police. One such disturbance took place in Brixton, London, and the report of the inquiry which followed, the Scarman report, recommended that there should be new methods for the police to consult with local communities. This resulted in legislation requiring local police authorities to set up consultative committees involving local authorities and representatives of local communities. The aim was to improve communications and understanding between the police and local communities, especially young people. Eileen Conn was an active community member throughout these years of one of the Police and Community Consultative Groups (PCCGs), and chaired it for a number of years. She is also studying community dynamics from a complex living systems perspective, and suggests that how all this evolved illustrates some interesting aspects of the nature of these committees, and the social ecosystem which they inhabit.

In the London boroughs the PCCGs were set up by the Metropolitan Police, the responsibility of central government, rather than by the local authorities as elsewhere in the country. At the time, the mistrust of the police was not confined to young people in the population. In London there was also antipathy towards them in parts of local government too, with resulting breakdown in communications. This antipathy between local councils and the police was in some cases so great that the only way to get the new arrangements established was for the police and the representatives from the voluntary and community sectors to work closely together. It took several years for each borough to have a fully fledged PCCG. Because of this environment, where there was little recent tradition of close co-operation between police and local councils, the new arrangements had to create their own culture. In most cases, they were set up deliberately at arm's length from the councils and so acquired a culture of independence which was closely guarded by the representatives from the voluntary and community sector who had contributed to their successful establishment. And they each had their own identity and name as 'groups', rather than being 'committees' in the local bureaucratic establishments as was the case outside London.

This created in London a unique landscape and environment for their evolution, not unlike in some respects the unique environments that Darwin studied in the Galapagos Islands which were separated from the mainland. In the early stages, new and constructive relationships developed between the police and community representatives, but in many cases not with their local councils. However, the improved working relationship between the police and community representatives, and the pressure from central government, together with external political changes, eventually after several years stimulated a cultural change in the local councils. The next phase of evolution then was the development of 'crime reduction' or 'community safety' partnerships between the police and local councils in each borough. The strengthening of the relationships between the two most powerful public agencies in each borough shifted the police attention away

from the PCCGs, thus distorting the groups' 'fitness landscape'. The concept of fitness landscape is an idea developed [13] to understand the way in which organisms adapt to changes in their environment. An organism's fitness landscape becomes 'deformed' by external changes and it needs to make internal and external changes to reposition itself in a better fit with its new landscape.

PCCGs responded to the threat to their existence by accelerating their moves towards collaboration across London through a new London-wide group, thus creating a new and wider scale in the arrangements. This stood them in good stead when the central government made the police of London responsible to a local government Metropolitan Police Authority. This new authority attempted then to impose similar restrictions and culture on the London PCCGs as experienced by many of the committees in other parts of the country. This was resisted with a strength which would not have been available to individual borough groups on their own.

The way in which the three sectors evolved in their relationships over these 20 years is an illustration of the co-evolution which is characteristic of the evolution of organisms in their environment. Within the PCCGs, the three sectors – the community, the police and the councils – all inhabited the same landscape, and were intimately interconnected with each other, the actions of each directly affecting the others. The initial alliance between the community and the police stimulated changes in the local councils; this in turn created new alliances between the police and the local authorities in the form of 'community safety' partnerships, which deformed the landscape of the community groups; they in turn found strength though their collaboration with each other at London level, and had to be taken notice of again by the police and community safety partnerships.

The adaptations which organisms make in response to their environment can in turn affect their environments. This dynamic interaction, termed 'structural coupling', between an organism and its environment is fundamental to its nature as a living system which self-organises [14]. These dynamics can also be seen in the ways in which the three groupings, involved in the PCCGs, coevolved together over the years. The internal nature of the police and the local councils, while in many ways different from each other, had similarities. Each has a large organisation culture that is related to exercising considerable power and authority, and managing large public budgets, which enabled them to link together in formal 'community safety' partnerships. The voluntary and the community sectors on the other hand are very different. They are comprised of numerous small groups, many of which have almost no budgets, no staff, no power and no authority to exercise in the public arena. These groups reflect the diverse social, ethnic, and religious groupings in their local area.

One of the emergent properties in the PCCGs was the new relationships which were made from the connections between people from such diverse backgrounds, with a mutual interest in the community safety of their common geographical area. Working together in this way nurtures trust, respect and reciprocity, essential for healthy living communities. These new relationships were a significant development enabling the PCCGs to use their potential to make constructive initiatives, as well as enabling other local connections of importance in the evolution of a more socially cohesive community. They are part of, and themselves nurture, the 'well connected community' [15]. This provides a strong and resilient web of voluntary connections from which emerge the community groups to engage with the public agencies. These groups share similar patterns and structures, though each will be unique in relation to its own local context, being fractals in the complex living system that is the community.

A human community, like any other living social system, is comprised of individuals who are living systems, each with their own internal natures which are 'structurally coupled' with their environments. So the internal emotions, feelings, psychology and spirituality of those individuals, affect the way they behave in these networks and the quality of their relationships. This in turn

affects the shape and working of the emergent community organisations. Traditional approaches to community consultation and community development tend to begin with forming structures and organisations, often a mechanistic, engineering kind of approach. But seeing the ways in which communities behave like other living systems – with groups and structures emerging naturally from the multitude of connections within the community – shows that what is needed is more akin to human horticulture than to social engineering. Community networks need to be cultivated to be strengthened, rooted and extended, and need to be supported to nurture the sensitive relationships they create.

9.2 The Scarman Trust

Also arising from the Scarman report was the establishment of the Scarman Trust. The Trust aims to change the culture of people and institutions to encourage them to behave very differently in fields such as health, education, drugs, etc. The people in the poor communities on the ground have ideas of their own and the object is to release those ideas, put them to creative use and feed them back to the institutions, Scarman having said that institutions must listen to people. To perhaps a unique degree for such a scheme the operation is based on trust and the results have been stunning. Small grants, no larger than £10,000, and considerable practical support are given for a large number of projects around the country – over 1000 in 2001. In only two was the trust misplaced and the money lost.

Ray Sheath, who was Managing Director of the Trust for a number of years, has an interest in complexity and adaptive systems. However, he has an aversion to viewing things in a complicated way and he believes that beneath a complex situation and guiding a complex system there is something simple at work. Looking for the underlying simplicity is part of the thinking about complex adaptive systems. Complex adaptive systems may be complex but they work to very simple rules. When disturbed they may remain static, become chaotic or stabilise to order. They may be physical, biological or social systems. For example ants, bees and other community creatures are phenomenal in their complex behaviour but work to very simple rules.

The social systems of humans in which we are interested are even more complex than that of ants or bees and Ray Sheath looks for the underlying simple rules. For example, the capitalist system works by the simple rule that someone puts money in and wants money out and company law deals with that rule and nothing else, certainly not a multiple bottom line. Changing the simple rule would change the results through a complex process. Understanding the simple rules behind a system can show the way to facilitating alternatives to develop. Just as in medicine understanding the genome helps to find cures for physical ills, so in society understanding the simple rules that govern it helps to find answers to its problems. Such ideas need not just remain in nor even originate in the academic arena; they are inherent in the approach of practitioners such as Ray Sheath [16].

9.3 Community study in Poland

There can be little doubt that the world of the 21st century is one of rapid change and volatility. Speed appears as a key force in the course of events – in travel, in communications, in the impact of events man-made or natural. Management and the ability to cope with the world about us has become increasingly difficult. The realisation that we are, as individuals, community or corporate, living systems having to recognise our interrelationships and our interdependence with other individuals, communities or corporations is slowly gaining favour. It was in recognition of this that an action programme was implemented in Poland by the International Business Leaders Forum which had been active in Poland since 1992. Funding came from the John Ryder Memorial

Fund and the programme, titled 'The Social Impact of Industrial Change', began in late 1997. The initial step was to gather information on how communities deal with the transition from a state planned communist command economy to a market economy.

Poland at that time was going through a severe economic downturn and foreign investment was minimal. The old labour-intensive practices were being dropped and people had to rely on their own resources to find a job in the new environment. Admittedly the 'black economy' was strong but the country was not coping with the social and health problems that had received some attention under the previous regime. Communities became more aware of the need to work together to solve some of the more pressing problems.

Some administrative and local government reforms were taking place and impending EU accession was changing the socio-economic development of the country. Adaptation to the market economy and the privatisation process was slow. Non-government organisations (NGOs) were emerging but there was uncertainty on their role. The transport, financial and legal infrastructures were in need of considerable change.

Action research was the chosen methodology, the emphasis being on solution oriented studies which would allow continuing intervention and interaction with the communities selected. The Cross Sector Partnership Conference was the chosen forum for stimulating discussion and facilitating action groups to create viable projects to solve the problems in a transition economy.

Demonstration projects were used to create dialogue and to build informal power networks. Among these were a residential programme for young unemployed to arrange work placements, anti-corruption projects, various infrastructure projects, manager shadowing, SME (small and medium enterprises) development, etc. One longer-term method used was capacity building including training, study units and strategic planning.

Cross-sector dialogue has played a valuable role in disseminating good practice. Other key findings have been the need for a regional approach and the business benefits of partnership, helping business to be more active in the development of future projects. Models have been developed and adapted on a countrywide basis for business to become more involved in the community. Qualitative benefits have been the development of democracy, growth of the NGO sector, local economy diversification, broadening of horizons and active adoption of the 'can do' mentality.

The recent history of Poland provides an example of ideologically based socio-economic structures that were imposed during the 20th century and which went wrong. The project described here is an example of efforts to rebuild those structures through the new way of participation and consent and of leaders informing and facilitating rather than directing. This is the way that emerged as we looked for the lessons of the living systems approach in changing society in the modern world.

10 Conclusion

In this chapter we have attempted to demonstrate the limitations of seeing human society and institutions as mechanically constructed organisations and to reveal some of the insights that can be gained from seeing our society as living systems having many analogies with systems to be found in nature. What started as a way of seeing companies holistically continued as an inquiry into how the same approach helps us to understand many aspects of human society as it adapts to the ever-changing conditions of both the environment and the results of human progress, whether technological, social or moral.

This is a continuing journey that reveals fresh questions and presents us with further unknowns, and indeed unknowables, along the way. During the past few centuries humanity has used the

human intellect, unique among Earth's species, to unravel many of nature's mysteries and to use the knowledge so gained for our own purposes. But as we make our discoveries the extent of our remaining ignorance is increasingly revealed to us. As we use our reason it is essential to have the humility to recognise that it cannot provide all the answers and nor is it infallible; there are many lessons to be learnt from the ways of nature.

As we look for these lessons there is a major trap to be avoided. The laws of nature revealed by science (as opposed to the simple rules of complex systems thinking) are so called because they provide an apparently invariable determinant of how material matter will behave under any given set of circumstances. They are not laws of obligation for beings capable of self-determination. The social Darwinians of the late 19th and early 20th centuries fell into this trap as they gave support to social philosophies of those such as the Nazis in Germany. Social Darwinians of that period also failed to recognise the role of diversity and the slow pace of Darwinian evolution, and some saw it pointing towards a laissez-faire approach. The lesson that we see in nature is not to determine for ourselves that which nature would select as the fittest nor is it to abdicate our own capabilities of self-determination, but rather to learn from the manner in which nature selects and to facilitate the conditions in which selections can be made in that manner. This has led us to put emphasis on the importance of relationships, lateral diversity and vertical harmony.

It is also necessary to recognise all that is meant by the human factor to which we have referred. The term 'survival of the fittest', coined by Herbert Spencer and often used to characterise Darwin's theory of natural selection, implies a complete absence of justice which is an important need for human society. Just how it becomes a need can be debated but a need it is. Selection involves a balance between all the relevant factors. Unreasoning nature ensures that balance in its own 'unjust' way and the challenge for the reasoning human race is to find the same balance with justice. Our reason is also used in a way that impacts on our environment. The problem is that our intelligence can be used to precipitate death as well as to sustain life and the great challenge facing us is how to use our gifts to ensure a sustainable existence for humanity and for our environment. As Nelson Mandela pointed out, respect and relationships are the key.

The whole of human history has been characterised by diversity and changes in ways of thinking. The diversity is represented by the variety of cultures to be found around the globe and change can occur with time. The pioneering work of philosopher Ken Wilber and psychologists Clare Graves, Don Beck and Chris Cowan has developed models for thinking about the evolutionary spiral of human cultures and values over time [17, 18], while the work of Dr Surinder Kaur [19] has also shown how our rationality is conditioned by our culture. Our ways of thinking are ways of using our gifts of intellect to understand ourselves and our universe. East and West, science and theology, ancient philosophy and modern philosophy may all seek truth in their own way, without any one being necessarily right whilst the others are wrong. They all have their place in the evolving spiral of human culture as it responds to changing life conditions. Human rationality may be varied and changing but, although the behaviour of the physical world may be less deterministic than was once thought, there is a constancy in the ways of nature which we observe to gain a better understanding of the systems of our own human creation.

Acknowledgement

About 70 people have contributed over the years to the discussions on which this chapter has been based. In addition to the two authors, those who have been particularly involved in these discussions are Peter Challen, Michael Collins, Pauline Graham, John Metcalf and Geoffrey Morris.

The ideas contained in this chapter are some which were presented or emerged during the discussions in which no attempt was made to reach a consensus or unanimous view on them.

References

[1] Cohen, J. & Stewart, I., *The Collapse of Chaos – Discovering Simplicity in a Complex World*, Penguin: London, 1994.
[2] Urry, J., *Global Complexity*, Polity Press with Blackwell Publishing: Oxford, 2003.
[3] Allen, P.M. & McGlade, J.M., Evolutionary drives: the effect of microscopic diversity, error making and noise. *Foundation of Physics*, **17**, pp. 723–728, 1987.
[4] Allen, P.M., Knowledge, ignorance and learning. *Emergence*, **2**, pp. 78–103, 2000.
[5] Porter, M., Clusters and the new economics of competition. *Harvard Business Review*, November–December 1998.
[6] Ruse, M., Sociobiology (Chapter 11). *Can a Darwinian be a Christian?*, Cambridge University Press: Cambridge, pp. 186–204, 2001.
[7] Madron, R. & Jopling, J., *Gaian Democracies*, Green Books: Totnes, UK, 2003.
[8] Challen, P. & Shakespeare, R., *Seven Steps to Justice*, New European Publications: London, 2002.
[9] Nixon, B., *Global Forces*, Management Books: Chalford, UK, 2000.
[10] Mandela, N., Independent lecture, Dublin, 12 April 2000.
[11] Bunzl, J.M., *The Simultaneous Policy – An Insider's Guide to Saving Humanity and the Planet*, New European Publications: London, 2001.
[12] Charter 99, co-ordinated by the One World Trust, http://www.oneworldtrust.org
[13] Kauffman, S., *At Home in the Universe*, Penguin: London,1995.
[14] Maturana, H.R. & Varela, F.J., *The Tree of Knowledge – The Biological Roots of Human Understanding*, Shambhala: Boston, USA, 1998.
[15] Gilchrist, A., The well-connected community: networking to the edge of chaos. *Community Development Journal*, **35,** p. 264, 2000.
[16] Gibb, J., *Deep Simplicity – Chaos, Complexity and the Emergence of Life*, Penguin/ Allen Lane: London, 2004.
[17] Wilber, K., *A Theory of Everything – An Integral Vision for Business, Politics, Science, and Spirituality*, Gateway (Gill & Macmillan): Dublin, 2001.
[18] Beck, D. & Cowan, C., *Spiral Dynamics – Mastering Values, Leadership and Change*, Blackwell: Oxford, 1996.
[19] Kaur, S., Culturally Bounded Rationality, PhD Thesis, Henley Management College, 1996.

Chapter 12

Searching for improvement

M.A. Atherton[1] & R.A. Bates[2]
[1]*School of Engineering and Design, Brunel University, UK.*
[2]*Department of Statistics, London School of Economics, UK.*

Abstract

Engineering design can be thought of as a search for the best solutions to engineering problems. To perform an effective search, one must distinguish between competing designs and establish a measure of design quality, or *fitness*. To compare different designs, their features must be adequately described in a well-defined framework, which can mean separating the creative and analytical parts of the design process. By this we mean that a distinction is drawn between identifying novel design concepts, or architectures, and the process of detailing or refining existing design architecture. In the case of a given design architecture, one can consider the set of all possible designs that could be created by varying its features. If it were possible to measure the fitness of all designs in this set, then one could identify a *fitness landscape* and search for the best possible solution for this design architecture. In this chapter, the significance of the interactions between design features in defining the metaphorical fitness landscape is described. This highlights that the efficiency of a search algorithm is inextricably linked to the problem structure (and hence the landscape). Two approaches, namely, *genetic algorithms* and *robust engineering design* are considered in some detail with reference to a case study on improving the design of cardiovascular stents.

1 Introduction

1.1 Search domains

The term *blue print* continues to be used, figuratively at least, long after the original device ceased to be widely used in engineering design. A blue print represented an expectation that the designer's intent would be faithfully reproduced in the finished artefact. It was not necessarily a plan of how to make the object but might indicate why any modifications to the original design had been made. Invariably these revisions of the blue print would be based on actual performance of the object and thus improved designs were often the result of trial-and-error. That is to say, the design process was heuristic.

'Blue prints' for every living thing on earth, uniquely encoded in the form of deoxyribonucleic acid (DNA), represent not only component parts but also their interrelationships within the total organism. Under changing circumstance, genetic code is said to adapt in order to survive over successive generations. This idea has been employed in engineering design process, notwithstanding the great differences in the respective timescales, economies and technologies between nature and engineering.

The description of an engineering system, embodied by its design, can be made on two levels: the basic operating principle of the system, i.e. its specific *technology* and then, within that technology, particular configurations of design features. Decisions made in the design process at both these levels determine achievement, or otherwise, of successful function (the solution) for a given application (the problem). For example, an innovative concept for a mechanism will not be successful if inappropriate materials and geometry are chosen, and conversely a rather crude mechanism can perform successfully if the detail is right. Parallels can be drawn with nature such as in the design of an eye. At the technology level, design could relate to whether the eye is of a refractive (e.g. human) or reflective (e.g. lobster) configuration. At the feature level, design could relate to the values of lens dimensions and pupil shape that are assigned when the operating principle is refractive. However, there are lower limits on the dimensions of a retinal eye, as at very small scales it cannot function, i.e. there are parametric constraints. This example again highlights that engineering design operates in two broad domains. Designs can be categorized in terms of the technology used and then, within each technology, competing designs can be thought of as a collection of features that are defined by *design parameters*, also called *design factors*.

In engineering at least, the process of designing a solution that utilises new technology is very different to that for deciding design parameter values. The former is usually addressed as a problem of creativity and the latter can be formulated as a mathematical search problem. Figure 1 illustrates this distinction between design in the technology domain and design in the feature domain.

In other words, the description of an artefact and the process by which its description is arrived at are inextricably linked. This means that an integrated engineering design process must concurrently create an operating principle and also identify workable design factor values, yet it must employ different methods in the two domains.

Systematic mathematical search tools are practically limited to design in the feature domain due to the difficulty of expressing creativity in a symbolic language. However, there are systematic methods for proposing viable technologies that use knowledge of successful design solutions (e.g. patents) to focus the search on a small number of operating principles for evaluation [1]. Essentially, these methods still rely on the creative ability of the designer to make a successful interpretation in the context of the problem. In this chapter we shall operate in the feature domain and consider it as a mathematical quest for improvement.

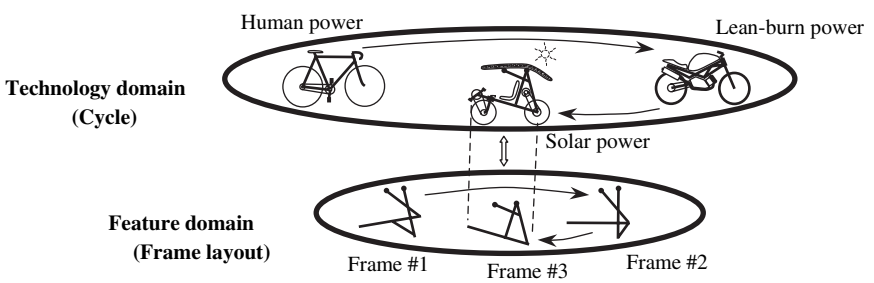


Figure 1: Abstract illustration of search within both technology and feature domains.

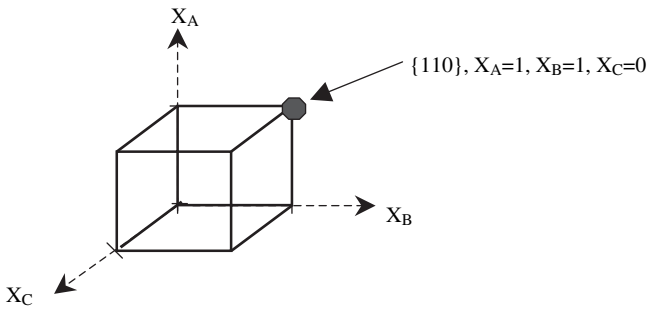


Figure 2: Design space for three design factors with binary levels.

Mathematical search is viewed to take place within a *design space* defined by the number of design factors and the set of possible values for them, rather than all possible solutions. This defines the dimension and limits of the design space. For example, a system described by three design factors X_A, X_B and X_C, each having two possible values (e.g. 0 and 1), could be represented as a point in a 3D design space, shown on a unit cube in Fig. 2.

An *efficient* search will rapidly converge on improved solutions regardless of the starting point. An *effective* search will yield significant improvements over existing designs. This implies that a good solution can be found without testing all possible solutions to the problem.

The above example describes a problem with *discrete* design factors. Each factor can take one of two possible values, 0 or 1 and consequently there are $2^3 = 8$ possible solutions. In practice, design factors can be either discrete or continuous (e.g. take any value between 0 and 1). In the latter case, the set of possible solutions ceases to be finite. The distinction between discrete and continuous factors may not appear to be important, but in fact it can have a significant effect on how the design space is searched. This will be discussed in more detail later on in Sections 2 and 3. Design factors such as the length or weight of a component may be continuous, whereas the choice of material for the component may be a discrete factor. However, even in this simple case, the factors can be difficult to classify. The component may only be available in a fixed number of lengths and weights and, conversely, if the material is defined by a factor such as Young's modulus, it may be possible to consider the material specification as a continuous factor. In all cases it is necessary to define the factors to accurately represent the problem to be solved. It may be the case that it is possible to manufacture customized components, allowing factors to be expressed as continuous, but this is an additional cost over and above the use of standard sizes. This needs to be taken into account when searching the design space for acceptable solutions.

1.2 Why use mathematical models?

Major issues in the design of any engineering system include cost, quality, reliability and demand. In this context, system optimisation can mean many things: minimising cost, improving reliability and so on. Each of these *objectives* will almost certainly be in conflict, and seeking design improvements that simultaneously satisfy them can therefore be a complicated process. Figure 3 outlines an example design scenario.

Systems that exhibit complex behaviour are often expensive to test during product development. This can immediately rule out a trial-and-error approach. An alternative strategy is to develop mathematical models of the system in order to gain understanding of the relationship between

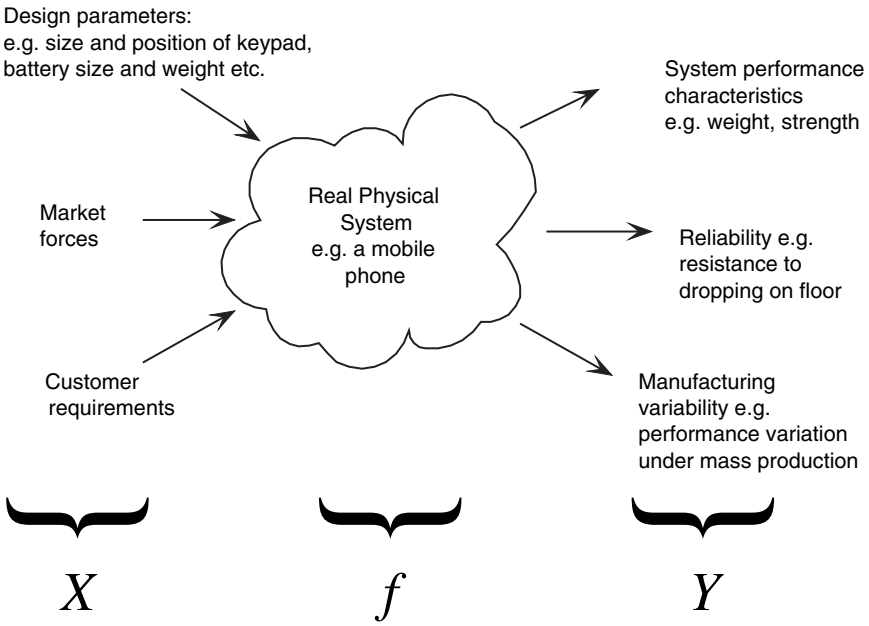


Figure 3: An example design scenario.

the design and its performance. Mathematically, the system is represented by an equation of the form:

$$Y = f(X). \tag{1}$$

The challenge here is to find f, in other words to find out how the inputs to the system (represented by X) affect the outputs of the system (represented by Y), as indicated in Fig. 3. There are two basic approaches to characterising f: *physical modelling* and *empirical modelling*. The latter approach is known as a *black box* method as the details of the system are treated as entirely unknown. A third way of characterising f, based on elements of both physical and empirical modelling, is known as *grey box* modelling.

1.2.1 Physical modelling

Complex engineering systems are generally designed from the principles of physics. This leads to mathematical models, often using differential equations, representing the system. An example of this is the analogue electronic circuit design where characteristic equations exist for each component of the circuit and these are combined to form banks of differential equations that need to be solved to deduce the *ideal* behaviour of the system. The word ideal is important here as these models are only approximations to the real system, and need refinement if they are to reflect non-ideal behaviour such as manufacturing variation and losses due to electrical resistance, friction or other, possibly unforeseen effects. These physical models are very important as the mathematical theory behind them forms the basis of computer-aided engineering software such as finite element analysis, computational fluid dynamics, and electronic circuit analysis. Software that incorporates this type of analysis is referred to as a simulator, as it can be used to define a physical system and simulate its behaviour on a computer.

1.2.2 Empirical modelling

Physical systems can be tested to gather information about their behaviour. The field of experimental design is concerned with the design of such tests in order to maximise the information gained while minimising the size of the test. The test involves observing the system at a carefully chosen set of design points, each point representing a particular set of design factor values. Referring back to Fig. 3, this means observing response features Y, made on the system f while varying the factors X. Once an experiment has been conducted, mathematical models are sought that fit the data gathered. These models are approximations of f, sometimes written as $\hat{f}$, and can be used both to estimate the relationships between factors and responses and to predict the response at untried factor values.

1.3 Building mathematical models

In general, whilst they both employ mathematics, physical and empirical modelling strategies have historically been separate but are becoming more closely linked via grey box modelling. For example, a computer model can be used to provide structural information on systems as a starting point for experimentation [2], and estimates of variation in design factors and responses can be used to make physical models more accurate.

In some cases, mathematical models of systems can themselves be complex, and systems are often modelled with powerful computer simulators as described. Complex computer simulations of systems can however be very costly in terms of computation time and in this case black-box modelling can be used to construct simpler empirical models that are faster to evaluate, this is known as the field of computer experiments. Such models are referred to as emulators, meta-models, low-fidelity models or surrogates and their main characteristic is that they trade off accuracy for speed (Fig. 4).

There is a close relationship between modelling and optimisation. The availability of mathematical models of complex systems opens the possibility of fully exploring the design space of all feasible combinations of factors to determine the best design, but even for small problems the dimension of the search space can be high and optimisation can still be difficult.

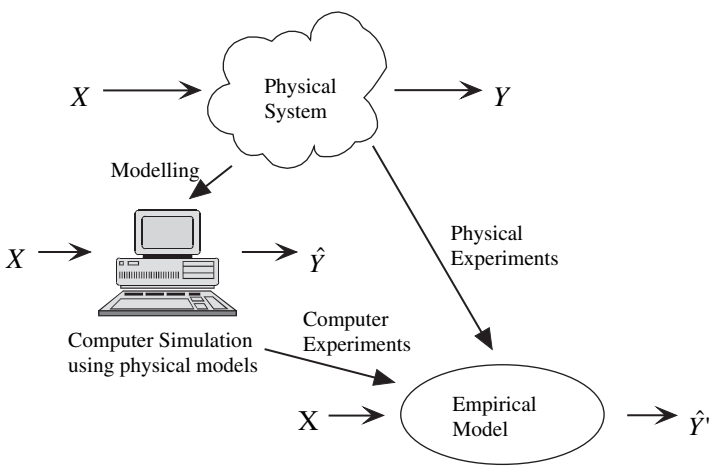


Figure 4: System modelling summary.

When considering the optimisation of a system or process there are several key decisions to be made about the nature of the search that will ultimately determine the level of success achievable.

1. *Parameterisation*: The system must be described in terms of design factors, X (the parameters), so that mathematical methods may be used for experimental design, modelling (both physical and empirical), and optimisation. The nature of each factor needs to be defined: whether they are discrete or continuous variables, the operating range, etc. This is perhaps the most important aspect of the formulation of a search and optimisation strategy as it determines the set of all possible design solutions.
2. *Experimental design*: Any empirical model of the system needs information on how the system responds to changes in design factor values. In neural network terminology this is referred to as the training set. There may be additional constraints on any experiments to be conducted on the system such as non-regular or non-feasible parts of the design space (constraints on certain combinations of factor settings).
3. *Modelling strategy*: Any approximate model of the system needs to be accurate.
4. *Objectives (or fitness functions):* The objective, or combination of objectives, sometimes referred to as the fitness function, is a statement of the goal of the optimisation process. A typical example would be to maximise strength whilst minimising weight. Even in this relatively simple case one can see that there is a trade-off to be made.
5. *Numerical optimisation*: Optimisation of the system (or a model of the system) can be either global or local in nature. Local methods seek to improve on previous solutions by changing factor values gradually, while global methods explore the design space more fully by making large changes to factor values. The two strategies can be combined by, for example, performing several competing local searches each at different starting points. Many optimisation algorithms exist, and choosing the right one for a given problem requires knowledge of the complexity of the problem. For example, are the functions to be optimised linear or non-linear? Similar consideration must be given to any constraints on inputs and outputs of the system, which will also have a functional form.

All the above decisions on how to conduct the search combine to determine the set of possible solutions that will be found. In fact it is the objectives that drive the optimisation process and determine which are the most suitable methods to use. In the simple example of the strength/weight trade-off, it may be desirable to explore sets of possible design solutions that place different emphasis on the two objectives so that a light and weak solution is compared with a heavy and strong one. If this is the case then it is preferable to have models of the system that can be adjusted quickly and efficiently so that the solution space can be explored effectively.

Alternatively, it may be the case that the overall objective is well known and a direct search of the system is appropriate. In this case modelling may be unnecessary, particularly if evaluations of the system are inexpensive.

1.4 Design robustness and variability

An important part of the quality of any design is the ability to cope with unwanted variation or noise. This may be in the form of variation in factor values, variation in manufacturing conditions or variation in the use environment. In the context of design improvement, robustness means the ability of a design to maintain performance in the face of such noise. In order to understand how noise affects a particular design, one needs to first characterise the noise and then see how the design behaves when subjected to it. Unfortunately this can significantly increase the burden of testing, as design factors need to be varied on both a macro-scale, to search for an improved

design, and a micro-scale, to identify how small changes in factor values affect performance. More detailed discussion of noise and robustness in design can be found in [3].

In terms of mathematical search, the design improvement problem changes from a deterministic problem, where exact factor values yield exact responses, to a more probabilistic formulation, where factor values are defined by statistical distributions and propagated through the system. Single response values then become response distributions that need to be optimised. Such a response distribution is most simply characterised by taking its mean and variance. Where previously the design improvement goal would be to maximise the response, now the goal might be, for example, maximising the mean of the response and minimising the variance of the response, this particular problem formulation is referred to as the *dual response* method [4].

The scenario just described, to minimise the variation in performance for a given level of input noise, is known as a *parameter design* problem. If one could imagine having control over the amount of noise the system is subject to, for example, by specifying more accurate (and more costly) components, an alternative problem formulation could be to find a design that meets given targets of response variation for minimum cost. This is known as a *tolerance design* problem. Modern search strategies recognise that parameter design and tolerance design are linked and that good design solutions can only be reached by considering them simultaneously.

2 Fitness landscapes and interactions

2.1 Feature domains and design performance

The previous section introduced the concept of the feature domain in order to characterise competing engineering design solutions. A design is decomposed first into features and then into design factors that define the design space. Each design can then be thought of as a point in the design space, which represents the full set of possible design solutions associated with the specified engineering problem. In order to compare different designs, specific performance characteristics, or *responses*, must be defined such as weight, strength, power output and so on. These responses, taken together, describe the overall fitness of the design for its intended purpose. If one could imagine knowing the full set of performance characteristics for every design in the design space then this would define the *performance space*, representing the same set of design solutions from the perspective of design performance, rather than design factor values. Figure 5 describes this for the simple case where there are two design factors, $\{x_A, x_B\}$, and two responses, $\{y_A, y_B\}$.

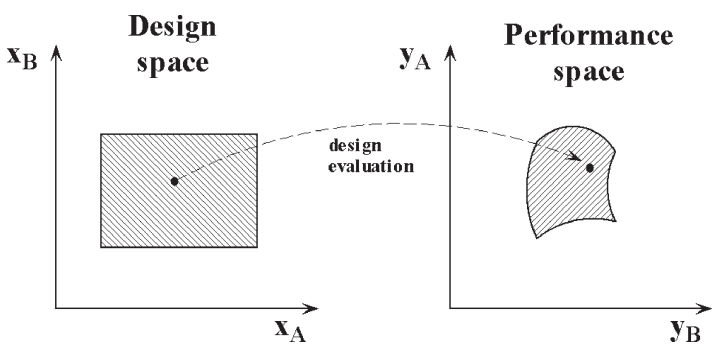


Figure 5: Design and performance spaces.

Of course it is easy to visualise such a case when there are only two design factors and two responses. However the concepts described are still valid for more complicated examples and can serve to describe techniques such as parameter design and tolerance design, mentioned in the previous section, as well as other design methods such as design centring and yield optimisation that we will not mention further here.

2.2 Fitness for multiple purposes

Beyond three dimensions, the limitations of visualisation mean that diagrams such as that presented in Fig. 2 can only provide a partial glimpse of *n*-dimensional design space. Representing performance requires an additional dimension. Therefore, in practice, attempts to plot design space search are abandoned for abstract mental images and instead simplified visualisations are used to plot performance against a subset of one (or two) design factor(s) or perhaps another performance objective. Indeed most engineering problems have more than one response to satisfy, i.e. they are multiple objective problems.

Satisfying multiple objectives is a challenge faced by organisms too, as we shall see. Invariably, due to interdependencies multiple objectives conflict with each other to the extent that as we increase satisfaction of one objective this typically results in decreasing satisfaction with the other objectives (Fig. 6).

Therefore compromise is usually inevitable for design confined to the parameter domain. Ignoring an improved concept design as a means of settling the conflict, trade-off is thus inevitable between multiple objectives and in parameter design two approaches are generally employed:

1. Selecting one of the main objectives and incorporating the others as *constraints* [5].
2. Employing a general 'portmanteau' unifying objective or *utility function* [6].

The utility function approach is often preferred for engineering robustness as it enables sensitivity analysis whilst in some cases with the former approach there is no feasible region of design space remaining after constraints are applied. The *desirability function* [7] is one such utility function. It transforms or maps each response into a desirability variable and then combines them geometrically into an overall desirability, *D*, which is effectively a continuous function of the responses. Thus a multivariate problem is expressed as a univariate one.

In biology, the overall performance of an organism is expressed as its *fitness*, in terms of its ability to survive and reproduce [8]. Fitness can be viewed as a utility function measuring survivability or level of adaptation. This level of adaptation can be likened to the elevation of a

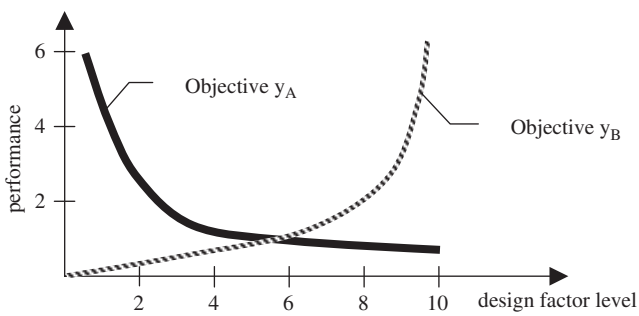


Figure 6: Two performance objectives plotted against a design factor.

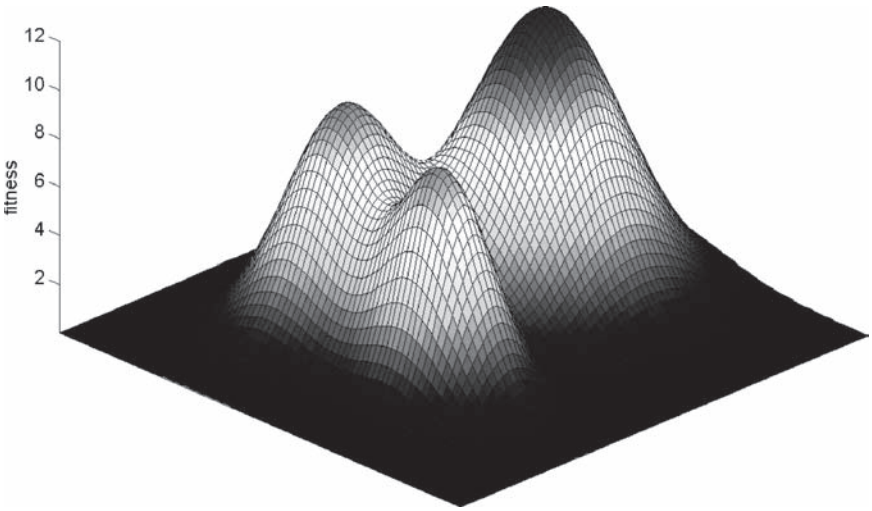


Figure 7: Theoretical fitness landscape.

landscape (Fig. 7) in which the peaks are populated by the higher living organisms. Here design factors and responses are combined in a single plot to indicate how adjusting the value of one factor can change the fitness, which is composed of response values.

This fitness landscape is a mathematical concept, not a literal terrain, but this vivid metaphor can be usefully manipulated. Imagine that the landscape is elastic and at the location for any particular organism the terrain deforms when the living conditions or the fitness of a contingent organism such as a predator, prey or parasite changes. Thus fitness has been termed a 'red queen effect' [9], described as a never-ending race merely to sustain fitness level amidst co-adapting competition. This notion is seen to apply to economic systems. Taking this further we envisage that due to advances of competition the desirability (e.g. its utility function) of a product can diminish whilst its performance remains unchanged.

Quality in human technology has an aspect roughly analogous to biological fitness [10] and stress has been laid on quality loss functions [11] as a powerful measure of utility in engineering problems. The general idea being that the ideal target product performance is one that incurs zero loss to society in terms of the cost of, for example, environmental damage, maintenance, injury, inconvenience or some other expense not directly related to the intended function of the product. We now consider how the co-adaptation analogy might combine with the quality loss function in dealing with the multiple objective optimisation of diesel engines.

The primary intent of a diesel-cycle internal combustion engine is to produce useful tractive power. On each cycle of the engine most of the fuel is completely burnt and produces useful energy. The remainder of the fuel is not completely burnt and therefore pollutants, such as particulate (smoke) and unburnt hydrocarbon (HC) emissions, are present in the exhaust gases. In both cases the quality loss function associated with these pollutants is 'smaller-the-better', as shown in Fig. 8. Loss is assumed to be a quadratic function of each pollutant, such that $L = ky^2$, where y is, say, the mean output of a pollutant, k is a coefficient that specifies the quadratic curve and L is measured in monetary units (e.g. British pounds (£)).

Engines from two rival manufacturers, A and B, are depicted above in relation to each other for the two pollutants (i.e. two objectives): smoke (S) and hydrocarbons (HC). The performances,

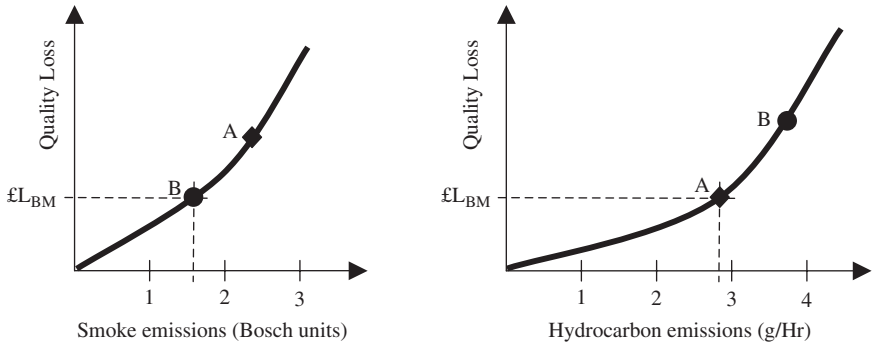


Figure 8: Quality loss functions of HC and smoke for product A and B showing a benchmark loss value, $£L_{BM}$.

y_S and y_{HC}, of each engine can be measured fairly straightforwardly. However, we cannot precisely define the quality loss functions for S and HC because the actual loss incurred for a given emission of pollutant, in terms of damage to health and property, reduced fuel economy and so on, is incalculable.

$$L_{HC} = k_{HC} y_{HC}^2, \tag{2}$$

$$L_S = k_S y_S^2. \tag{3}$$

But as competition contributes to the notion of a fitness landscape for organisms, so can competition help to define the quality loss functions in this case, as follows. Using the best performance (*benchmark*, 'BM') for each pollutant, an arbitrary loss value, say $£L_{BM}$, can be assigned to both and thus k_{HC} and k_S, the coefficients of the two quadratic functions, determined from rearranging eqns (2) and (3) respectively. Both pollutants are correlated as they are products of not-completely burnt fuel, this enables a portmanteau objective to be calculated for each engine performance, such as the overall 'fitness', total loss, $L = L_{HC} + L_S$. This loss changes if a new benchmark performance is reached or if a difference in the cost weighting between S and HC emerges. There is a trade-off relationship between the two pollutant emissions and determining quality loss functions by virtue of competitive benchmarking penalises products that stand still [12]. Thus the 'loss landscape' for each pollutant behaves as if it were elastic, changing according to competitive forces.

According to Goodwin [13] more sophisticated descriptions of landscapes tend to move away from the use of such non-generic fitness functions and towards language such as attractors and trajectories, attempting a unification of biology, mathematics and physics through the study of complexity.

2.3 Multi-criteria decision making

Instead of combining individual objectives into a single fitness function, an alternative approach is to keep each performance measure separate. This leads to the idea of a performance space, described in Section 2.1, where it is then possible to show sets of competing design solutions. This is useful in trade-off situations where one objective is in conflict with another. In order to rank competing solutions, the idea of *Pareto optimality* [14] can be used, which involves the

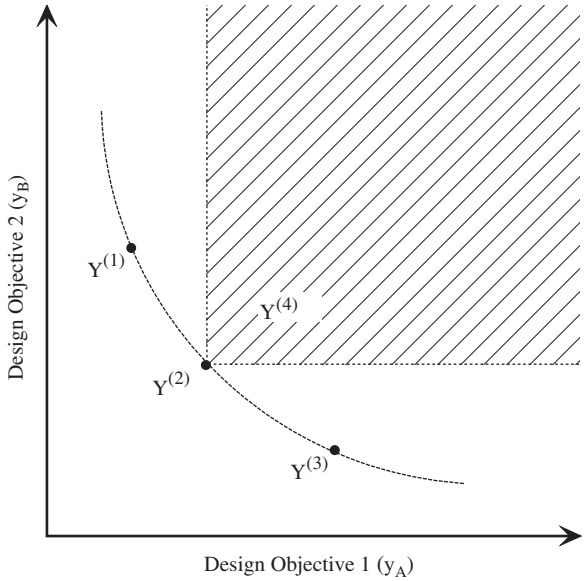


Figure 9: Pareto boundary.

concept of dominance, where one solution is said to dominate another if one or more of its objective function values are better, and none are worse. This is graphically described in Fig. 9 which shows four design solutions, $Y^{(1)}, \ldots, Y^{(4)}$, for a problem where the aim is to minimise two design objectives, y_A and y_B, and Y represents some combination (or function) of the two objectives, $Y = f(y_A, y_B)$.

The shaded area in Fig. 9 represents the region dominated by solution $Y^{(2)}$, therefore solution $Y^{(4)}$ is said to be dominated by $Y^{(2)}$ as its values for both design objectives are worse. The dotted curve shows an estimate of the Pareto boundary or *Pareto front*, which represents the set of all non-dominated solutions. In this simple example, the Pareto front is assumed to be convex, but this may not necessarily be the case. Armed with such information, a designer would be in a good position to decide how to trade off one objective against another in the search for the design factor values which represent the best design solution.

2.4 Coupling and search

The key to understanding the scope of natural selection theory depends on understanding the structure of the fitness landscape explored by an adapting population. For example, whether it is smooth and single-peaked, rugged and multi-peaked, or just completely random. One must also consider the mechanism by which the population adapts. The fitness landscape of Fig. 7 is composed of two design factors and a fitness function, all of which can vary their values on a continuous scale. If a population is described in binary terms, such as a genetic encoding, then the design space becomes discrete and the relationship between one design and its nearest neighbour in design space is not well defined. One could say that the geometry of the search space has been weakened or even destroyed and therefore search strategies need to cope with this.

Genotype spaces are vast. Consider organisms with N different genes each of which has two versions, or alleles, 1 and 0. For a haploid population, such as that of *E. coli*, there are apparently

3000 genes [9]; therefore genotype space is 2^{3000} or 10^{900}. For a diploid population such as in plants that may have 20,000 genes then genotype space is $2^{2000} \times 2^{2000} = 10^{12000}$. Therefore let us consider walks across simpler fitness landscapes.

A genotype with three genes, N, each having two alleles, A, has $A^N = 8$ possible genotypes – {000}, {001}, {010}, …, {111}. Each genotype is 'next to' those that differ by changing any one of the three genes to the alternative allele value. Figure 10 shows fitness values arbitrarily bestowed on each genotype. The arrows point uphill to fitter neighbours.

An adaptive walk on this random fitness landscape only moves to a fitter variant from amongst the three immediate neighbours. In some cases these walks end at local peaks, e.g. {101} and {110}, as shown in Fig. 11.

In this simple example there is one global best peak and two local peaks but in a large genotype space the number of local peaks on a random landscape is $2^N/(N+1)$. Hence for $N = 100$ there are more than 10^{28} local peaks. Thus adaptive search on random landscapes is difficult because finding the global peak by uphill search becomes almost impossible. Searching the entire design space could feasibly exceed even the most generous estimates of the age of the universe unless more intelligent methods exist. Figure 10 also highlights the lack of geometry of problems posed in this way as the position of each genotype is plotted arbitrarily (in this case to echo the shape of the fitness landscape in Fig. 7).

From any initial arbitrary point on a landscape, adaptive walks reach local peaks in a number of steps. The expected length of such walks to local peaks are very short (ln$N$) as any initial point is very close to one of the local peaks, which trap the adapting population and prevent further

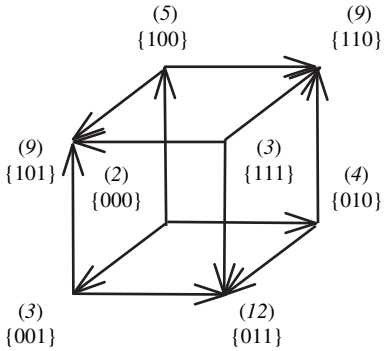


Figure 10: Genotype space (showing fitness values of each genotype).

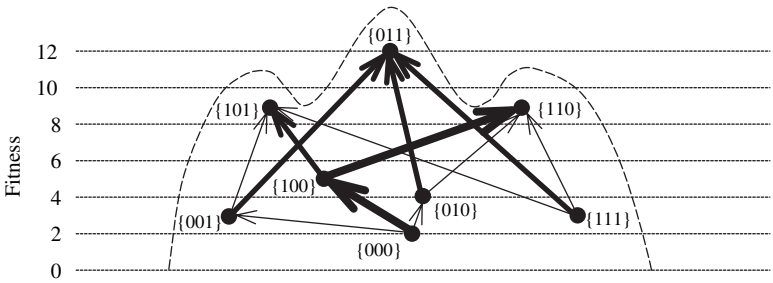


Figure 11: Fitness landscape showing walks of genotype space.

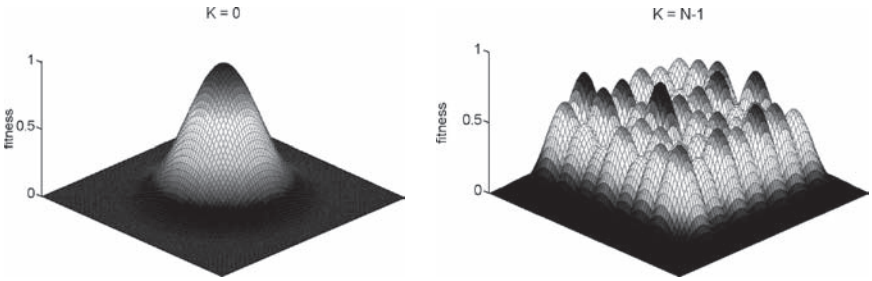


Figure 12: Effect of epistasis on the ruggedness of a fitness landscape.

search for distant higher peaks. Moreover, the higher the fitness, the more difficult it is to find improvement, as each step upward requires twice as many options being searched. However, real landscapes are not random, they are correlated, i.e. nearby points tend to have similar heights.

Gene epistasis or epistatic coupling is where the contribution of one allele of one gene to overall fitness of an organism depends in complex ways on the allele of other genes. Thus a network of epistatic interactions might exist. The *NK* model [9] captures such networks, where K reflects the degree to which nodes on the landscape interact. $K = 0$ represents total independence between nodes. $K = N - 1$ represents the highest possible value of K where all nodes interact with each other. In a more general sense, when $0 \leq K \leq N - 1$ then the K genes assigned to interact with each gene are chosen at random. In effect K alters the ruggedness of the landscape. When $K = 0$ we have a single smooth-sided peak and as K is increased – genes are more interconnected – more conflicting constraints exist and so the landscape becomes more rugged with more local peaks (Fig. 12).

Many rugged peaks occur because the best states of the shared epistatic inputs for one gene will be different than for its partner and thus in conflict – there is no way to satisfy both as much as if there was no cross-coupling between their epistatic inputs. In other words, as K increases there are so many constraints in conflict that there are a large number of compromises rather than a single best solution. As landscapes become more rugged, adaptation finds it more difficult to make the crossing. K is like increasing the compression of a compressed computer program. With $K = 0$ changing any gene can only change the genotype fitness by at most $1/N$. Therefore the side of the peak is smooth and from any random starting point the number of directions uphill reduces by only one with each step. This dwindling of options is in sharp contrast to random landscapes where the number of uphill options reduces by half at each step. Gradualism works only on such a smooth single-peaked landscape. Thus as K increases the number of peaks increases, ruggedness increases, peak heights drop and locality of search increases. More interestingly, at moderate degrees of ruggedness, the highest peaks can be selected from the greatest number of critical positions, i.e. high peaks have the largest surrounding slopes [9].

3 Some methods for design improvement

Here we describe and compare in some detail two methods for searching the design space for improved designs. The first method, robust engineering design, is built on the traditional field of design of experiments and has both a classical and a more modern approach. The second method defines the search problem in more biological terms and uses genetic algorithms to search for improvement.

3.1 Robust engineering design

Robust engineering design (RED) seeks to make engineering products robust to variation in both manufacture and use. A key aspect of RED is to understand the significance of each design factor on system performance through a highly structured search (Fig. 12). Exposing the design to representative noise conditions and subsequently observing its behaviour are fundamental to the method. The design space can be searched directly, using physical prototypes or indirectly using a representative model such as a simulation model. Some parallels can be drawn with the search in *genetic algorithms* (GA) (see Section 3.2) but in general, for RED, very careful selection and arrangement of design factor values is required.

Figure 13 shows three main stages in the RED methodology: experimentation, analysis and design improvement, or optimisation. Experimentation involves choosing the type and size of an *experimental design plan* that will be used to evaluate different designs. Depending on the type of experiment chosen, the analysis stage interprets the results and provides information on the relationship between the design factors and the responses. This information is carried forward to the optimisation stage, where improved designs are sought. Choosing and executing an experiment appears, on the face of it, to be the first step in applying RED methods. However, this can only be done once the method of design improvement has been decided. The first step is in fact to determine the design objectives. This will define how each design solution is judged and will point to the type of analysis method and therefore the type of experiment required.

There is no single method for performing RED; rather, there are many different methods that can be used in the three stages described. One important distinction between different methods is whether a model is built to describe the relationship between factors and responses as part of

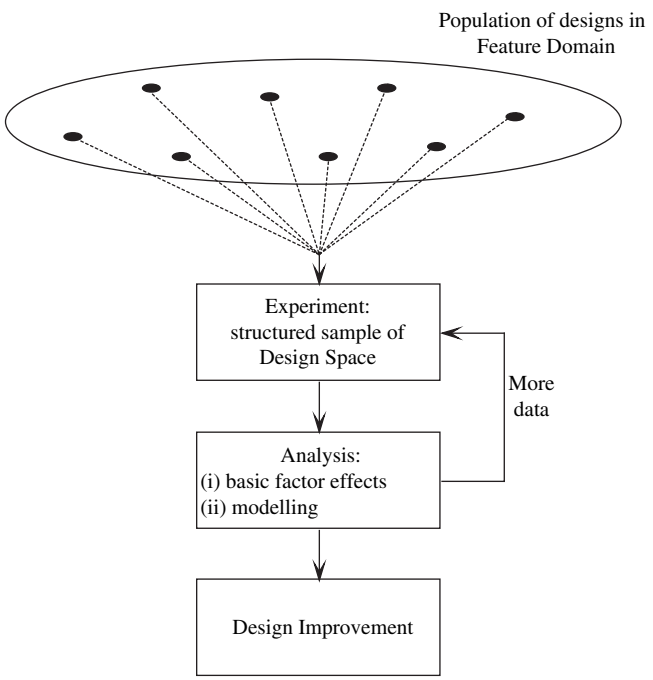


Figure 13: General RED procedure.

the analysis stage (see Fig. 4, Section 1.3). This is sometimes referred to as *model-based RED*. Another important distinction for experimentation is whether, for model-based RED, the structure of the analysis model needs to be specified beforehand (model-based experimental design), or whether the data collected is used to determine the model structure (model-free experimental design).

3.1.1 'Classic' RED

In relation to eqn (1), $y = f(x)$, the 'classic' approach to RED assumes that a simple *additive* relationship exists between the design factors (x) and some transformation, η, of the response (y). That is, *classic RED* is 'model first' in that usually a first- or second-order polynomial is budgeted to search the design space. An additive relationship is represented in eqn (4) for three design factors:

$$\eta = g_1(x_A) + g_2(x_B) + g_3(x_C). \tag{4}$$

It is important to note here that, in general, additivity with respect to η does not imply additivity with respect to y [15]. 'Additivity' is so central to classic RED, that the avoidance of interactions or cross terms (e.g. $x_A x_C$) between the chosen design factors is a dominant issue because they can render the assumed model unreliable.

3.1.1.1 Orthogonal arrays An *orthogonal array* (OA) is a predetermined matrix commonly used for coding the design factor levels to be used in a set of classic RED experiments (Table 1). It is the experiment plan.

Each column of an OA represents the values a particular design factor will take. The allocation of levels in each column is balanced with the other columns such that between any two columns each factor level is paired an equal number of times with the levels of the other columns and vice versa. The effect of this *orthogonality* is to search design space efficiently and also enable the average value of η for each design factor level to be compared. Data is collected for each experiment under discrete conditions of noise. Figure 14 illustrates the nature of the search and also highlights how each design factor is tested evenly against changes in the levels of other design factors.

From Table 1 it can be seen that, in terms of η, the average effect of design factor A at level 0 is calculated, according to the first column, as the mean η of the first two experiments (eqn (5)):

$$\overline{\eta}_{A0} = \tfrac{1}{2}(\eta_\alpha + \eta_\beta), \tag{5}$$

and so on for all factor levels yielding six *mean design factor effects*, which are all the permutations of the two-value combination means from η_α, η_β, η_γ, η_δ (illustrated in Fig. 15). For comparison

Table 1: Simple OA (L_4).

	Design factor A	Design factor B	Design factor C	High noise data	Low noise data	Data transformation (unspecified)
Experiment α	0	0	0	$y_{11}\ y_{12}\ y_{13}$	$y_{14}\ y_{15}\ y_{16}$	η_α
Experiment β	0	1	1	$y_{21}\ y_{22}\ y_{23}$	$y_{24}\ y_{25}\ y_{26}$	η_β
Experiment γ	1	0	1	$y_{31}\ y_{32}\ y_{33}$	$y_{34}\ y_{35}\ y_{36}$	η_γ
Experiment δ	1	1	0	$y_{41}\ y_{42}\ y_{43}$	$y_{44}\ y_{45}\ y_{46}$	η_δ

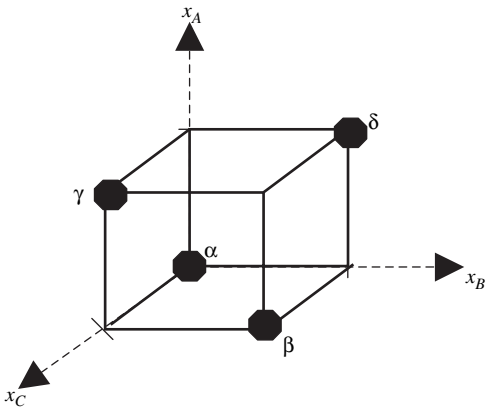


Figure 14: Balanced search of 3D-design space by OA.

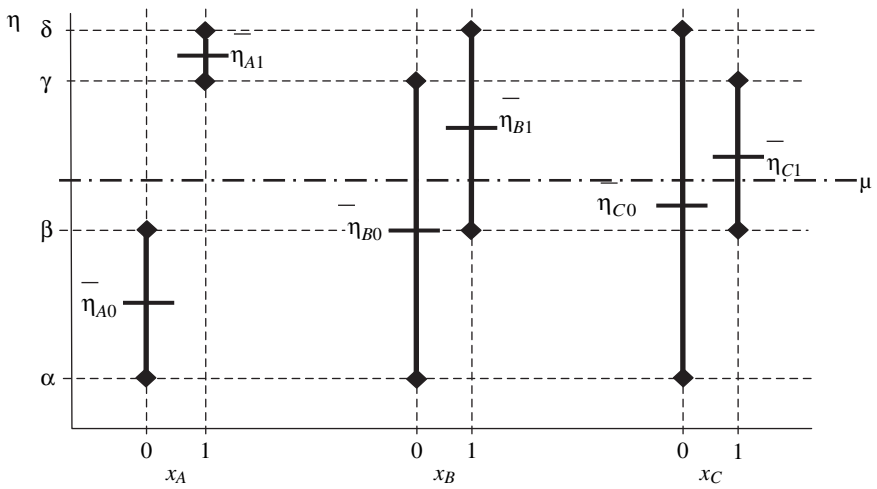


Figure 15: Mean design factor effects.

with Fig. 5, experiment δ could be described as $A1\ B1\ C0$ or design $\{110\}$; therefore, η_δ could be expressed as η_{A1B1C0} or more simply η_{110}.

Consider predicting the value of η_{jkl} for an untried configuration x_{Aj}, x_{Bk} and x_{Cl}, where j, k and l signify the levels of each design factor. From eqn (4) each of the three terms, e.g. $g_1(x_A)$, can be viewed as a contribution to η, and from Fig. 15 this can be developed into eqn (6):

$$\eta_{jkl} = \left(\overline{\eta}_{Aj} - \mu + \frac{\mu}{3}\right) + \left(\overline{\eta}_{BK} - \mu + \frac{\mu}{3}\right) + \left(\overline{\eta}_{Kl} - \mu + \frac{\mu}{3}\right) = \overline{\eta}_{Aj} + \overline{\eta}_{Bk} + \overline{\eta}_{Cl} - 2\mu. \quad (6)$$

This is of more direct use with the OA for prediction than the more familiar general form of eqn (4).

There is an underlying assumption inherent to the OA/additive prediction model combination expressed in the above equations, i.e. the effects associated with all of the OA columns account

for the system performance within acceptable confidence limits. In other words, any significant design factors or interactions not handled by the columns that vary will corrupt the predictive power of the classic RED method. Therefore design factors are carefully selected, grouped and allocated to an OA in accordance with the additive model being used. In effect, these design factor assignments centre on the issue of interactions.

3.1.1.2 Interactions Dealing with interactions in classic RED has two schools of thought. One school [11] advocates saturating the OA columns with design factors or combinations of factors. These assignments are judged to be independent of each other. The other school [16] allows some columns to be unassigned, in effect allocating these *degrees of freedom* to tracking the effects of potential interactions.

Modifying the general form of the simple additive model (eqn (4)) to include an interaction term:

$$\eta = g_1(x_A) + g_2(x_B) + g_3(x_C) + g_4\left(\frac{x_A}{x_C}\right). \tag{7}$$

This now means that with an incremental change in x_A, say Δx_A, the contribution to η, say $\Delta\eta$ is also dependent on x_C and the coefficients g_1, g_3, and g_4. Indeed, the net effect of Δx_A on $\Delta\eta$ might be in the opposite direction to that without the interaction (eqn (4)). In such cases this is termed negative or *antisynergistic* interaction, and if not included in the experiment plan, renders predictions unreliable for yielding improvement. Where interactions boost the effect of the design factors involved this is termed positive or *synergistic* interaction. The term *superadditivity* has also been used to describe the effects of design factors boosted by interactions.

Interactions may, to some extent at least, be an artefact of the scale, units or metric, or distribution of the original data. In such cases the interaction is considered to be *transformable* and a *data transformation*, expressed as η above, is considered to offer the potential to improve additivity [11, 17, 18]. Thus we seek a suitable transformation (eqn (8)).

$$\eta = h(y). \tag{8}$$

3.1.1.3 Transformations In classic RED it is desirable, when relevant, to differentiate between factor levels that most influence mean effects (*location effects*) and factor levels that minimise variability (*dispersion effects*). Therefore the transformations used often seek to reflect both the mean response and the variability in the response and are sometimes termed *noise performance measures*.

In statistical terms data transformations attempt to enhance three statistical properties of the data [16, 18, 19]:

1. independence between mean and variance of each experimental trial,
2. simplicity of the mathematical model and
3. normality of error distribution.

Non-linear transformations such as $\eta = log(y)$ dominate those used, but have little effect unless the ratio y_{max}/y_{min} of all the data is greater than two or three.

The signal-to-noise ratio (SNR) is a transformation that has been widely used in classic RED although it does not escape statistical criticism [15, 19]. But it does help to simplify the analysis and roughly demonstrates the statistical properties above. Moreover, it is linked to quality loss functions such as eqn (2).

For a set of quality characteristic readings, $y_1, y_2, \ldots, y_n$, the average quality loss, Q, is:

$$Q = \left(\frac{1}{n}\right)\{L(y_1) + L(y_2) + \ldots + L(y_n)\} = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{9}$$

For a 'nominal-is-best' (NB) problem, where M is the target value and the mean is μ, it can be shown that when n is large, Q approaches:

$$Q = k\{(\mu - M)^2 + \sigma^2\}, \tag{10}$$

i.e. the quality loss has two components:

1. $k(\mu - M)^2$, an *accuracy quality loss* proportional to the deviation of the mean from the target;
2. $k\sigma^2$, a *precision quality loss* proportional to the mean squared deviation about the mean.

If Q is adjusted to bring the mean (μ) on target (M) then this first component will disappear and **the** second will be modified by the adjustment. This represents a two-stage optimisation philosophy [9], which is also addressed later in this chapter in model-based RED. The adjustment is to increase each reading by M/μ, which adjusts Q to the quality loss after adjustment, Q_a:

$$Q_a = k\left(\left[\frac{M}{\mu}\right]\sigma\right)^2 = kM^2\left(\frac{\sigma^2}{\mu^2}\right). \tag{11}$$

Attention need only be focused on (μ^2/σ^2), since for a given quality characteristic, k and M are constants. This is the SNR, and as σ^2 is the effect of noise factors and μ^2 is the desirable part of the data, then it can be viewed conceptually as the ratio of power of signal to power of noise.

Therefore, minimising Q_a, the quality loss after adjustment (or sensitivity to noise), is equivalent to maximising the inverse measure of variability proportional to mean, (μ^2/σ^2). It also converts what is in effect a constrained optimisation problem into an unconstrained one as there is only one metric to optimise rather than two, however this conversion does not allow for a thorough search of solution space, as described in Section 2.2. In view of Table 2, a $\log_{10}$ transformation could improve the additivity of the main effects, although generally this is sometimes applied thoughtlessly and is of questionable validity when it is. Thus, the SNR_{NB} based on eqn (11) is expressed in decibels as:

$$\eta = \text{SNR}_{\text{NB}}(\text{dB}) = 10\log_{10}\left(\frac{\mu^2}{\sigma^2}\right). \tag{12}$$

3.1.2 'Model-based' RED

The goals of model-based RED are the same as for classic RED. The difference addressed here is that experimentation is used to build and validate an empirical model of the system that will then be used for engineering design. In the previous section there was some discussion about interactions and their effect on designing experiments. In this section, interactions are considered more generally as part of the experimental design and modelling problem.

We have already discussed the motivation for modelling when direct evaluation of the target system is not possible or feasible given constraints on time and resources. In the case of robust design the motivation for modelling is even stronger as we shall see.

3.1.2.1 Definitions of robustness The subject of robustness in an engineering sense can cover a wide range of concepts such as the ability of a system to cope with unexpected inputs or failure of subsystems or components. The Santa Fe Institute, a research organisation committed to understanding complexity, has a working list of definitions [20]. We repeat here one definition of robustness which can be embodied by the following constrained optimisation problem:

1. attain a target level of performance, subject to
2. minimising variation around that target.

This is related to the SNR of the previous section, but by redefining the problem in this way we retain generality over the problem and can look to the many algorithms available in numerical optimisation to help solve this type of constrained optimisation problem, including GA, global random search algorithms and local optimisation methods such as steepest descent. Of course these algorithms require many evaluations of the system at different settings in their search for optimal solutions, which is why emulators that are fast and accurate statistical models of systems are important in this field.

3.1.2.2 Building accurate emulators By definition, the target system is complex and expensive to evaluate. Complexity in this case means that the relationships between the design factors themselves may be non-linear and in turn their relationship to the systems response(s) may also be non-linear.

In the process of designing an experiment, an early decision to be made is whether to specify the emulator model ahead of performing the experiment or not. If this is possible then it is natural to ask the following question: 'Given a particular emulator, and a fixed number of trials, what is the best experimental design?'

By 'best experimental design' we mean a plan that will extract the maximum amount of information for a given cost, in this case the number of trials in the experiment. This leads to the field of *optimal design*, where certain characteristics of experimental design and model are optimised in order to maximise the efficiency of the experiment. For example, given the following polynomial emulator:

$$y = \phi_0 + \phi_1 x_A + \phi_2 x_B + \phi_3 x_A x_B + \phi_4 x_A^2.$$

We could then ask the question: 'What is the best 7-point design to identify this emulator?'

We can start with a set of seven points placed randomly with values in the range $[-1, +1]$ and optimise them with respect to the chosen desirable characteristic to find the best experiment design. Figure 16 shows the results using *D-optimal* design theory, where the determinant of the information matrix is maximised [21].

As we have discussed, it is generally the case that there is some knowledge of the system, but often not enough to confidently rule out possible interactions between factors. Indeed, it is often the case that even if there is some knowledge of interactions, these assumptions should be tested via experimentation. So, given that it may not be possible, or even desirable, to specify a particular model in advance of experimentation, the question to be asked becomes: 'What is the best experiment design for a fixed number of trials, given no prior assumptions on the model?'

In this case, the best experimental design is one that fills the design space in the most efficient way. Two standard space-filling designs are *latin hypercube sampling* (LHS) designs [22], and *lattice* designs [23]. These experimental designs seek to distribute observations evenly throughout the entire design region. The rationale is that we do not know anything about the behaviour of the system in the design region, so the best we can do is sample this space as evenly as possible.

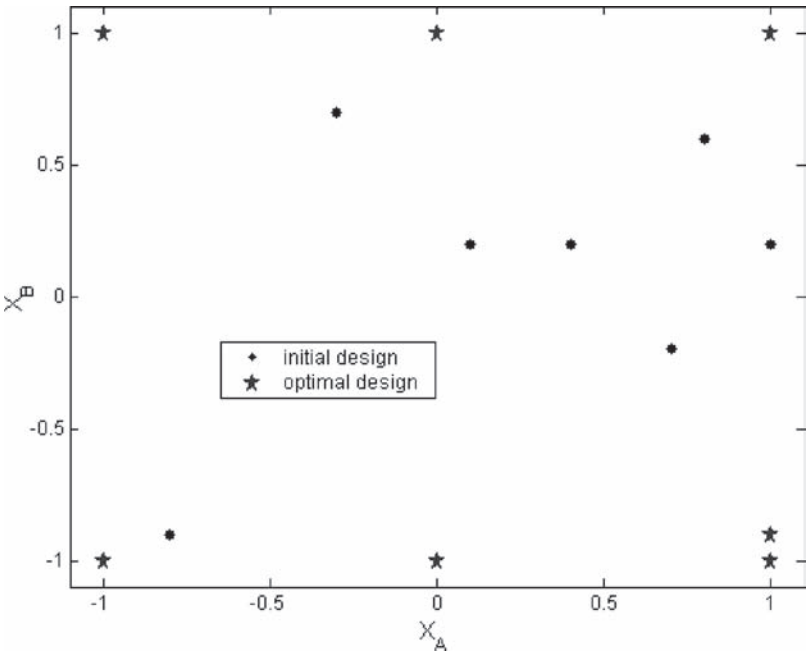


Figure 16: An example D-optimal experiment design.

Other strategies, such as sequential design methods, may also be useful as information gathered by an initial experiment can be used to direct subsequent observations.

Using space-filling designs leads to the use of alternative emulator types, often referred to as spatial models, which seek to characterise the response surface in terms of the distance between observations.

These models do not require any assumptions on the relationships between factors to be made prior to experimentation, and are generally more adaptive than polynomial models. Figure 17 shows an example LHS design with seven points and two factors.

From this brief discussion one can see that there is a strong relationship between experimental design and modelling.

3.1.2.3 Emulator validation Once constructed, the emulator models need to be validated to assess their accuracy. If it is not possible to conduct further trials, then statistical methods such as generalised cross-validation can be used to estimate the accuracy of the emulators [24]. Otherwise additional experiments can be conducted at previously untried settings and the results compared with the equivalent emulator estimates to estimate prediction accuracy.

3.1.2.4 Using emulators for RED After conducting experiments and performing the emulator building and validation process, the emulators can be used for RED. They can be evaluated directly at any point within the design region. In addition sensitivity analysis on the emulators themselves can be used to provide estimates of variability. The main advantage is that this can be achieved quickly, with an evaluation taking seconds, or even less, to perform. This means that designers are more inclined to perform what-if analysis, and a systematic search of the design space (e.g. using a global optimiser) will be more likely to find a globally optimal solution to the design problem.

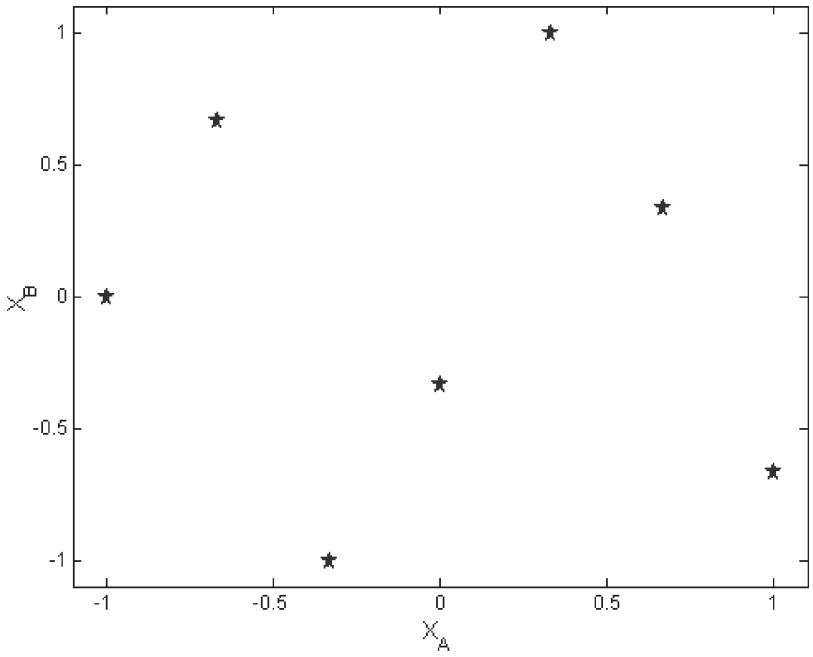


Figure 17: An example LHS design.

3.2 Genetic algorithms

GA are founded on the theory of 'survival of the fittest' combined with the information exchange processes of natural genetics [23, 24]. This information exchange, which is structured yet pseudo-random, forms the basis of the search method. GA rely upon the assumption that in nature, complex non-linear relationships between design factors have to be processed efficiently. Therefore the system under investigation is considered to be a black box in which there are only two aspects of interest, namely the coding of the design configuration and its performance or 'fitness'. The GA procedure is illustrated in Fig. 18.

The starting point is an initial random sample population but too small a sample size risks the GA converging at a local optimum. Fixing of operator values in a GA is difficult as it depends upon problem type, population size, coding and other issues. Thus, wide ranges of values are quoted in the literature [25, 26].

For brevity, let us consider a simple example. An initial sample in a simple design experiment comprising four two-level design factors could be coded as shown in Table 2.

Reproduction progresses typically in terms of giving the design configuration (string) with a higher fitness a greater role in spawning a subsequent generation until fitness values converge at a maximum value.

3.2.1 Matching

One method of matching is to allocate a higher probability of contribution to a dominant string based on its percentage of the total fitness for the generation ('sample' in Fig. 18), as shown in Table 3.

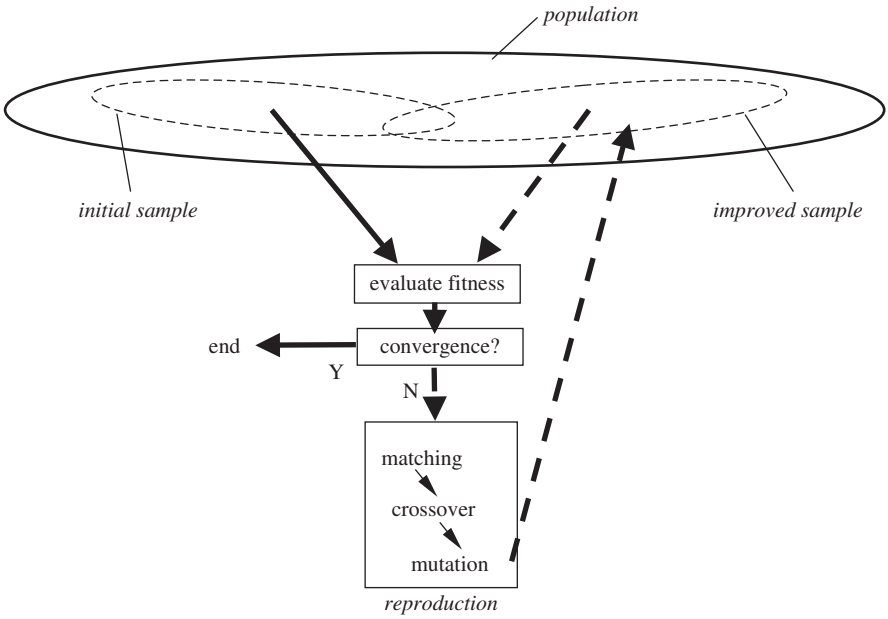


Figure 18: General GA procedure.

Table 2: Initial random GA coding.

	Design factor *A*	Design factor *B*	Design factor *C*	Design factor *D*	Fitness
Experiment 1	0	1	1	0	31
Experiment 2	1	0	1	0	76
Experiment 3	1	1	1	1	48
Experiment 4	1	1	0	0	104

Table 3: Initial random GA coding with matching probability values.

	Design factor *A*	Design factor *B*	Design factor *C*	Design factor *D*	Fitness	% of total
Experiment 1	0	1	1	0	31	12.0
Experiment 2	1	0	1	0	76	29.3
Experiment 3	1	1	1	1	48	18.5
Experiment 4	1	1	0	0	104	40.2
					259	100.0

Table 4: Crossover of the GA coding.

First generation (parents)					
	A	*B*		*C*	*D*
Experiment 2 =	1	0	|	1	0
Experiment 4 =	1	1	|	0	0
Second generation (offspring)					
	A	*B*		*C*	*D*
Experiment 2′ =	1	0	|	0	0
Experiment 4′ =	1	1	|	1	0

Strings selected for reproduction are entered into a mating pool. In Table 3, experiments 2 and 4 would have a relatively high probability of forming a mating pair based on their superior fitness.

3.2.2 Crossover

Crossover tends to pass on desirable traits. A position along the string is chosen as a crossover point, say in between *B* and *C* in Table 5. Codes on either side of this crossover point are then swapped between the mating pair, as indicated in Table 4 below.

3.2.3 Mutation

This plays a secondary but important role in producing a 'random walk' through design space by virtue of an occasional alteration of the value of a design factor.

For example, if the first offspring in the second generation above underwent a random mutation of design factor *A* then perhaps Experiment $2' = \{0000\}$. The incidence of mutations is generally limited to the order of between one per thousand and one hundred per thousand crossover transfers.

In general, further generations would be evaluated until the improvement in fitness converged to the desired level. As the generations unfold it enables the identification of successful combinations of design factors to be identified. These schema or building blocks can then be fixed, which focuses subsequent searches of design space.

3.2.4 Schemata and epistasis

Comparing the code for Experiment 2 and Experiment 4 in Table 3 reveals that two alleles are common to both, namely, *A*1 and *D*0. This 'coadapted' set of alleles can be an indication of significant *epistasis* (interaction) between the two design factors.

A *schema* is a template incorporating a metasymbol, '$*$', to represent all the strings that contain the epistasis in question, i.e. {1**0} for this case. Furthermore, *building blocks* are particularly fit, short schemata and play an important role in the GA. The matching operator tends to be biased towards building blocks that possess higher fitness values thus ensuring their representation. Crossover and mutation have the ability to promote new building blocks but this tends to diminish with the crossover of similar strings. Building blocks tend to increase exponentially as a proportion of the sample population as the search continues – a fact apparently unique to GA and called *implicit parallelism*. Tracking the development of the best schema provides an estimate of the rate of the convergence of the GA.

Thus coding of interactions, i.e. building blocks, is critical to the performance of the GA. For example, simply placing the crossover between interacting alleles will destroy a schema.

Table 5: Crossover of the GA coding with dominance.

First generation (parents including reserved diploid code)					
	A	*B*		*C*	*D*
Experiment 2 =	1	0	|	1	0
	1	*–1*	|	*–1*	*1*
Experiment 4 =	1	1	|	0	0
	–1	*0*	|	*–1*	*0*
Second generation (offspring including reserved diploid code)					
	A	*B*		*C*	*D*
Experiment 2′ =	1	0	|	0	0
	1	*–1*	|	*–1*	*1*
Experiment 4′ =	1	1	|	1	0
	–1	*0*	|	*–1*	*0*

3.2.5 Diploidy and dominance

A *diploid* code is based on the double-stranded chromosome of DNA as opposed to the single strand of haploid organisms, which tend to be relatively uncomplicated life forms. The additional strand provides a mechanism for remembering useful alleles and allele combinations.

Effectively the redundant memory of diploidy permits multiple solutions to the same problem to be carried along with only one particular solution expressed. This helps the diploid population to adapt more quickly, particularly to changes in environment over time, compared with haploid coding.

Dominance identifies which allele takes precedence (is expressed) in genotype–phenotype mapping. This mapping should be allowed to develop.

A three-alphabet or triallelic scheme, -1, 0, 1 combines allele information and dominance mapping at a single position (Table 5). Here 0 dominates -1 and 1 dominates 0.

Comparing Table 5 with Table 4, the resultant code for offspring Experiment 4′ is {1111} instead of {1110} due to the reserved allele $D1$ dominating $D0$. In addition, the reserved status operator shields such alleles from harmful selection in a currently hostile environment. A famous example is the peppered moth where the original white camouflage for lichen covered tree trunks was held in abeyance whilst a black form dominated in areas where trees had been darkened by the industrial revolution.

Mutation places a 'load' on the adaptive plan through its random movements away from the optimal configuration. Therefore it is desirable to keep mutation rate as low as possible, consistent with mutation's role of supplying missing alleles and without affecting the efficiency of the adaptive plan. Under dominance a given minimal rate of occurrence of alleles can be maintained with a mutation rate that is the square of the rate required without dominance. In other words, the robustness of search is enhanced by dominance.

3.3 Comparing model-based RED and GA for the design of cardiovascular stents

3.3.1 Background

It is common for human arteries to become blocked (a stenosis) by disease that can severely restrict blood flow to vital organs. Mechanical cage-like devices, known as cardiovascular stents, are often inserted to dilate these blockages and restore the blood flow. Unfortunately, without the intervention of drugs there is a significant risk that a stented artery will become re-blocked

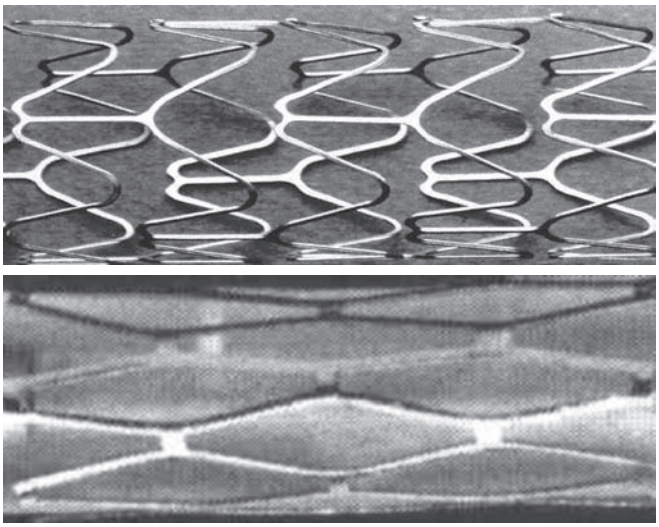


Figure 19: Top: Guidant/ACS Multilink™ stent; bottom: Palmaz Schatz PS153™ stent [27].

(a restenosis). Numerous investigations have identified the flow pattern over the stent to be a key factor and as a consequence elaborate stent patterns have been designed for less disruptive effects on the blood flow. The two successful stent patterns in Fig. 19 can be seen to differ quite markedly, which raises the question 'Are there better untried stent pattern designs?'

Experimenting with new stent designs in live patients is not only a sensitive subject but it is also very difficult to gather flow measurements. *In vitro* experiments are more workable but are also time-consuming and costly. Therefore computer simulations are an attractive option in order to test a large number of stent patterns.

A reasonable first approximation is to model say a 3 mm-diameter artery as an idealised cylinder, however the ratio of overall size to important stent detail, typically 30, severely limits mesh discretisation in the computational fluid dynamics (CFD) model (Fig. 20).

We can simplify this model in two ways in order to improve this meshing. Firstly, assuming that the stent pattern is repeating, the model can focus on a single segment of the pattern (Fig. 21).

Secondly, as the stent diameter is much larger than the thickness of material it is made from, then we can construct a flat model of the partial stent (Fig. 22).

Comparing Figs 20 and 22, the mesh discretisation and hence the fluid flow detail can be observed to be much finer in the partial model for similar computer memory allocation.

3.3.2 Parameterisation for computer experiments

Stents employ a variety of patterns, some elaborate, and the inference is that there are thousands of potential designs. In order to systematically explore the range of possibilities using computer models we must 'parameterise' the pattern, i.e. identify a number of key features or design factors that sufficiently capture the scope of stent design. Continuing our simple approach we can describe the generic repeating stent pattern using five design factors, namely:

1. *Strut thickness:* The thickness of the material from which the stent is cut and having a range of 0.08 mm to 0.10 mm.
2. *Strut section ratio:* Expressed as the ratio of width to thickness, ranging from 1 : 1 to 1.5 : 1.

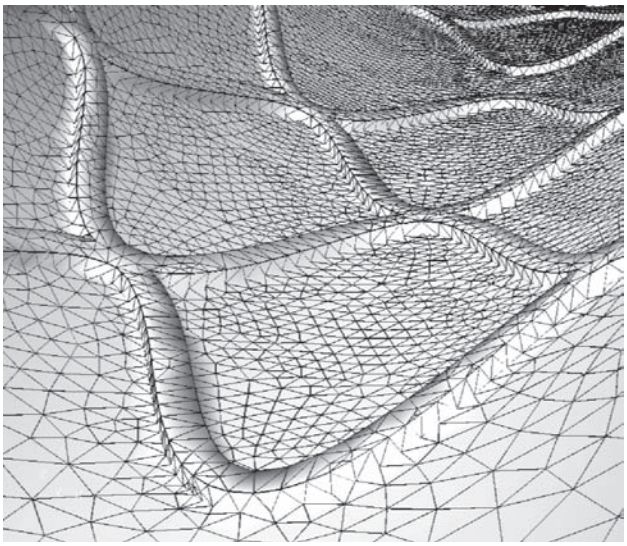


Figure 20: CFD mesh discretisation for full 3D-stent model of PS153™ [28].

Figure 21: Partial model of PS153™ stent cut from a full stent [28].

3. *Pattern skew:* See Fig. 23, defined by the relative position of the peak within one pitch (distance 1.0). Thus a value of 0.5 defines a symmetrical curve and 0.9 produces distinct asymmetry.
4. *Repeating pattern:* Specifies whether a longitudinally adjacent stent segment is merely a copy or a mirror image of the existing segment, i.e. two levels.
5. *Shape order:* Defines the degree of curvature of a segment. Two levels were used.

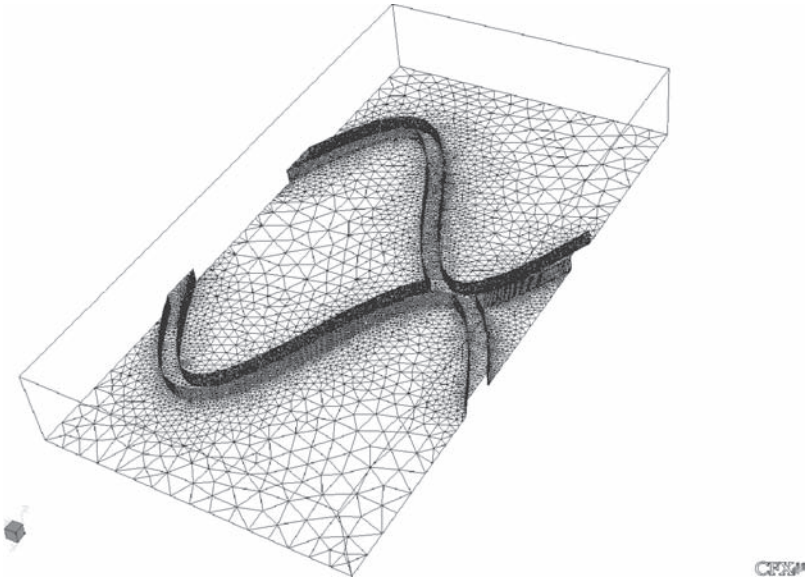


Figure 22: CFD mesh discretisation for partial model of stent.

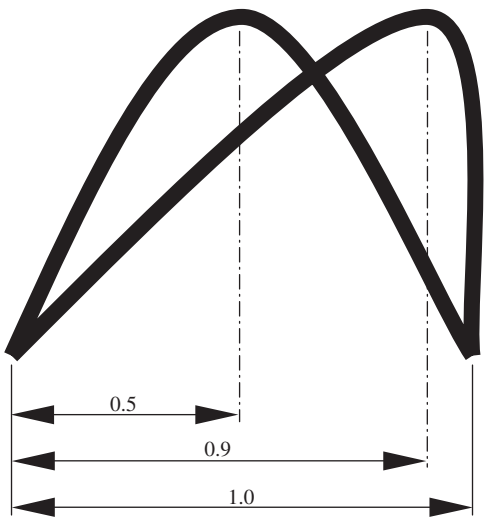


Figure 23: Range of pattern skew.

Repeating pattern and shape order for a symmetrical pattern are illustrated in Fig. 24. The pattern on the left has a sharper '1st order' shape curve and is mirrored, whilst that on the right is a smoother '2nd order' shape curve copied longitudinally. Note that a copied pattern requires a link for structural integrity.

Noise factors were also considered in the model. Firstly, the degree of *strut embedding* in the artery wall, which has the effect of reducing the strut thickness in the CFD model. Secondly, the

Figure 24: Effect of repeating pattern and shape order [28].

Table 6: Chromosome encoding of stent design.

Strut section ratio		Strut thickness		Pattern skew			Copy	Order
0	1	1	0	0	1	1	1	0

flow inlet angle to the partial stent model characterises the different flow conditions a stent design will experience depending upon patient and location.

Flow velocities or wall shear stresses in a 3D-flow field need to be summarised succinctly in order to quantitatively assess the performance of each stent design. We devised a scalar quantity that averaged wall shear stress over the whole surface, termed dissipated power [28] that was inspired by an observation that the diameter of arteries as they branch into smaller arteries do so according to minimisation of energy losses rather than a conservation of total area. Thus a minimum value for dissipated power was sought.

3.3.3 Genetic algorithms

Table 6 summarises the alleles used in the GA 'chromosome' for encoding stent designs. The two alleles used for both strut section ratio and strut thickness have the capacity to represent four values but only three are required. Therefore incorporating a dummy level renders the fourth value in the allele sequence equal to the third. With this encoding the number of unique stent design combinations possible is 288.

With such a short chromosome length and a high simulation cost, the GA parameter settings used in order to avoid extreme local convergence were a population size of 10, crossover probability of 0.75 and mutation probability of 0.02. In our search 11 generations passed before convergence (Fig. 25), involving 20 mutations and 40 crossovers. A total of 27 unique designs were tested under four noise conditions (a total of 108 CFD simulations), covering approximately 10% of the design space available.

The stent design solution at convergence is defined as shown in Table 7.

3.3.4 Model-based RED

Following on from Section 3.1.2, a model-based RED approach using an emulator. This requires the design and noise factors to be continuous parameters and so the two discrete design factors, shape order (1st order) and repeating pattern (mirror), were fixed at the values already confirmed to be the best in initial studies. Thus only 12 trials were necessary (Table 8) in order to predict the response for any set of values for the three design factors and two noise factors.

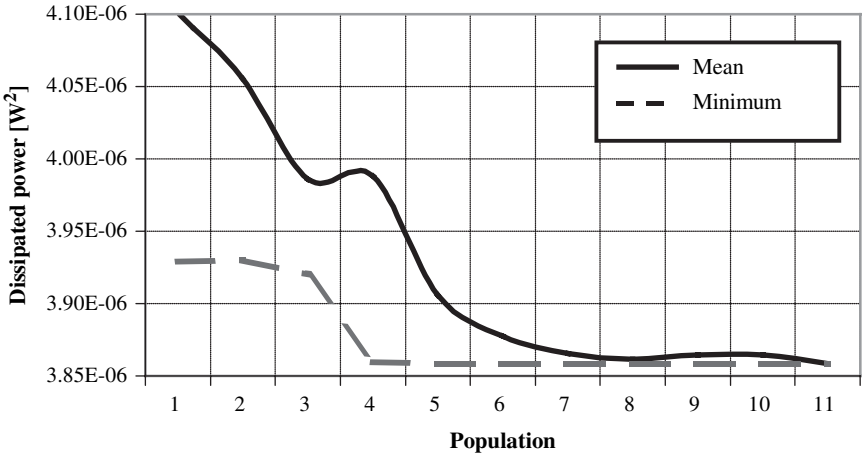


Figure 25: GA convergence [29].

Table 7: 'Optimum' design resulting from GA.

Parameter	Value
Strut section ratio	1 : 1
Strut thickness	0.08
Pattern skew	0.5
Repeating pattern	Mirror
Shape order	1st
Dissipated power (W^2)	92.25×10^{-6}

Table 8: Experimental plan – continuous factor setting [30].

Run no.	Skew	Thickness (mm)	Width ratio	Embedding (%)	Inlet angle (degree)
1	0.65	0.087	1	36.36	60
2	0.57	0.089	3.55	7.27	0
3	0.79	0.093	1.73	21.82	10.91
4	0.72	0.095	2.45	50.91	54.55
5	0.54	0.1	2.09	43.64	32.73
6	0.9	0.098	3.91	58.18	27.27
7	0.68	0.091	4.64	14.55	43.64
8	0.86	0.096	2.82	65.45	21.82
9	0.5	0.08	1.36	0	16.36
10	0.83	0.082	4.27	72.73	5.45
11	0.76	0.084	5	29.09	38.18
12	0.61	0.086	3.18	80	49.09

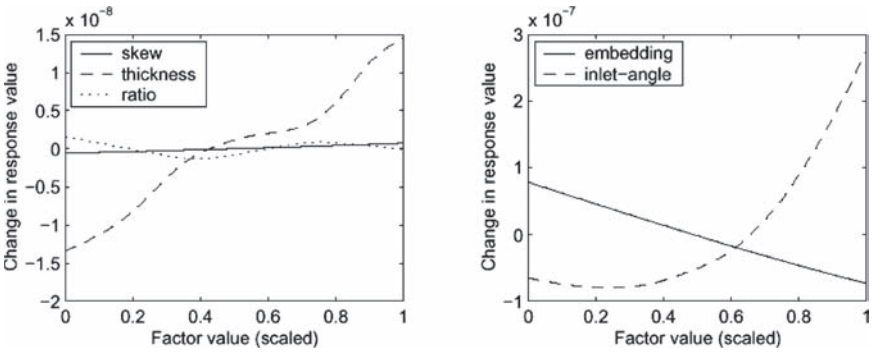


Figure 26: Emulator main effects plot (response values in W^2) [30].

Table 9: 'Optimum' design resulting from RED.

Parameter	Value
Strut section ratio	1 : 1.5
Strut thickness	0.08
Pattern skew	0.518
Repeating pattern	Mirror
Shape order	1st
Dissipated power (W^2)	91.07×10^{-6}

The emulator is combined with a global optimiser in order to determine the values for the three design factors that yield the lowest value for the sum of squares of the dissipated power at the four noise factor settings. Assuming the same treatment for all four discrete factor settings this equates to a maximum of $12 \times 4 = 48$ CFD simulations.

The results plotted in Fig. 26 show the main effects of the factors on the response and it can be seen that the inlet angle noise factor and the strut thickness design factor both have non-linear effects, which is of interest in identifying design solutions that are robust to noise.

The optimum configuration (Table 9) is very similar to that found by the GA but has a slightly better performance.

3.3.5 Discussion

GA and RED treat noise and robustness differently. The use of two levels of noise (high and low) for the GA immediately assumes that noise has a linear effect on the design response, whereas the model-based RED shows that inlet angle has a non-linear effect.

The GA treats continuous design factors as discrete, which restricts the search for an improved design and does not enable an understanding of how the design factors affect the response. However, RED treats the continuous factors as continuous and searches a larger space of designs as a result – but discrete factors must be considered separately. In addition RED provides insight into the design problem through analysis, and this may aid the designer in understanding the design problem and help in finding improved design solutions.

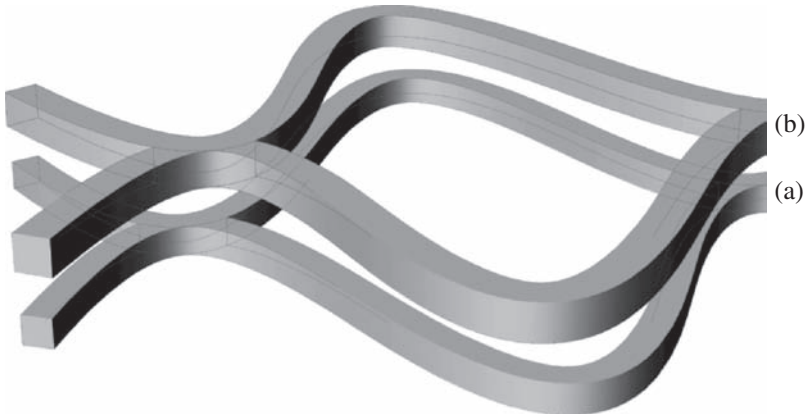


Figure 27: Comparison of the two 'optimum' designs. CFD performance (dissipated power): (a) GA $= 92.25 \times 10^{-6}$ W^2; (b) RED $= 91.08 \times 10^{-6}$ W^2.

Both the GA and the RED searches found improved designs, the RED design giving slightly better results (Fig. 27).

It is also interesting to note that the GA search required $27 \times 4 = 108$ CFD simulations, whereas the RED search required a maximum of $12 \times 4 = 48$ CFD simulations. In this medical engineering example, the stent pattern has to accommodate variations in artery geometry between patients. A more dynamic solution, if it were possible, would adopt whatever shape necessary in order to minimise disruption to the flow for each patient.

A large design space is produced by the few design factors considered. However, neither GA nor RED can search technology options, rather they are parameter searches for the improvement of an existing working principle, in other words adaptation. The GA search converged effectively within a few generations although choosing appropriate values for mutation and crossover was an additional uncertainty in configuring the search.

These studies both highlight some of the challenges involved in automating the redesign process and the importance of incorporating noise into the design process more generally. If sources of noise are not taken into account, then an improved design will not necessarily be robust to them and may fail as a result. The amount of time and resources available to the designer strongly influences the search method. In this regard, the RED method is more efficient as it required less design evaluations than the GA approach, and also achieved a marginally better result. The method of coding a design and the choice of performance measure are critical to the success of any strategy for design improvement. The use of dissipated power as a measure inspired by nature and solely used for improving the design seems to work well.

4 Summary

The powers and limitations of the theory of natural selection are not fully understood even 140 years after Darwin's thesis. How stripes and spots appear is not explained by natural selection [13], it merely suggests that once there the pattern will stay if it offers an advantage. The popular image of natural selection, tirelessly sifting for useful variations among random mutations as the primary source of order, has in extreme cases led to a belief in gene survival as the principal driver above that of the host species. Goodwin [13, 31, 32] insists that the gene's eye view cannot be

complete as some aspects of an organism's form persist in spite of natural selection, not because of it. In the early 20th century, Thompson [33, 34] raised the point that form was not selected, it was inevitable, an argument not inconsistent with Darwin's. Neither of these statements are 'Lamarckian heresy', i.e. the theory that evolution is a response to the environment. However, whilst Thompson was unable to persuade most of his peers of the importance of form and pattern formation, they have begun to remerge in the past two decades as an identifiable field of study.

The explosion in computer power has helped theoretical ideas about patterning that are difficult to test experimentally and the study of complex natural systems has begun to benefit engineering design. Kauffman [9] highlights two limitations to neo-Darwinian theory without self-organisation: Firstly, some systems change their behaviour massively with minor changes to detail. Secondly, accumulation of minor improvements does not always hold. For example, a maximally compressed computer program has no redundancy and therefore it is very fragile to change. Hence starting with a long program becomes progressively more difficult to compress with an evolutionary search because as redundancy is squeezed out there are fewer and fewer clues as to where to search next. Not only that but a minimal program cannot be found by searching every possible configuration, as it could take aeons. Thus redundancy appears to be an essential element in assembling complex systems by adaptive search, and the processes of adaptation and product development are seen to be deeply similar. Therefore design factors, objective function(s) and search methods are intimately linked on the fitness landscape topology, and competition effectively renders the landscape elastic. Adaptive walks progressively worsen on the more rugged landscapes that result from strong interactions between design factors. This helps to explain why complex adaptive systems appear never to reach an endpoint.

What does this mean for the process of designing complex engineering systems?

Mechanisms of robustness are very different between nature and conventional engineering; this is an issue of complexity. Engineering systems have tended to improve robustness through added complexity, which therefore produces new sensitivities. This approach to robust design will persist whilst we wait for the science of complexity to mature. Non-equilibrium may be more prevalent in engineering systems than we realise. Therefore more consideration should be given to the use of data transformations in design experiments to reveal hidden pattern. For example, phase space plots in determining design factor levels dynamic systems or including a term for the rate of entropy production for dissipative systems in key performance indicators. For optimising complex systems, not only is it impractical to search the entire design space but the true best design remains an unknown and so improvement is often a sufficient description and a realistic goal under changing circumstances, and that is why the two words are used interchangeably in engineering.

Which search methods should be employed in engineering design?

It is tempting to see the relevance of our favourite theorems in a complex problem but the *no free lunch theorems* (NFL) [35] show that the *average* performance of *any* pair of algorithms across *all* possible problems is *identical*. This means that if the structure of a problem is not incorporated into a search algorithm then there are no formal assurances that it will be more effective than even a random trial-and-error approach. Generally calculus-based, enumerative and random methods are ruled out because they are too demanding of knowledge and time. We have been unable to consider the full range of search methods based on natural phenomena such as *simulated annealing*, *tabu search* and *ant colony search* [36]. Elements of these may be incorporated into improved search methods and, notwithstanding the NFL theorems, a general theory of optimisation based on nature may yet emerge.

Engineering design is not limited to searching parameter values for improvement. In engineering design, improved global search, limited to the concept design level, has been made by classifying

patented inventions so that an appropriate working principle can be matched to a given problem [1]. We have considered GA and there is a beauty in their global performance through local action. The major shortcoming of GA is their complete dependence upon the 'detectors' (performance measures) to determine the coding, which risks a search too inefficient for expensive engineering experiments. In RED we have seen that optimisation is about finding the underlying system function through physical and empirical modelling. In other words, information gathering is a more overt aspect of RED than GA. Modelling and optimisation can therefore be closely related in engineering design, which accords with the NFL theorems. Apparent conflict between additivity, interactions, orthogonal search and fitness landscapes are tackled differently by RED and GA methods. The issue of interactions needs to be addressed carefully. In classic RED the use of linear models is dominated by additivity concerns, which restricts this approach to smaller regions of design space than the model-based RED approach.

There are several criteria for engineering design algorithms that emerge from the above consideration of search, namely:

1. Design factors often need to be coded as discrete values rather than remain as continuous variables in order to configure a design space.
2. In practice, by optimisation we mean improvement by virtue of selecting the best solution in the search space we have defined rather than the very best from all possible solutions.
3. A useful method for engineering improvement is a trade-off between the more general global search methods and the specialised local search algorithms.
4. It is important to efficiently search a large number of possible solutions without getting stuck at local optima.
5. Probabilistic rules dominate the decision process, which are enhanced when populations rather than individuals form the basis of each search step.
6. Directed search approaches are favoured from amongst the many optimisation methods and algorithms available to engineering, such as GA and RED.

Finally, one must consider the overall resources available and the complexity of the system under investigation when embarking upon a search for an optimal design solution. If there are unlimited resources for experimentation and the system is highly complex (and therefore difficult to model), then a simple random search may prove effective. If only a few observations of system performance are possible, then a more considered approach, involving a carefully designed experiment is likely to be the most appropriate path to follow.

References

[1] Altshuller, G., *The Innovation Algorithm: TRIZ, Systematic Innovation and Technical Creativity*, Technical Innovation Center, 1999.
[2] Atherton, M.A. & Bates, R.A., Bond graph analysis in robust engineering design. *Quality and Reliability Engineering International*, **16**, pp. 325–335, 2000.
[3] Atherton, M.A. & Bates, R.A., Robustness and complexity. *Design in Nature, Volume 1: Nature & Design*, WIT Press: Southampton, pp. 63–84, 2004.
[4] Vining, G.G. & Myers, R.H., Combining Taguchi and response surface philosophies: a dual response approach. *Journal of Quality Technology*, **22**, pp. 38–44, 1990.
[5] Dieter, G., *Engineering Design*, McGraw-Hill, 1983.
[6] Thurston, D.L., Carnahan, J.V. & Liu, T., Optimisation of design utility. *Transactions of ASME Journal of Mechanical Design*, **116**, pp. 801–808, 1997.

[7] Derringer, G. & Suich, R., Simultaneous optimisation of several response variables. *Journal of Quality Technology*, **12(4)**, pp. 214–219, 1980.
[8] Bak, P., *How Nature Works: The Science of Self-organised Criticality*, Oxford University Press: Oxford, pp. 118–121, 1997.
[9] Kauffman, S., *At Home in the Universe: The Search for Laws of Self-organisation and Complexity*, Oxford University Press: Oxford, pp. 216–217, 1995.
[10] Vogel, S., Cat's *Paws and Catapults: Mechanical Worlds of Nature and People*, Penguin Science: London, pp. 246–247, 1998.
[11] Taguchi, G., *Introduction to Quality Engineering*, Asian Productivity Press and UNIPUB, p. 121, 1986.
[12] Atherton, M.A. & Wynn, H.P., Multiple quality objectives in robust engineering design. *Proc. of the 1st Int. Conf. on Quality and its Applications*, Newcastle-upon-Tyne, pp. 537–544, 1991.
[13] Goodwin, B., How *The Leopard Changed Its Spots*, Phoenix: London, pp. 92–96, 1997.
[14] Pareto, V., *Manual of Political Economy*; trans. A.S. Schwier & A.N. Page, Macmillan: London, 1972.
[15] Parks, J.M., On stochastic optimization: Taguchi methods™ demystified; its limitations and fallacy clarified. *Probabilistic Engineering Mechanics*, **16(1)**, pp. 87–101, 2001.
[16] Grove, D.M. & Davis, T.P., *Engineering Quality & Experimental Design*, Longman Scientific & Technical: Harlow, 1992.
[17] León, R.V., Shoemaker, A.C. & Kacker, R.N., Performance measures independent of adjustment. *Technometrics*, **29(3)**, pp. 253–265, 1987.
[18] Box, G.E.P., Signal-to-noise ratios, performance criteria and transformations. *Technometrics*, **30(1)**, pp. 19–27, 1988.
[19] Box, G.E.P., Hunter, G.H. & Hunter, J.S., *Statistics for Experimenters*, John Wiley & Sons: New York, 1978.
[20] Santa Fe Institute, http://discuss.santafe.edu/robustness/stories/storyReader$9, accessed 23 November 2004.
[21] Pukelsheim, F., *Optimal Design of Experiments*, John Wiley & Sons: New York, 1993.
[22] McKay, M.D., Conover, W.J. & Beckman, R.J., A comparison of three methods for selecting values of input variables in the analysis of output from a computer code, *Technometrics*, **21**, pp. 239–245, 1979.
[23] Holland, J.H., *Adaptation in Natural and Artificial Systems: an Introductory Analysis with Applications to Biology, Control, and Artificial Intelligence*, MIT Press: Massachusetts, 1992.
[24] Goldberg, D.E., *Genetic Algorithms: In Search, Optimisation, and Machine Learning*, Addison-Wesley Longman: Massachusetts, 1989.
[25] Fang, K.-T. & Wang, Y., *Number-Theoretic Methods in Statistics*, Chapman & Hall: London, 1994.
[26] Jobson, J.D., *Applied Multivariate Data Analysis, Volume II: Categorical and Multivariate Methods*, Springer-Verlag: New York, 1992.
[27] Serruys, P.W. & Kutryk, M.J.B., *Handbook of Coronary Stents*, 2nd edn, Martin Dunitz: London, 1998.
[28] Atherton, M.A., Tesch, K. & Collins, M.W., Effects of stents under asymmetric inflow conditions. *Biorheology*, **39(3–4)**, pp. 501–506, 2002.
[29] Tesch, K., Atherton, M.A. & Collins, M.W., Genetic algorithm search for stent design improvements. *Adaptive Computing in Design and Manufacture V*, ed. I.C. Parmee Springer-Verlag: New York, pp. 99–107, 2002.

[30] Atherton, M.A., & Bates, R.A., Robust optimization of cardiovascular stents: a comparison of methods. *Engineering Optimization*, **36(2)**, pp. 207–217, 2004.
[31] Goodwin, B., Development as a robust natural process. *Thinking about Biology*, eds. W.D. Stein & F.J. Varela, SFI Studies in the Sciences of Complexity, Lecture Note, Vol. III, Addison-Wesley: New York, pp. 123–148, 1993.
[32] Goodwin, B.C., Kauffman, S. & Murray, J.D., Is morphogenesis an intrinsically robust process? *Journal of Theoretical Biology*, **163**, pp. 135–144, 1993.
[33] Thompson, D.W., *On Growth and Form* (abridged), Cambridge University Press: Cambridge, 1997.
[34] Chaplain, M.A.J., Singh, G.D. & McLachlan, J.C. (eds.), *On Growth and Form: Spatio-temporal Pattern Formation in Biology*, John Wiley & Sons: Chichester, 1999.
[35] Wolpert, D.H. & Macready, W.G., No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation*, **1(1)**, pp. 67–81, 1997.
[36] Song, Y.-H. & Irving, M.R., Optimisation techniques for electrical power systems: Part 2 Heuristic optimisation methods. *IEE Power Engineering Journal*, pp. 151–160, 2001.

This page intentionally left blank

Chapter 13

Living systems, 'total design' and the evolution of the automobile: the significance and application of holistic design methods in automotive design, manufacture and operation

D. Andrews[1], P. Nieuwenhuis[2,3] & P.D. Ewing[4]
[1]*Sustainable Transport Research Centre, London South Bank University, London, UK.*
[2]*Centre for Automotive Industry Research, University of Cardiff, UK.*
[3]*ESRC Centre for Business Relationships, Accountability, Sustainability and Society, Cardiff, UK.*
[4]*Department of Mechanical Engineering, Imperial College of Science, Technology & Medicine, London, UK.*

Abstract

During the latter decades of the 20th century as products became increasingly complex it became necessary to formalise engineering and product design methods, reputed exponents of which include Gerhard Pahl and Wolfgang Beitz, L. Bruce Archer, Nigel Cross and Stuart Pugh. Pugh's 'total design' method is described as a linear activity in that product manufacture and use are considered but not what happens to the product at the end of life. This chapter discusses the need to update the 'total design' model by comparing the product life cycle and that of the automobile in particular with systems and cycles in the natural world.

1 Introduction

A tool is 'a device or implement . . . used to carry out a particular function' [1]. The term tool is usually associated with hand-held devices but, in the broadest sense, all labour-saving devices that have been produced to make the execution of tasks easier, more efficient and precise may be described as tools.

Tools are used by several species of animal including members of the Ape family although none are as complex or sophisticated as those used by human beings (*Homo sapiens*). The simplest tools are natural objects such as stones and cactus spines, the former being used by Egyptian vultures to crack ostrich eggs while the latter are used by the woodpecker Finch to extract grubs from trees [2].

Other found objects may be modified or processed in order to perform a particular task and require greater physical dexterity and mental ability for production. Wild chimpanzees, for example, trim blades of grass to facilitate the extraction of termites from mounds [3] while early humans similarly learned to shape stone to make spearheads for hunting. The use of tools contributed to the development of the human race, society and culture while man's advancing knowledge simultaneously encouraged the development of increasingly complex tools and manufacturing processes. In this context, tools may be seen as a demonstration of man's ability to solve problems.

Some tools function as extensions of the body so that the hammer for example is an extension of the fist, the washing machine – the hands, the bicycle and car – legs and feet and the computer – the brain. In other instances tools enable the completion of tasks that would not otherwise be possible such as cutting diamonds, viewing DNA material through an electron microscope, the prolonged storage of food in a fridge or freezer and flying in an aeroplane. The production of each object involved conscious thought to determine the structure (what the object is), the behaviour (what the object does) and function (what the object is for) [4] and can therefore be said to have been 'designed'.

In traditional craft-based societies, objects are conceived or 'designed' through making and it is unlikely that any drawings or models are produced prior to realisation. The Industrial Revolution encouraged a shift from making by hand to manufacture by machine which in turn necessitated planning (design) prior to production. Consequently in industrialised societies, design and manufacture became separate processes and encouraged the development of the modern design profession.

The design profession is comprised of many specialist areas including fashion, textiles, graphics, illustration, film, theatre, architecture and interiors, industrial, product, automotive and engineering although many projects now involve multi-disciplinary teams. The emphasis of this chapter is holistic automotive design and manufacture and so considers the industrial, engineering product and automotive design disciplines. Although difficult to determine an exact date when these disciplines were first practised, they evolved throughout the 20th century in response to and as a result of the development of automated manufacture and mass-production processes. As products became increasingly complex it became necessary to formalise engineering and product design methods, reputed exponents of which include Gerhard Pahl and Wolfgang Beitz, L. Bruce Archer, Nigel Cross and Stuart Pugh. There are parallels and similarities in each of their design methods and the above experts' respective publications have influenced engineering and design education and the design profession.

Each model advocates an integrated approach to the design activity although, following the publication of *Total Design, Integrated Methods for Successful Product Engineering* in 1990, Pugh is usually associated with the term 'total design' which he defines as

> the systematic activity necessary, from the identification of the market/user need, to the selling of the successful product – an activity that encompasses product, process, people and organisation [5].

In this chapter we use the term 'total design' to describe a holistic approach to design rather than just the method associated with Pugh and we refer to and study other models.

Design methods were developed to help designers, engineers and design managers to create the most appropriate design solutions in response to particular parameters. The British Standards model for design management systems (Fig. 1) [6] describes this complete system. However, the most important element within the system is the product, because without it there would not be any other elements or businesses. Although the emphasis of the above authors' publications is the product development process (PDP) (which influences all other elements), because the system

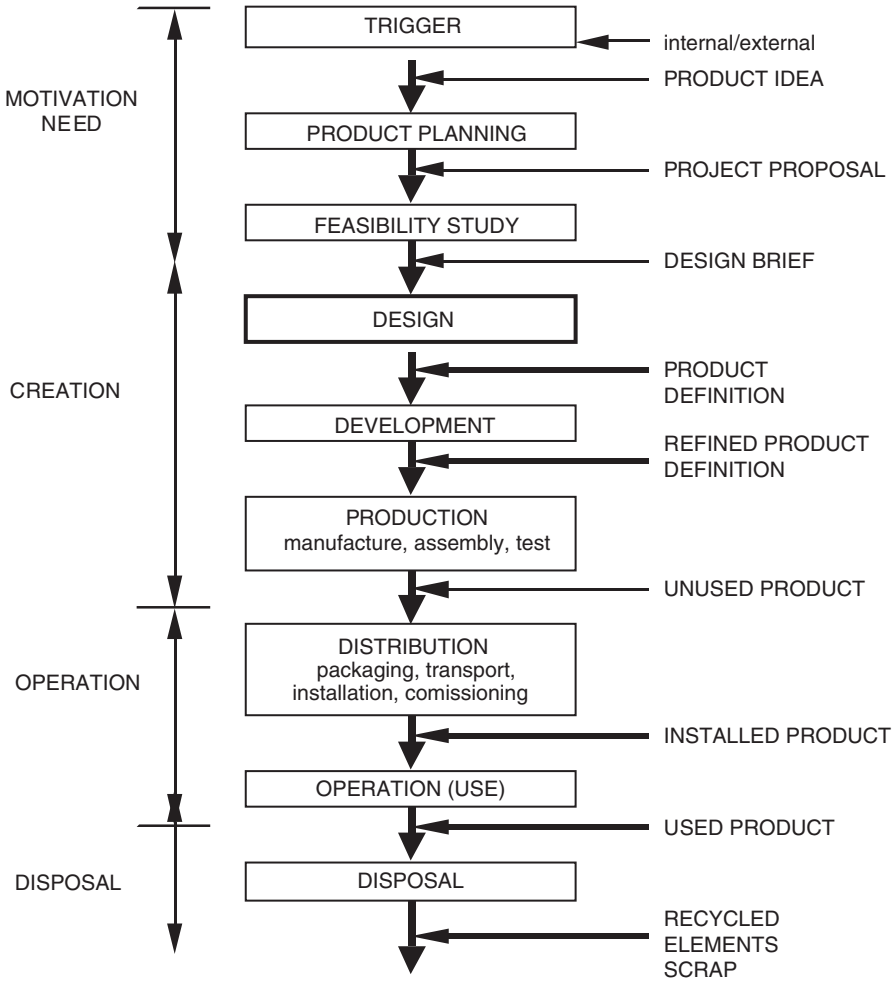


Figure 1: Design management systems based on BS 7000 (Part 10 – 1995: Glossary of terms used in design management), London: BSI (1995).

is *integrated* all other elements should have *some* bearing on the development of the product. However, as will be made evident in our study of the design and manufacture of automobiles, this presumed bearing of all elements is not always borne out in practice.

Holistic and whole systems are not of course discrete to the man-made environment or products. Many systems have been identified and investigated in the natural world and recognition of their significance led to the formation of the International Society for the Systems Sciences (ISSS). Living systems range in size from the sub-atomic to the galactic and have the special characteristics of life, one of which is the ability to maintain a steady state in which the entropy (or disorder) within the system is significantly lower than its non-living surroundings. For our purposes, we may postulate that the essence of life is *process*, expressing the fact that living systems interact with their environment through information and material-energy exchanges and when these processes end, life also ends. Regardless of complexity, every system depends upon the same essential

processes in order to survive and thus continue the propagation of their species or types beyond a single generation. All living systems are self-organising and (most importantly in the context of this chapter) may be described as 'open' because they evolve and adapt in response to change in order to survive [7].

In our chapter entitled 'The evolution of land-based locomotion: the relationship between form and aerodynamics for animals and vehicles with particular reference to solar powered cars' (Volume 1 of *Design in Nature* series) we showed that the forms of the fastest animals on land, in the air and in water are both aerodynamic and energy-efficient and that they evolved in response to and as a result of their natural environment. We further showed that solar powered cars adhere more closely to the laws of nature than mass-market personal transportation vehicles in that they too are aerodynamic and energy-efficient and may be viewed therefore as optimised design solutions. The development of mass-market cars, however, was very different to that of solar cars and their design is inefficient in many respects. Levels of car ownership and car use are going to continue to rise until at least 2050 and, because their overall impact on the environment is detrimental, it is both pertinent and necessary to consider changes in existing automotive design and manufacturing paradigms and the processes that maintain them.

In this chapter then, in order to illustrate the importance and relevance of a holistic (total) approach to design we first discuss several examples of living systems and the value of biomimesis as an aid to innovation. The PDP, the need to update 'total design' methods and life cycle assessment (LCA) are then described in detail. The history and development of the motor car is then compared to those of two other products (namely the radio and personal stereo) in order to appreciate why the contemporary motor car is as it is. A discussion about the need for change and various current trends in the automotive industry follows. Although 'total design' cannot rectify all wrongs in existing products and in the motor car in particular, we propose that if adopted it would improve the car in its own right and so lead to improvements in the environmental system as a whole. We then investigate alternative automotive manufacturing paradigms including the Rocky Mountain Institute 'Hypercar' and GM 'AUTOnomy' and examples of automotive environmental assessment.

We conclude that a great deal can be gained from adopting the principles of living systems and argue that these principles must be incorporated within the 'total design' process in order to solve some of the problems deriving from current automotive design, manufacture and operation and perhaps even for the survival of the automotive industry.

2 Living systems, biomimesis and the 'closed loop' economy

A *system* is a set of connected things or parts of things forming a complex whole [1] and many have been identified in the natural world. These systems have formed the basis for study in various fields and have applications, for example, in medicine and economics. Natural systems vary in size and complexity and range from the sub-atomic to galactic. Regardless of size, the properties (or behaviour) of a system as a whole emerge as a result of the interaction of the components comprising the system. Living systems have the special characteristics of life, the unique properties of which derive from DNA, RNA, protein and some other complex organic molecules, none of which are naturally synthesised outside of cells. Living systems are also self-organising and interact with their environment through information and material-energy exchanges.

We will introduce discussion of Living Systems Theory via the model devised by James Grier Miller. It is a general theory that describes how all living systems 'work', how they maintain themselves and how they develop and change and proposes that there are eight metaphorical levels of living systems (Fig. 2):

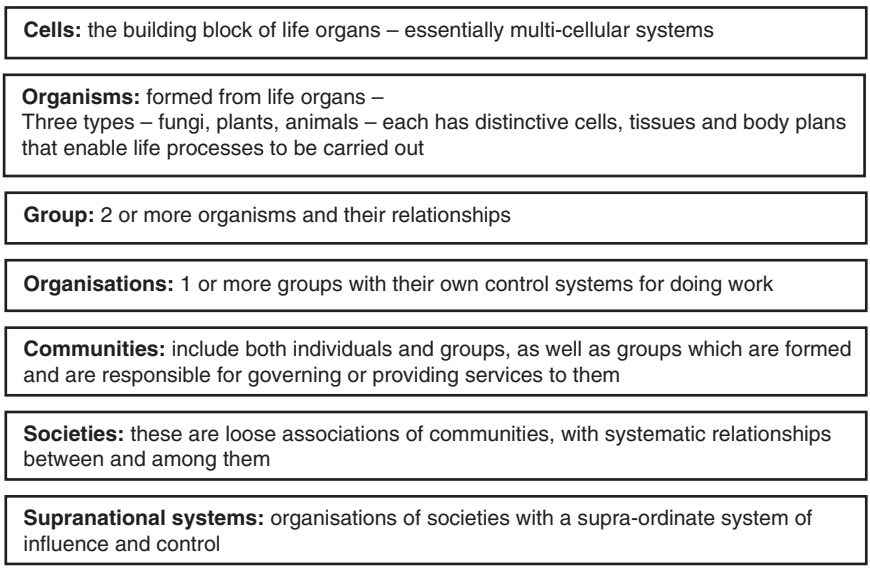


Figure 2: Living Systems Theory: James Grier Miller [7].

Grier also states (as we have already postulated for the current study) that the essence of life is process and that when the processing of material-energy and information ends, life also ends. Likewise whether comparatively simple or complex, each system is dependent on the same essential subsystems (or processes) in order to survive and to thus continue the propagation of the species or types beyond a single generation. Processes function as input–throughput–output, some of which deal with material and energy for the metabolic processes of the system. Other subsystems process information for the co-ordination, guidance and control of the system while some subsystems and their processes are concerned with both. In general, with the possible exception of computers, living systems process more information than non-living systems and maintain their energic state by taking in the required material-energy and information inputs from the environment and, as a result of various processes, discharge information and material-energy back into the environment [7]. Living systems respond to changes in their surroundings and adapt and evolve in order to survive, and in this sense may be described as 'open'. They are also elements within 'loops' or closed systems in that taking, for example, the end of life, residual material will form nutrients and therefore energy for elements within the same or other systems.

2.1 Human physiology and homeostasis

There are many examples of self-organisation and regulation in living systems in the human body. The human body is a highly complex organism, comprised of nine interrelated and interactive systems, namely the endocrine (hormonal), circulatory, digestive, muscular, nervous, reproductive, respiratory, skeletal and immune systems, the normal functions of which are described as physiological processes. These systems and the overall body function are regulated to ensure that a stable equilibrium between the independent elements (known as *homeostasis*) is maintained [8, 9]. Homeostatic mechanisms operate at various levels within organisms and these mechanisms regulate and are regulated by molecules and proteins, cells and tissues and the behaviour of the entire organism. Homeostatic mechanisms also compensate for changes in the environment [10]

so that, for example, a constant average body temperature of 37°C is maintained in humans. Perspiration is one manifestation of the homeostatic control of body temperature.

2.2 Life and reproductive cycles

All living organisms have a finite lifespan. A generalised overview of this begins with division of a nucleus or fertilisation followed by germination or birth, growth and maturation, reproduction and eventual death. All organisms have the potential to reproduce when mature and although some organisms are able to produce either sexually or asexually (examples are water fleas (*Daphnia*), aphids and some fungi) the majority have evolved to exploit one means or the other. Both means of reproduction perpetuate the species and while asexual reproduction allows for rapid population growth and promotes uniformity (except in rare mutations), fertilisation during sexual reproduction combines chromosomes and DNA and thus promotes variability. Both reproductive processes are closed loops, as shown in Figs 3 and 4. The overall life cycle forms another example of a living system closed loop.

In an ideal environment and unless early death is caused by a predator, disease or some other external factor, lifespan varies in length and complexity according to species. Different organisms age at different rates so that, taking three contrasting examples, the average lifespan of *Homo sapiens* is about 75 years, that of a Californian redwood tree (*Sequoia sempervirens*) is 2,500 years and for an 'annual' plant like the wild cornflower (*Centaurea cyanus*) lifecycle lasts for one to three seasons and is complete within one year [11].

During life, organisms undergo an ageing process. This may be genetically pre-determined as suggested by the senescence theory or it can be caused by free radicals. The free-radical theory purports that ageing occurs because certain chemicals (free radicals) are produced as a by-product of biological activity. These are particularly harmful to healthy cells and gradually destroy cells until they can no longer function; as a result whole organ systems break down and eventually the

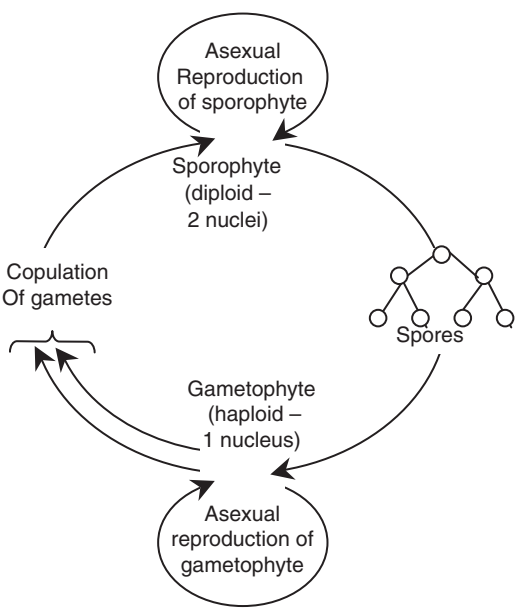


Figure 3: Asexual/sexual reproduction [11].

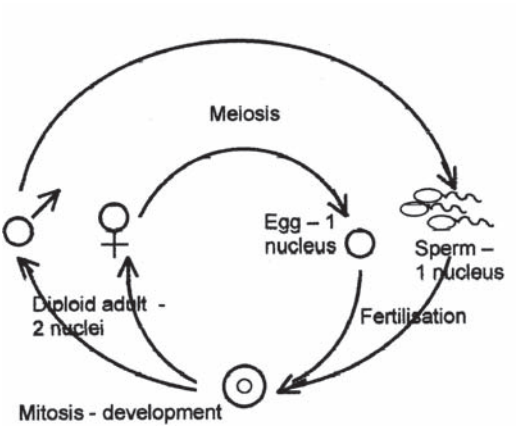


Figure 4: Sexual reproduction in animals and plants [11].

organism dies [12]. Following death, organic plant and animal matter decompose and become nutrient material for other organisms, thus again closing the loop.

2.3 Gaia theory

Many scientists and philosophers have hypothesised that the earth is a living entity and during the 3rd century BC an early exponent, Plato of Athens, described the cosmos as a 'Creature' [13]. One of the more recent proponents of this concept is James Lovelock (a British atmospheric chemist), who developed the Gaia theory during the 1960s in conjunction with Lynn Margulis, an American microbiologist. The Gaia theory is named after the Greek goddess of the earth and here we quote from Lovelock's 1979 exposition *Gaia: A New Look at Life on Earth*:

> ... the physical and chemical condition of the surface of the Earth, of the atmosphere, and of the oceans has been and is actively made fit and comfortable by the presence of life itself. This is in contrast to the conventional wisdom, which held that life adapted to the planetary conditions as it and they evolved their separate ways [14].

Lovelock postulated that the planet is essentially self-regulating and that the earth's surface temperature, atmospheric composition and oceanic salinity are systemically controlled. The processes are highly complex but are described here in simple terms. Lovelock knew that, since life began on earth and despite the fact that the sun's heat rose by 25%, the temperature has remained more or less constant thus suggesting the influence of a regulatory mechanism. Collaboration with Margulis (who was then studying the processes by which living organisms produce and remove gases from the atmosphere) revealed this and other feedback loops, all of which could act as regulating manner.

Now the atmosphere consists of a highly unstable mixture of reactive gases, the composition of which is maintained by the continual removal and replacement of these gases by living organisms. The early atmosphere had a higher concentration of CO_2 than that of today and the consequent greenhouse effect warmed the earth beneficially. A subsequent increase in methane and oxygen was produced by living organisms and cycled through oceans and rocks. This cycling is maintained by evolving species while atmospheric CO_2 is minimised by the biological 'pumping down' of carbon. Without greenhouse gases the earth's surface temperature would be approximately

−19°C. Similarly cells cannot tolerate salinity above 5% and it is apparent that oceanic salinity is similarly regulated to around 3.4%. Regulation derives in part through evaporite beds and lagoons where limestone deposits created by marine organisms are buried. In addition Lovelock suggests that this process also contributes to tectonic movement so that salts are moved from the oceans to landmasses [14].

In essence, Lovecock's theory promotes the idea that the earth is a homeostatic entity (in that self-regulation brings equilibrium) and evolution is the result of co-operative rather than competitive processes. Margulis' later modified the hypothesis suggesting that the earth tends towards homeorhesis: in other words the earth survives because it is in a state of persistent creativity where new organisms evolve continually while others become extinct [15]. Consequently the earth is not a living organism that can live or die all at once, but rather a community of trust which can exist at many discrete levels of integration and where all inhabitants are members of a symbiotic union [16].

Other scientists (including Richard Dawkins) have attacked and criticised the Gaia theory arguing that under no circumstances would natural selection (and Darwinism) lead to altruism on a global scale because such a system would require genetically predetermined foresight and planning by organisms. The existence and evolution of regulatory feedback loops are similarly disputed, as is the concept that Gaia is alive, this latter of course because it (or she) is incapable of reproduction. Further arguments state that the theory is not scientific because it cannot be tested by controlled experiment. Although Lovelock responded by developing a computer model called 'Daisyworld' as mathematical evidence to refute most of these criticisms, the debate continues [17]. Other debate derives from Judeo-Christian theory in which the earth is described as a garden. In this context human beings are defined as custodians of the earth with particular responsibilities [18] rather than as elements within a system. Even if the Gaia theory is inaccurate or inconclusive, the existence of other closed loop and self-regulatory living systems is proven some of which have served as models for man-made systems. We now discuss the importance of what can be learned from nature and *biomimesis*.

2.4 Biomimesis

One of the key factors that distinguishes man from animals is that man knows that he knows something. For example, a horse may know where to place a foot safely to avoid falling but a man knows that he knows where to place his foot safely. For millennia artists, designers and engineers have been inspired by the natural world, some of the earliest remaining examples of which are cave paintings of animals and date from the Ice Age. These images are said to represent more than their depicted subjects because the artists who created them believed that the images had spiritual power.

In addition to the ability to copy and replicate (like a bird learning how to build a nest) *Homo sapiens* also have the ability to abstract and learn from their surroundings, all of which are both natural and progressive. The exploration of nature is necessary for survival but as man moves from substance to meaning and back again, he is also searching for immortality and a meaning in nature itself [19]. Thus painting, sculpture and poetry have become known as the *mimetic arts*, not because they make literal copies but because the surrounding world and ideas inspired their creators. Aristotle discussed this concept at length and in his context, although the Greek *mimesis* literally means to imitate, this definition is somewhat over-simple when applied to the creative arts. When Aristotle wrote that all human action was *mimesis* he meant that it is not merely the holding of a mirror up to nature but that it is 'producing something in the fashion' of nature. He also discussed *poiesis* (which includes all arts and crafts and skills of manufacture as well as

poetry as well as what we understand as poetry in the 21st century) and defined it as 'a general name for calling something into being out of what was not'. Similarly, *techne* was defined as a way of proceeding, and accumulating experience and expertise thereof. He stated that they too are ways of making imitations of nature but like *mimesis*, transcend the literal and copying [20]. It can be appreciated how comprehensively Aristotle studied the range of activities including *our* use of the word design.

Biomimesis (sometimes called bio-mimicry) derives from the Greek *bios* meaning 'the course of human life' (although in modern scientific terms this is extended to include all organic life) and from *mimesis*. But as in the mimetic arts, biomimesis is more than copying nature, rather that it is conscious reference to and inspiration from the natural world, the formalised study of which is becoming increasingly popular. Biomimesis is said to value what can be learned from nature rather than what can be extracted from it and the science of biomimesis regards the natural world as a database of proven solutions from which designs and processes can be imitated and abstracted to solve human problems [21]. Biomimesis may also be political, an example of which is bioregional democracy, wherein political borders conform to natural eco-regions rather than human cultures or the outcomes of prior conflicts. Biomimesis is also used as a guide for economic reforms as advocated by Amory Lovins and others in *Natural Capitalism* where efficient use of energy and materials is rewarded and waste is harshly punished [22]. Based on the premise that all natural life forms minimise energy and material needs as a matter of survival, such a hypothesis mimics the larger process of ecological selection where failures are removed by eco-regions and ecological niches.

Although almost all engineering could be said to be a form of biomimesis, the modern origins of this field are usually attributed to Robert Buckminster Fuller (1895–1983) and its later codification as a field of study to Lynn Margulis. An inventor, architect, engineer, mathematician, poet and cosmologist, Buckminster Fuller is perhaps best known for the development of the geodesic dome. Although his prediction that poverty would be conquered by 2000 remains unfulfilled, the success of the geodesic dome is unquestionable and is the apotheosis and physical embodiment of Buckminster Fuller's philosophy. This structure is cost effective and easy to construct and can cover more space without internal supports than any other enclosure while becoming proportionally lighter and stronger as size is increased. Buckminster Fuller successfully exploited technology to create 'more and more life support for everybody, with less and less resources' [23].

Other examples of biomimesis include the now ubiquitous Velcro, developed by George de Mestral. In 1948 de Mestral noticed that adhesive plant burrs were covered with microscopic stiff hooks. Working with a weaver from a textile mill in France, he perfected the hook-and-loop fastener, which was patented in 1955. Further the layer by layer development of antlers, teeth, bones and shells through bio-mineralisation led to the development of computer-aided rapid fabrication processes. Finally the Wright brothers were birdwatchers and their successful aeroplanes were undoubtedly inspired by their subject [24].

In addition to the development of architectural and engineering solutions and products the natural world also serves as a source of inspiration for the development of non-material and conceptual solutions, one example of which is systems. We have already learned that when life ends, residual material will form nutrients and thus energy for elements within the same or other systems. This principle can also be applied to products at the end of life, but it is important to remember that when dead leaves are dropped from a plant, for example, they cannot be absorbed by that plant or other plants until they are processed by various micro-organisms. Similarly secondary materials produced from product reuse and recycling may need to be processed or to be used for different applications. We now discuss the 'closed loop' economy with particular reference to the automotive industry.

2.5 Learning from living systems and the 'closed-loop' economy

In the theoretical closed-loop (or 'dematerialised') economy, no new raw materials are added, only the existing pool of secondary materials is used, reused and recycled and any energy used during reuse and recycling has to come from renewable non-fossil sources. Similarly any net increase in emissions is not permitted and only readily absorbed emissions can be tolerated, an example of the latter being CO_2-neutral energy generation where the fuel is harvested coppice wood. The CO_2 released in combustion is balanced by absorption of CO_2 by growing coppice. Given the quantity of secondary materials currently available to world economies, with judicious recycling and reuse, a car could indeed be made without extracting additional raw materials, in other words by recycling what has already been extracted in the past. Macquet and Sweet [25], however, challenge the view of 'closed loop' recycling and highlight a number of issues that make it unworkable if discrete to individual industrial sectors. They state that it is only practicable in less complex industries (such as the newsprint industry), although regulators still often adhere to this narrow view.

In the automotive context the degradation of many materials during repeated reprocessing is a major issue. Thus coated steel sheet used for bodies (which is one of the most recyclable of automotive materials) can rarely be economically recycled back to body sheet without adding some percentage of virgin material. Similarly, aluminium sheet is normally downgraded and used in castings after recycling as are many plastics. While thermoplastics are downgraded for less demanding applications, thermoset materials are notoriously difficult to recycle, although they can be used in other ways. In natural processes genuine 'closed loops' are also rare. As stated above, in practice several intermediate steps are needed before the same matter can be reused by its original source. Thus, viewing the 'closed loop' in relation to the wider ecosystem or economic system is more realistic. Jones would describe such systems as 'open loop' [26]. Despite this taking a wider macroeconomic view and applying it to an entire economy, one could still argue in favour of the term 'closed loop'. We therefore use the term 'closed loop' in a wider sense, reflecting a situation whereby the whole of the economy would close the loop, rather than an individual sector.

2.6 Summary

It is evident that, in spite of the philosophical and scientific debate surrounding the Gaia theory, living systems do exist and that the natural world has served and will continue to serve as a source of inspiration and information for designers, engineers and other experts. It is recognised that all natural life forms minimise energy and material needs for survival, that living systems adapt and evolve in response to changes in the environment and that nature involves the principle of macro-economic 'closed loop' and self-regulating systems. We propound that there are great opportunities to learn from and employ these factors as a means of developing more efficient and sustainable products.

In the introduction we noted that man-made and mass-produced products in particular can be said to have been 'designed' and that formal design methods were developed to help designers, design engineers and managers to create the most appropriate solutions in response to given parameters. We now discuss the design process and several specific methods. We include the approach developed by Stuart Pugh known as total design and explain what can be learned from living systems in order to update these models, so re-expressing them as 'closed loop' systems.

3 Total design, process and methods

3.1 The design process

In the design management systems diagram (Fig. 1) we have already recognised that 'design' (PDP) is the most important element because without it, no other elements would exist. 'Design' is a complex and demanding process that can be assisted and enhanced through use of design methods. In reality it is often difficult to separate process and methods because design is an experiential activity and it is not possible to design without doing and therefore to use a method of some sort. Design methods are in the broadest sense procedures, techniques, aids or 'tools' that assist designers and design engineers; it is fair to say that if a product is successful, whatever the means or method used to develop the product can also be viewed as successful. It is important to note, however, that adherence to a design method does not automatically guarantee development of a successful product at the end of the process: the best designers and engineers have innate ability and although ability can be developed through education and practice, it must be present *ab initio*.

The simplest model of the design process consists of four principal stages (Fig. 5) but this is only a descriptive overview and the process is better seen in Archer's prescriptive model (Fig. 6) [27]. This is more systematic and therefore more representative of a design methodology. Design process models vary in complexity but all include the activities as summarised by Pahl and Beitz [28]:

- Clarification of the task: collect information about the requirements to be embodied in the solution and also about the constraints.
- Conceptual design: establish function structures, search for suitable solution principles; combine into concept variants.

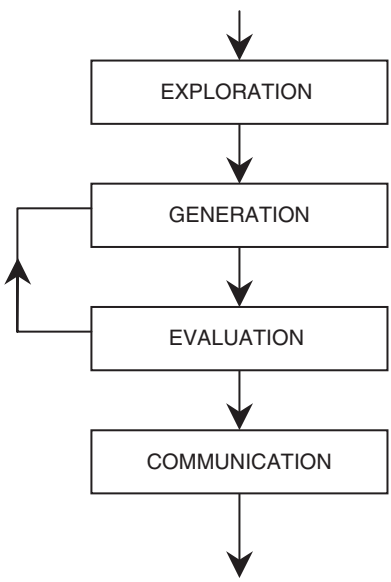


Figure 5: Four-stage design process [5].

- Embodiment design: starting from the concept, the designer determines the layout and forms and develops a technical product or system in accordance with technical and economic considerations.
- Detail design: arrangement, form, dimensions and surface properties of all the individual parts finally laid down; materials specified; technical and economic feasibility rechecked; all drawings and other production documents produced.

Most models suggest that the creative process is a linear activity, which is rarely the case because design involves both divergent and convergent thinking as well as 'flashes' of inspiration and serendipity. In reality designers explore and develop solutions simultaneously as illustrated in Fig. 7 ([40], p. 58). This integrative model describes the essential nature of the design process in which understanding of the problem and the solution co-evolve and contains the seven stages of

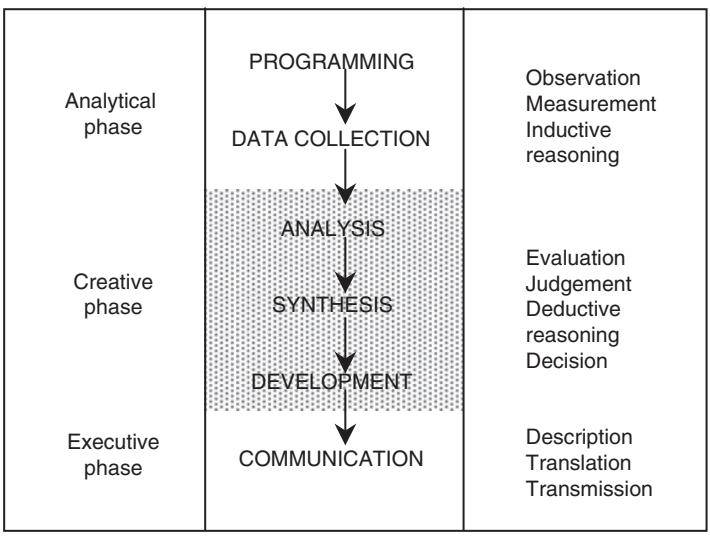


Figure 6: Three-phase model of the design process [27].

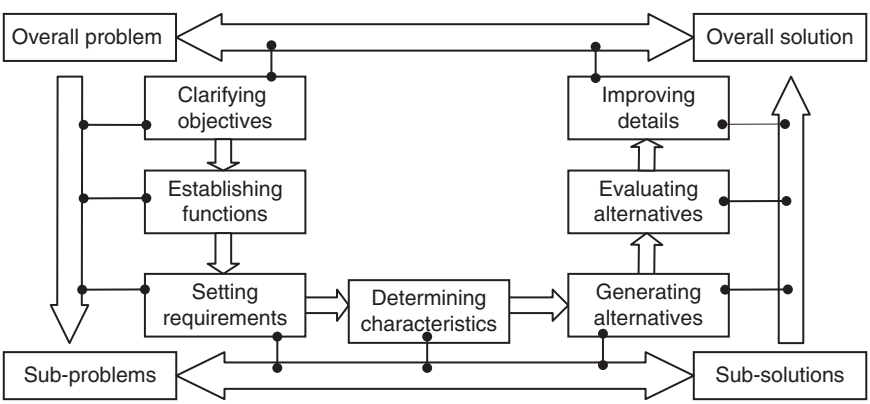


Figure 7: The symmetrical relationships of problems/sub-problems/sub-solutions/solution in design [6].

the design process. This model also describes the design process as a loop and therefore correlates with systems found in the natural world.

3.2 Design methods

The increasing complexity of products and projects, the need to eradicate errors and minimise lead-time from concept to market all lead to a rationalisation of the design process and formalisation of design methods. Design methods combine long-standing or 'traditional' design techniques (e.g. drawing) and concepts from other disciplines such as operational research, decision theory and management sciences. In addition to drawing (and in some instances sketch-modelling), creative and rational methods are used.

Creative methods include brainstorming in groups and synectics, which is analogous thinking that identifies parallels and connections between apparently dissimilar topics and therefore includes biomimesis. Creative methods are particularly appropriate tools for dismantling pre-conceptions about products thus widening the search and potential for innovative ideas. There are numerous accounts of sudden creative insight among highly creative individuals and so the importance of serendipity and the 'flash of inspiration' must also be acknowledged although these instances are invariably the results of background work, prior knowledge of the subject and experience.

Rational methods actively encourage teamwork and group participation and also serve as a checklist to ensure that nothing is overlooked. Different types of rational design methods can be utilised at each of the seven stages of the design process in Fig. 7.

- Clarifying objectives: Objectives tree
  Aim: to clarify design objectives and sub-objectives and the relationships between them
- Establishing functions: Function analysis
  Aim: to establish functions required and the system boundary of a new design
- Setting requirements: Performance specification
  Aim: to make an accurate specification of the performance required of the design solution
- Determining characteristics: Quality function deployment
  Aim: to set targets to be achieved by the engineering characteristics of a product in order to satisfy customer requirements
- Generating alternatives: Morphological chart
  Aim: to generate a complete range of alternative design solutions for the product and thus widen the search for potential new solutions
- Evaluating alternatives: Weighted objectives
  Aim: to compare the utility values of alternative design proposals on the basis of alternatives against differently weighted objectives
- Improving details: Value engineering
  Aim: the contemporary approach is concerned with increasing the product value for the same or lower cost to the producer ([40], pp. 57–58).

3.3 Total design

Pugh defines 'total design' as

> the systematic activity necessary, from the identification of the market/user need, to the selling of the successful product to satisfy that need – an activity that encompasses product, process, people and organisation. ([5], p. 5).

The design core (comprised of market (user need), product design specification, conceptual design, detail design, manufacture and sales) is central to the activity ([40], pp. 57–58). This model (Fig. 8) again describes design as a linear activity, which, as we have already discussed, is not an accurate representation of the design process.

Moreover since the publication of *Total Design, Integrated Methods for Successful Product Engineering* in 1990, awareness of and concern about the environmental impact of the manufacture, use and sustainability of products has grown. It is therefore both pertinent and necessary to update the total design activity model and to consider the product life cycle and 'end of life' within

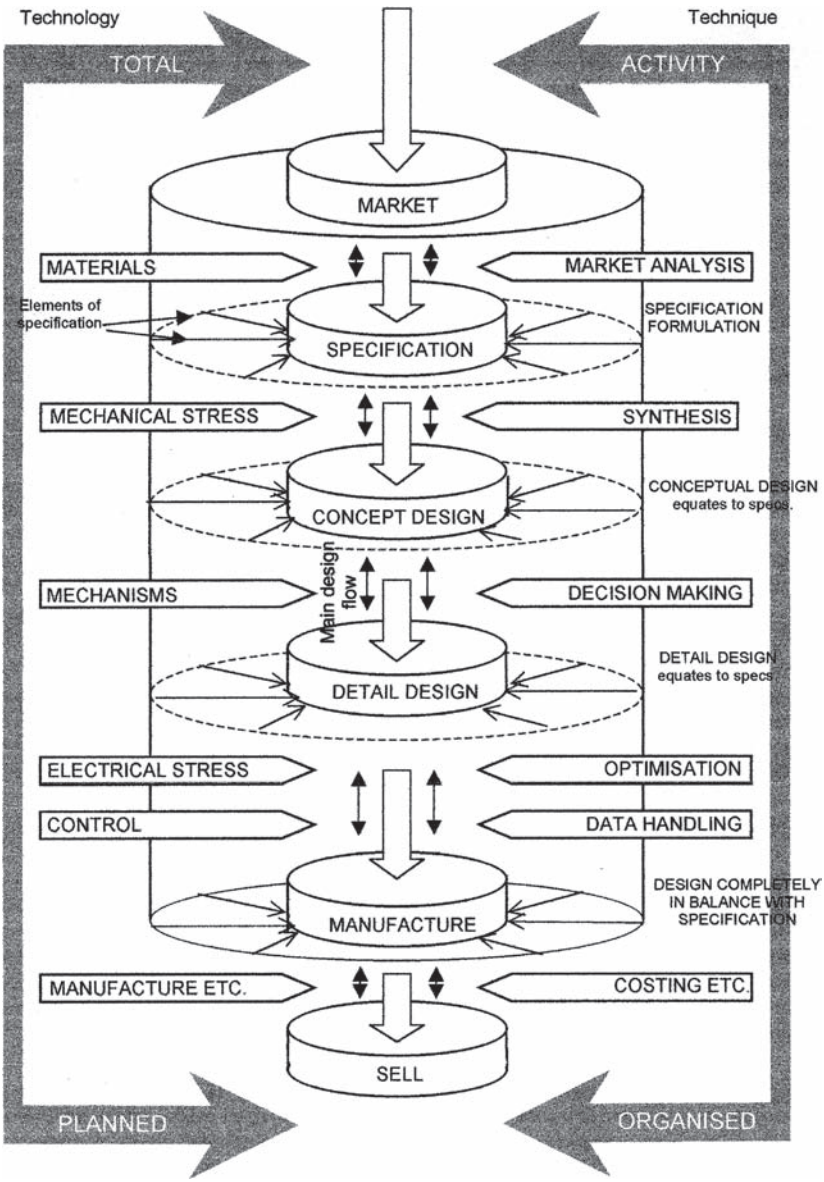


Figure 8: The total design activity model (5).

the process. 'Disposal – recycled elements and scrap' (or 'end of life') is included as the final element of design management systems (Fig. 1) but in many instances, this is currently viewed as additional rather than integral to the design process.

'Closing the loop' and the holistic approach to design are not new concepts and several products have been designed to facilitate dismantling, recycling of materials and reuse of components. Others embody a comprehensive LCA, which incorporates consideration of resource consumption and emissions at all stages of production and use. These are, however, a minute percentage of all products in production. In order to raise designers' awareness of 'end -of life', life cycle assessment issues, this element must now be included within the product design process and 'design core' and could, for example, be placed in the centre of Cross's model (Fig. 7). 'End -of life' and LCA will subsequently have a significant impact on materials selection, manufacturing and assembly processes.

We have already seen that closed loop and self-regulating systems are beneficial in nature where the waste from one process becomes the input/raw material for the next one. This will also be the case when the design loop is closed but will only happen when life cycle assessment becomes a key consideration within the design activity. It is important to note that recycled materials may not necessarily be used to re-make the same component or product but end-of-life product components and materials should be used and will thus be part of a larger system. Nevertheless, only when the loop is closed will the design activity genuinely warrant the term 'total' design.

4 Sustainability and Life Cycle Assessment

The term 'sustainable development' was first coined by the United Nations in 1972 and was later defined by Baroness Bruntland as

> meeting the needs of the present without compromising the ability of future generations to meet their own needs [29–31].

and is therefore a process of 'trade-offs' involving interaction between biological, economic, and social systems [31] in which the ideal is achieved when all three values coincide (Fig. 9).

Life cycle assessment is used to identify what improvements can be made to reduce the environmental impact of existing products or processes and as a guide for the development of

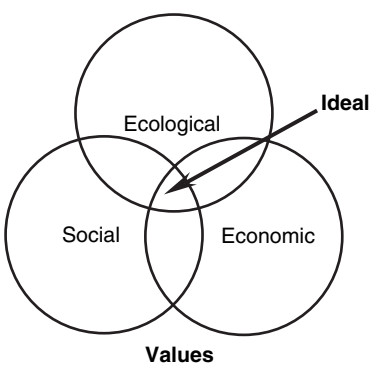


Figure 9: Sustainability [31].

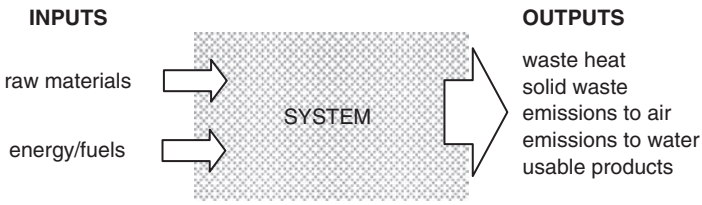


Figure 10: Life cycle analysis based on INCPEN (The Industry Council for Packaging and the Environment) [electronic version], retrieved August 2002 from http://www.incpen.org/html/lca.htm.

new products. The former involves the collection of data to produce an inventory while the latter includes evaluation of the inventory. These techniques determine the environmental aspects and potential impacts of existing and new products and processes at all stages of life from the acquisition of raw materials through manufacture, use and disposal (Fig. 10). Resource (materials and energy) use, human health, and ecological consequences are considered to determine the extent to which a product is sustainable.

In 1998 a series of international standards (ISO 14000) was published and an increasing number of companies officially comply with the directives. ISO 14000 (Environmental Management) specifies requirements for establishing an environmental policy, determining environmental aspects and impacts of products, activities and services, planning environmental objectives and measurable targets, implementation and operation of programmes to meet objectives and targets, checking and corrective action and management review. ISO 14000 was criticised because the measurement of environmental impact is not standardised but the family of standards are regularly updated and improved. Either way the fact that many manufacturers now recognise and adhere to the guidelines is a positive consequence.

Several computer programs and tools have also been developed to assess product life cycle, and perhaps not unexpectedly the results from these differ considerably. This difference is due to variation in product complexity, the number and type of components used and the inevitable complexity of environmental systems. As yet there is no overall agreed methodology for LCA but a consensus is nonetheless beginning to emerge. Until such a consensus is fully reached, providing that the reports from various studies state the methodology used and the assumptions made, LCAs provide a useful indication as to where improvement is needed [32–34].

5 Three product case histories

We have discussed life cycle assessment and established the importance of its inclusion within the ‘total design’ process. In order to appreciate how ‘total design’ can improve existing products (and the automobile in particular), it is necessary to first discover why certain products are ‘the way they are’ and so we now briefly review the history and development of the radio, the personal stereo and the automobile.

5.1 The radio

Initially dependent on face-to-face speech and the written and printed word, several inventions transformed the speed of verbal communication over distance during the latter part of the 19th century. In detail the electric telegraph and Morse code were first demonstrated in 1844,

Alexander Graham Bell patented the telephone in 1876 and Thomas Alva Edison patented his phonograph in 1877.

In 1895 the electrical engineer (and later Nobel prize winner) Guglielmo Marconi succeeded in sending wireless signals over a distance of one and a half miles to become the inventor of the first practical system of wireless telegraphy. In 1896 he took his apparatus to England, and after introduction to the engineer-in-chief of the post office, was granted the world's first patent for the wireless telegraphy system. Marconi's experiments continued and the distance between transmitter and receiver gradually increased leading to wireless communication across the English Channel in 1899. Marconi then proved that radio waves were not affected by the curvature of the earth and transmitted signals across the Atlantic Ocean from Cornwall to Newfoundland in 1901 [35].

Wireless telegraphy was initially used for military and maritime purposes but became an increasingly popular amateur pastime following World War I. By 1922, broadcasting was so popular that a licensing scheme was established to regulate broadcasting in the UK and the British Broadcasting Company (later the British Broadcasting Corporation) was founded. Public wireless radio broadcasting dramatically extended the dissemination of information eventually bringing live entertainment and information into the homes of millions of people [36, 37].

The radio may be considered a *dynamic* product in that, when first conceived, it had no precedent and was therefore a new invention although the radio as we now know it evolved over several decades. The first sets were built by enthusiasts and were essentially examples of technical equipment (Fig. 11). As popularity and the demand for radio receivers grew, manufacturers began to produce sets for domestic use and sought a new aesthetic. Furniture makers were employed to build cabinets for early domestic receivers and so they resembled familiar household items such as linen chests and tallboys. In 1928 Broadcasting House, the purpose-built home of the BBC, was opened in London. This monument to radio technology impressed one manufacturer in particular and E.K. Cole subsequently commissioned one of the architects of Broadcasting House to design a new type of radio. Wells Coates exploited Ekco's plastics manufacturing capability and developed the AD65 Bakelite model in 1932 (Fig. 12) [37]. This circular radio proved seminal and in addition to influencing the design of many other radios, gave the radio a specific product identity.

Innumerable radios were manufactured in various shapes, but until the invention of the transistor in 1948, size was determined by the use of valves. The transistor revolutionised established radio

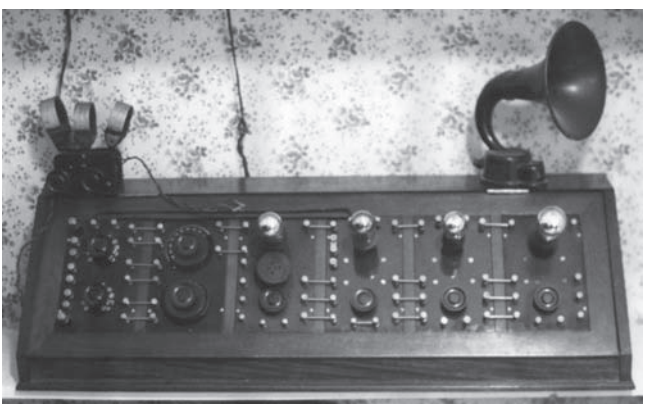


Figure 11: 1919–20 panel set with separate speaker.

Figure 12: Ekco AD65, 1932.

Figure 13: First British transistor set Pye PAM 710, 1956 (at rear).

design by decreasing size and weight and increasing portability. The first pocket model appeared in the US in 1954, followed by Sony's first pocket set in 1955 and the first British transistor radio in 1956 (Fig. 13). Ease of portability and lower purchase costs changed the way in which radio was used so that it became as much an accessory as a product.

Invented in 1936, the television became increasingly popular in the UK during the 1950s and 1960s. Comparatively low purchase cost means that ownership rather than non-ownership of a television is now considered to be the norm and this in conjunction with the development of the transistor lead to its replacement of the radio as a domestic focal point or 'household god' [36].

Nevertheless, whether in the background or foreground, broadcast sound has become ubiquitous in the 'developed' world.

5.2 The personal stereo

Dr Fritz Pfleumer and AEG began to construct magnetic tape recorders in 1931 but the Sony Walkman personal cassette player was not launched until 1979. The first public demonstration of sound tape recorders took place in 1935 and the first live concert was recorded on tape in 1936. Reel-to-reel magnetic tape technology was further developed and led to the introduction of pre-recorded tapes in 1954. Stereo hi-fi technology was introduced to the public in 1958 followed by stereo headphones in 1959 and Philips demonstration of its first compact audiocassette in 1963. This technology was initially used in dictation machines and Philips failed to anticipate the demand for blank tape used for personal music recording. In 1966, American auto-manufacturers began to install eight-track cassette tape players in cars. This technology was swiftly superseded by Philips smaller cassette playback and recording technology, which, along with CDs replaced the LP record to become the dominant consumer music format by 1988 [38].

The Sony Walkman is analogous to the radio in that it is now ubiquitous and *dynamic* because prior to invention there was no comparable product. Sony presents itself as a paradigm of a 'design-led' organisation, which both responds to and creates consumer 'needs' in a highly flexible manner. Many manufacturers invariably carry out market research before developing a product but, like the radio, the Walkman was the result of technological progress rather than market strategy. There are various accounts of the invention of the Walkman concept, one of which propounds that the founders of Sony wanted to be able to listen to music when playing golf and flying. Another account suggests that it was the result of intense internal competition as departments strove to miniaturise and personalise products. Either way the forerunner to the Walkman was The Pressman, a small mono-recorder used by journalists. The fundamental difference between these two products was that the Walkman did not have a recording facility but played in stereo. Sony's founders believed that the personal stereo would succeed and so urged engineers to develop smaller lightweight headphones to complete the product.

Although conceived as a consumer product, the original TPS-L2 model closely resembled office equipment and standard cassette players. While technological developments meant that The Walkman II (WM2) was both smaller and lighter than its antecedent, the Walkman aesthetic evolved so that the cassette was concealed inside a beautiful metallic box. The controls were also moved from the side to become details on the front of the product. These changes all ensured that the personal stereo developed a specific product identity and as a result of 2.5 million sales, like 'Biro' and 'Hoover', 'Walkman' became a generic term. Widespread use of the Walkman has also made an impact on expectations about behaviour. Even though the transistor radio enabled consumers to transport music into public places, listening to the radio and stereo system was regarded as a private activity until the introduction of the Walkman. This product offered the same choice of listening as the domestic stereo but took private listening into the public domain and so blurred the boundaries between the public and private spheres [39, 40].

5.3 A brief history of the motor car

The first petrol-powered, internal combustion-engined car was invented approximately 120 years ago but it was not a completely novel concept. Various antecedents were developed prior to the invention of the car as we now understand it. These include clockwork vehicles built during

the Renaissance, Dutch-built wind-powered land yachts based on Chinese precedents and several generations of steam-powered vehicle. The bicycle could also be regarded as an immediate ancestor of the car and so we will return to this point.

The invention of the car by Daimler and Maybach on the one hand and Carl Benz on the other was made possible by the coming together of a number of enabling technologies in the 1880s. First, the internal combustion engine had been developed by Lenoir, a Luxembourger living in France in the 1860s. This was improved and promoted in Germany primarily by N.A. Otto of Cologne who, using a principle first suggested by Frenchman Beau de Rochas, introduced the four-stroke cycle in 1876. Otto's firm became the engine manufacturer Deutz (today KHD), which also employed Wilhelm Maybach [41]. These new faster running engines provided more power at lower weight, making them suitable for automotive use for the first time.

While Daimler and Maybach first put their engine in a two-wheeler in 1885 – which thus became the first motorcycle. Benz used modern bicycle technology for the first three-wheeler Patent Motorwagen in 1886 (Fig. 14). The bicycle concept contributed bent tube technology, which made the construction of light but stiff chassis frames possible. It also contributed wire wheels, rubber tyres, chains, bearings, steering systems, freewheels and other basic building blocks (Fig. 15). This was historically followed by Daimler's first four-wheeled car built in 1887, which incorporated a wooden chassis and was based on traditional carriage building technology [42].

Reaching maturity just as the car started life, the bicycle was a key enabling technology for the car with Starley's Rover safety bicycle (the prototype of the modern diamond frame bike) first appearing in 1886. It was also a key social and cultural enabler, for it established the market for mechanised personal transport. The bicycle also established the notion of freedom of movement and the 'open road'; concepts still associated with the car today. However, the bicycle is in many respects superior to the car. It is probably the only means of land-based transport where man has improved on nature; cycling is the most energy efficient means of human propulsion, and (as discussed in 'The evolution of land-based locomotion, *Design and Nature*, Vol. 1') even surpasses walking [43].

Frederick Lanchester appears to be the first person to have designed a car from *first principles*. This was described by his brother George in the following terms:

> . . . after absorbing all that was being done on the Continent, he came to the conclusion that they were all crude adaptations of cycles or coach engineering, and decided to start 'de nove', working from first principles [44].

George Lanchester also refers to a remark in *The Autocar* [44] stating that his brother Fred was responsible for 18 of the 36 primary features of the modern car at that time (Fig. 16). Among other advanced features, the Lanchester Company were pioneers of a key enabler of mass production, namely the interchangeability of components and subassemblies. This was also a significant departure from the ways of nature, where no item is directly interchangeable in that nature does not produce exact duplicates. Unlike a replacement car part, no replacement in nature is identical even if a part has the same function. Rather nature aims to achieve functional equivalence without identical replacement. Examples are plentiful and include a lizard's replacement tail (which, although usually shorter, is functionally equivalent to the original), a broken bone that heals (although the development of scar tissue may result in a slight difference in shape), and plant re-growth after pruning. Thus part replacement in the natural world can be seen to resemble the traditional 'craft' approach to manufacturing (where each replacement part is made by hand to fulfil an equivalent function to the part it replaces), rather than mass-production processes that replicate identical components. Although this delayed the introduction of the cars to the market, Lanchester saw the longer-term advantages [45]. He also introduced the use of counterbalancer shafts in order to reduce mechanical vibration, which he considered destructive. Lanchester outsourced very little

Figure 14: Benz 1886.

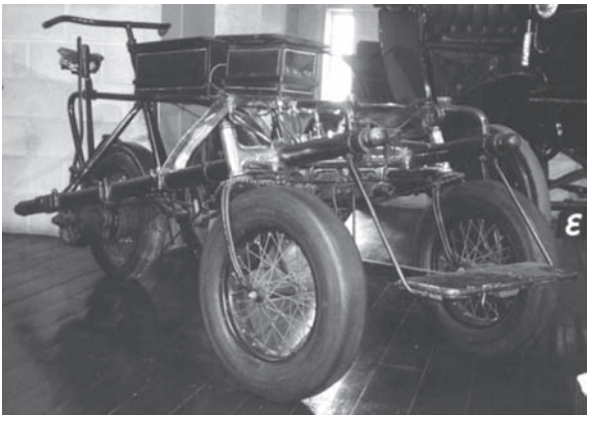


Figure 15: Pennington autocar 1896.

Figure 16: Lanchester 1895. Published with permission from Chris Dowlen.

and Bird [46] describes the first Lanchester as 'the first car to be designed entirely by one man, owing nothing to another source' and so this was probably the first example of 'total design' in a car.

Interchangeability of components became a major feature of industrialisation in the USA. The country lacked the craft skills needed for large-scale production, yet a growing territory and population created an insatiable demand for manufactured goods, which appears to be the essential catalyst for the development of mass-production processes in the USA. Cadillac first used interchangeable components in cars, a system more famously adopted by Henry Ford of Detroit, Michigan, who became the chief advocate of the mass-produced car. Introduced in 1908, the Ford Model 'T' was essentially a craft-built car. These traditional 'craft' techniques were then adapted to mass production as the production system was gradually designed and built around the car itself. For example, in 1913 Ford introduced the moving assembly line without materially changing the design of the car itself (Fig. 17).

Mechanising the production of composite ash-framed, metal, fabric or leather-clad bodies was difficult and so Ford initially failed to mass-produce and construct car bodies successfully. Moreover the main bottleneck in car production at that time was the painting of bodies, which effectively prevented the mass production of complete cars. Firing painted bodies in drying ovens frequently led to the burning of the ash frames. Natural drying was therefore preferred, but this could take weeks, and lighter colours in particular took longer because they contained less pigment. Ford's famous preference for black derived from this because darker colours only took several days to dry. Thus outsourcing most bodies gave Ford greater control over his production process.

The mass production of bodies was therefore a major breakthrough in the establishment of the modern mass car production system. This innovation is due to Edward Budd and his partner Joe Ledwinka [47, 48] who developed the all-steel welded body patented in 1914. This technology became predominant in the industry because it facilitated the development of the unitary body, or monocoque, and the consequent abandonment of the chassis on which Ford had relied. Unitary body technology now dominates the way cars are made and a modern car assembly plant is essentially a machine for making car bodies. In our natural analogy this move represents a move from the internal or endoskeleton of vertebrate structures to invertebrates with an external carapace or exoskeleton. From that point the production system was no longer adapted to the car: the car had to adapt to the production system and cars had to be designed so that they could be made

Figure 17: Ford Model 'T', 1915. Published with permission from Chris Dowlen.

using Budd technology. Consequently design optimised for use was downgraded as a criterion and examples of 'total design' are therefore rarely seen in cars these days.

The very high fixed investment costs of Budd's all-steel body technology mean that costs need to be recovered by making large numbers of cars, an approach to which the technology is ideally suited. Budd technology works best when identical cars are built in very large numbers. Minimum economies of scale occur at volumes of over 100,000 a year, maintained over several years. Optimum economies of scale occur at around one to two million a year although this figure was never actually achieved [49]. Nevertheless, after 120 years of history we have achieved a situation where cars are designed to be produced, rather than to be used; a far from ideal situation from a design perspective. This is also far removed from the way in which nature operates in that our man-made world produces large numbers of absolutely identical individual units.

This design and manufacturing paradigm still dominates modern car design with the exception of those small-scale carmakers who never adopted 'Buddism'. These carmakers still tend to use a separate chassis, maintaining the vertebrate approach to car design. Bodies are made of plastic composites, or hand formed metal. Both techniques allow much greater flexibility of form. This is one of the reasons why TVR's sports cars, for example, have such adventurous forms. Plastic composite unitary, or monocoque, bodies have been made, but are still rare. The first example was the Lotus Elite of 1959; some UK kit cars still use this technique, notably the Welsh Davrian and Darrian club racing cars including the Quantum Coupé. The same holds true for all modern formula one racing cars, which use carbon fibre monocoque construction to make a central 'tub' which holds the driver and to which all mechanical ancillaries are attached.

However, after 120 years of mainstream car design and production we have reached a situation in which large centralised factories, with vast global supply networks, make large numbers of somewhat baroque personal transport devices of debatable fitness for purpose. These are distributed through a complex global logistics system and sold via an extensive world-wide network of largely independent localised dealers. This rather unwieldy organisational activity is accompanied by the fact that few parties in this system are very profitable [48]. Indeed, as is clear from Pelfrey's account of General Motors' history, profitability has been declining over time [50] indicating that the whole system is becoming increasingly economically, as well as environmentally, unsustainable.

5.4 Summary

The radio, Walkman and automobile are all ubiquitous products and the millions sold have encouraged social and cultural change. All three products are similar and were not the result of market forces, rather they were the result of engineering and design inspiration, although the 'inventors' all had prior knowledge and experience as described in Section 3.2. The most significant difference between the radio and Walkman and the automobile is the way in which they evolved: although all were dependent on technological development, neither the radio nor the Walkman had precedents and so can be viewed as new inventions. It must not be forgotten that the radio and the Walkman are considerably smaller and less complex than the car. The number of mass-produced products sold must exceed investment in tooling in order to make a profit, and because initial costs for a radio and a Walkman are lower than that of a car, the cost of redesign and change is far lower than that associated with change in the automotive industry. Moreover mass-production of radios did not really begin until the late 1920s and manufacturers probably therefore learned from more established mass-production processes (including car manufacture).

By the time that the Wells Coates was commissioned to design a radio for E.K. Cole, the influence of the Bauhaus (established in 1919) and Modern Movement in general was spreading. Principal Bauhaus ideologies were integrity of materials, economy of form and that 'form follows function' while modernism in general consciously strove to break with the past and to develop a machine aesthetic [51].

The main component of Wells Coates design for the Ecko AD65 was a press-moulded phenol-formaldehyde (bakelite) body, the form of which was determined in part by both the manufacturing process and behaviour of the material. Although some plastics manufacturers produced objects that aped a craft aesthetic, plastics manufacturing processes were not based on traditional craft processes. These factors and the influence of modernism encouraged Wells Coates to seek a new aesthetic that reflected this new radio technology ([37], pp. 204–206), thus establishing an aesthetic paradigm for audio equipment. When the Walkman was launched in 1979, the public were familiar with comparatively hi-tech products and expected an appropriately hi-tech aesthetic. Similarly mass-production processes had become highly sophisticated and the practice of integrated design and manufacturing thinking became the norm.

Man has used land-based, animal-assisted transport for more than 5,000 years as recorded in Mesopotamian pictographs with engraved images of carts dating from 3200–3100BC [52]. As the direct descendants of horse-drawn carriages and carts the earliest automobiles resembled and were sometimes described as horseless carriages. Bicycle technology influenced and was exploited in car design and manufacture as were coach-building materials and manufacturing processes. Mass-production processes were first developed in the USA in response to the desire and market for consumer goods and Henry Ford among many others recognised and supplied this market. We have already explained the fact that Ford's fundamental error was to design a mass production process around a vehicle initially designed to be made using 'craft' and hand-making techniques rather than adopting first principles and redesigning the vehicle and production process simultaneously. Despite again as we have seen, Budd and Ledwinka's contribution of the all-steel welded body, to this day we have suffered the consequences of energy-inefficient and uneconomical design solutions. Moreover these solutions are contrary to the premise that all natural life forms minimise energy and material needs as a matter of survival.

Just as the earliest radios lacked product identity and were no more than boxes of components with external controls the earliest automobiles were little more than carriages without horses. This is evident in Figs 18–20. Both open and closed models soon developed a more distinctive visual identity as cars in their own right (Fig. 21). Product identity continued to evolve but whether they were luxury vehicles (the Alfa Romeo B 'Lungo' (Fig. 23)) or mass-motoring vehicles (the Austin 7 (Fig. 22) and Hillman Minx (Fig. 24)) all were very definitely 'motor cars'.

The overall product has proved so popular that the number of cars on the road globally now exceeds 600 million [53, 54] and the World Bank has predicted a further rise to 1 billion by 2030 [55]. The level of car ownership is expected to stabilise in developed countries (such as the UK) by 2020 but will continue to rise in some parts of the 'developing' world until 2050 and beyond [56]. The popularity of the motor car and its extensive use have both positive and negative outcomes. We now discuss how and why the car became so ubiquitous, the benefits from and negative outcomes of car use and the urgency for and importance of change in automotive design, manufacture and operation.

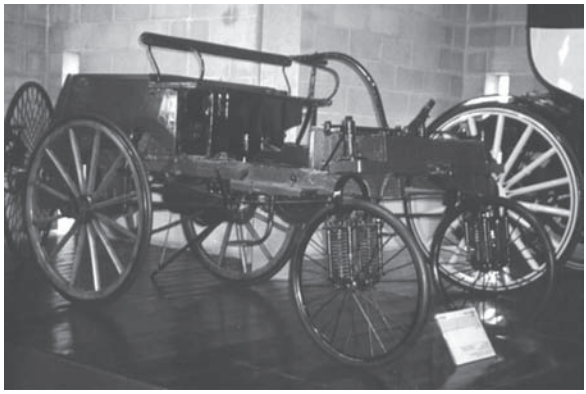


Figure 18: Knight, 1895.

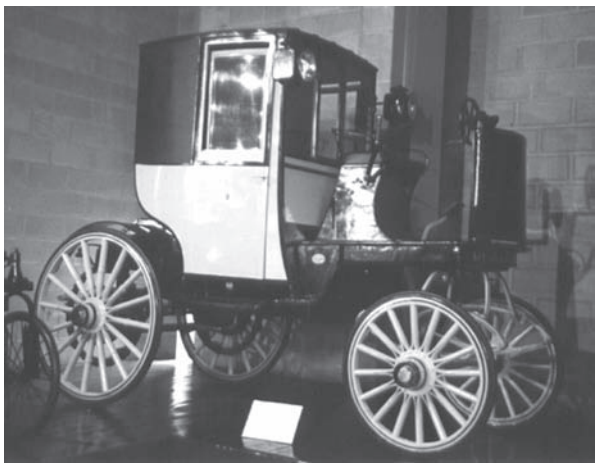


Figure 19: Bersey electric cab 1897.

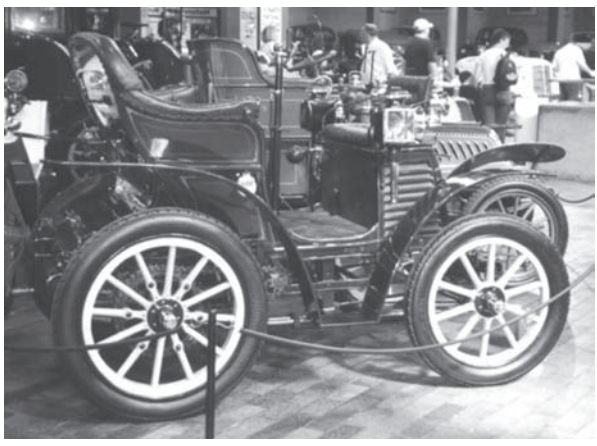


Figure 20: Fiat 3 1/2 hp, 1899.

Figure 21: Bugatti type 15, 1910.

Figure 22: Austin 7 Pearl Cabriolet, 1938.

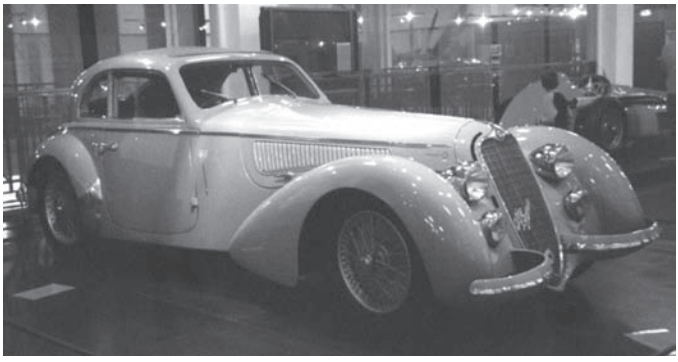


Figure 23: Alfa B Lungo, 1938.

Figure 24: Hillman Minx Magnificent, 1938.

6 The need for change in automotive design, manufacture and operation

Car ownership was facilitated by availability which proved to be synonymous with increased income [57] and following World War II the world economy grew rapidly; between 1951 and 1972 average weekly earnings increased by 83% in Britain. Ownership of registered vehicles in the UK increased concurrently from 2,034,400 at the start of the war, to 3,525,800 in 1955 and 14.7 million in 1975 [58] 11.8 million of which were private cars [59]. Since then the number of cars on the road in the UK has almost doubled and there are now approximately 25 million.

Car use has risen concurrently with car ownership so that in the UK, for example, between 1970 and 1999 the number of annual passenger vehicle miles travelled increased 133% to more than 387.5 billion miles (620 billion km) while the number of journeys by public transport and on foot have either remained static or decreased [60]. The motor car has initiated and contributed to cultural, societal, economic and environmental changes, both positive and negative and many of the latter are related to the way in which cars are designed, manufactured and fuelled.

6.1 The economics of automotive manufacture and use

Mobility has contributed to and is a consequence of economic growth in industrialised countries, while in less developed countries it is viewed as a necessary requirement for sustaining economic growth and a perceived benefit of such growth [57]. Car ownership is therefore indicative of both national and personal success and consequently something to which non-owners aspire. The extent to which individuals value mobility is reflected in their expenditure on it. Although not the highest in Europe, UK expenditure on motoring accounts for approximately 15% of total expenditure on consumables while non-motoring travel accounts for 2% and other household expenditure 83% [61].

Automotive manufacture and use have contributed to global economic development and many national economies are now dependent on both the manufacturing and associated industries while others are reliant on income from oil production. Automotive manufacture has become the largest industry in the world and, until the demise of the Ford plant in Dagenham, for example, the annual turnover of the UK motor industry represented more than 5% of GDP. In the EU automotive and allied industry represents 1.61% of GDP and 8.2% of the total employment of

manufacturing industry. Hence at least 55 million vehicles were manufactured globally in 1999 and almost 2.5 million cars were sold in Britain in 2001 [62]. In addition to revenue from vehicle sales, fuel sales also generate revenue and, the British government collected £2.6 billion from the oil and gas industry in 1999 [60]. Significant change in automotive manufacture and use would create economic instability and is therefore unlikely.

6.2 The role of the car

The car fulfils various roles, all of which contributed to its rise in popularity and growth in use and ownership. These roles can be grouped into two main categories, namely practical and psychological.

Practical roles and practicality include

- increased travel speeds which change perception of time and distance,
- transportation of passengers and goods ([58], p. 39),
- freedom to travel when and where desired [63, 64],
- providing a comfortable travel environment [65],
- rapid and convenient refuelling [66], and
- increased work and leisure opportunities [67].

Psychological roles include

- status symbol – an object of desire so that ownership is an aspiration [68],
- costume [69],
- thrill – changes in perception when travelling at speed provoke physiological responses [70],
- sense of power and control – 'an aphrodisiac' [69],
- a liberator – reduces people's sense of isolation [71], and
- a place of safety, privacy and a 'home from home' [65, 72].

Such is the popularity of the car that the majority of car users claim that they are unwilling to change their habits [73] and perceive the car to be indispensable [74] and the majority of the young people want to own a car as soon as they can [63]. All evidence indicates that that there will not be any decline in levels of car ownership and use.

6.3 The negative outcomes of car use

Extensive car ownership and use has created many environmental and other problems. These include

- changes in infrastructure that
  - damage and destroy ecosystems and the environment [75],
  - limit work and leisure opportunities for those without access to a car or private transport [76],
  - result in social exclusion and isolation [77];
- congestion which
  - prolongs journey times [78],
  - causes stress and contributes to 'road rage' [79], and
  - costs an estimated £20bn per annum in the UK alone [80];

- traffic accidents, injury and death [81, 82]
- detrimental emissions which
  - affect people's health [83, 84],
  - damage both the built environment and natural ecosystems [85], and
  - include CO_2 and other 'greenhouse' gases that contribute to global warming and climate change [86].

All of the latter factors account for enormous hidden costs, which will rise concurrently with car use. For example, the World Bank estimates that traffic accidents alone cost the global economy about 500 billion US$ annually resulting in losses between 1 and 3% of a nation's gross domestic product [87]. Annual recorded road traffic accidents account for 1,171,000 deaths and 10 million injuries globally although the actual figures will be higher because in many parts of the world, road accident fatalities are unreported and so actual costs will also be higher [84]. Nevertheless, all evidence shows that the public currently tolerate these negative factors and will continue to do so in the foreseeable future.

In addition to the social and environmental costs of car use, automobile manufacture requires considerable resources (raw materials and energy) some of which also produce pollutants and emissions, as does vehicle disposal at the end of its life.

6.4 Resource consumption – propulsion fuels

Internal combustion engine vehicles obviously require fossil fuels during use and with average UK car consumption at 30 mpg [88], a total of approximately 2,500 gallons of fuel and 50 gallons of oil [89] will be used during a car's life. Total UK transport fuel consumption in 2000 was equivalent to 55 million tonnes of oil [88] which accounted for 46% of energy consumed in the UK, while the transport sector consumed 32% of all energy used [60].

Oil is a finite resource and, although estimates vary about when known reserves will near exhaustion, many experts believe that this could be by 2050 [90–92]. Current data shows that in the UK, the transport sector consumes by far the greatest proportion of petroleum [93] (Fig. 25) so research into alternative fuels and energy supplies is ongoing the results of which will have a significant impact on the design of future vehicles.

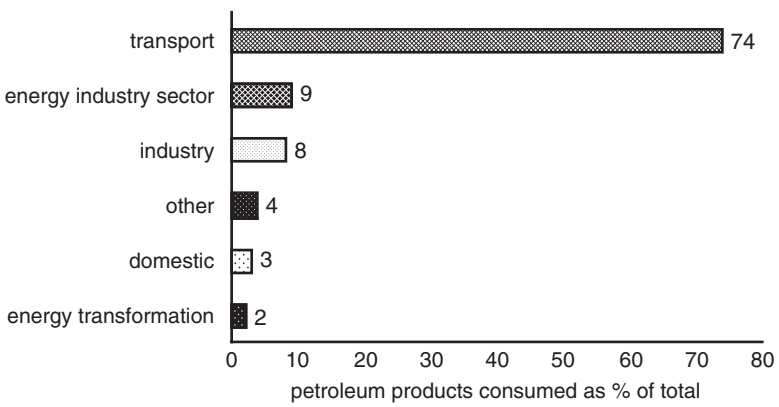


Figure 25: Petroleum products used for energy in UK, 2000 [60].

Current design and manufacturing processes mean that cars are highly inefficient. This is despite the fact that during the past 10 years in particular, manufacturers have sought to improve fuel economy through more efficient engines and transmission and by reducing weight. The latter is achieved by increasing the percentage of plastics, composites and aluminium used. Nevertheless, only about 2% of energy consumed for operation is used to transport the average number of passengers (1.5) while the remaining 98% is used to transport the vehicle itself [94]. Similarly no more than 20% of available energy is used to turn the wheels while the remaining 80% is lost as heat and exhaust ([22], p. 24). This is a reflection of the second law of thermodynamics consequence for equilibrium heat engines rejecting a substantial amount of the energy input in the form of heat.

6.5 Automotive manufacture – materials and energy consumption

In 1999, 16,734,250 trucks and 38,946,820 cars were sold globally, meaning that approximately these numbers were manufactured [95], thus accounting for a considerable quantity of raw materials. A typical vehicle is constructed from high-strength carbon, stainless and other steels, iron, plastics and composites, lubricants and fluids, rubber, aluminium, glass, copper and zinc. Cars vary in weight from the small 2-seat Smart car (MCC Smart City Coupe) at 720 kg (1,584 lbs) to the Rolls Royce Silver Seraph at 2,350 kg (5,170 lbs) as does the ratio of materials used. In order to reduce car body weight and improve fuel economy, during the past 10 years in particular, manufacturers have increased the percentage of plastics, composites and aluminium used. The percentage breakdown of materials by weight in a typical current European model is shown in Fig. 26.

Reports about the amount of energy used during car manufacture vary considerably according to source and definition of product life cycle [96, 97] with lower estimates being 6% [98] and higher estimates at exceeding 16% [99] of total energy consumption during vehicle life. Nevertheless, the average amount of energy consumed in 1999 was equivalent to 0.61% of the total annual global primary energy consumption of 380 quads [100] (1 quad = 1 quadrillion British thermal units or 25,199,577.242763 tonne of oil equivalent).

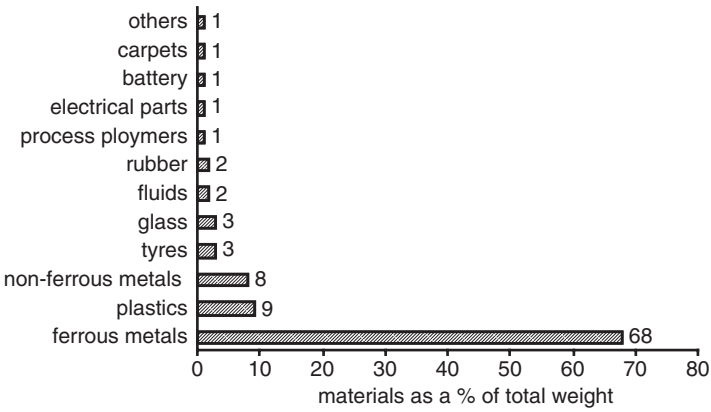


Figure 26: Average 1998 car – material breakdown by weight.

6.6 Vehicle disposal at end of product lifetime

Between 1.3 and 1.8 million end-of-life vehicles (ELVs) are scrapped in the UK every year [101] with total ELV scrap in the EU generating between 8 and 9 million tonnes of waste [102]. At the end of useful life cars were generally sold to a vehicle dismantler but changes in policy now mean that dismantlers are usually paid to dispose of vehicles. It has long been the case that engines, gearboxes and similar components have been removed and re-manufactured in what is known as the 'core' recycling business. Dismantlers also have to remove potentially environmentally polluting materials such as operating fluids and batteries, and parts that can be sold on to undergo a shredding operation where the hulk is broken into fist sized parts. In subsequent processing the ferrous and non-ferrous metals are extracted and recycled, and the residual non-metallic material is sent to landfill sites. In the UK between approximately 74% of a vehicle is currently recycled (64% of materials are recycled and 10% of parts are reused) while approximately 490,000 tonnes of the remaining total material from ELVS is buried in landfill sites annually [103].

6.7 Summary

It is evident that current cars are inefficient in that the majority of fuel consumed during operation is used to transport the vehicle rather than goods and passengers. Similarly mainstream automotive propulsion technologies require fossil fuels, which are a finite resource and produce detrimental emissions and pollutants. These latter impact on people's well-being and quality of life and damages the environment at local, national and global levels. Although automotive manufacture, oil and associated industries employ millions of workers and are considered essential for economic growth, a considerable percentage of income generated is consumed by the hidden costs deriving from extensive car ownership and use. Current automotive design and manufacturing processes also account for a notable percentage of global energy production while current policy means that approximately 490,000 tonnes of material from ELVs ends up as landfill in the UK every year. Nevertheless, all evidence shows that the car is integral to 'developed' societies and its ownership an aspiration in 'developing' societies and that it is therefore here to stay.

It is therefore essential to reduce the environmental impact of vehicles for personal transportation as much as possible. We have already discussed the development of current design and manufacturing paradigms and will now consider the impact of impending changes and what further beneficial changes should be introduced. It is increasingly evident that these changes must include the adoption of a truly 'total design' process.

7 Current trends in automotive design and manufacture

In Section 5.3 we discussed the development of the automobile. Some sources (including Pugh) state that the conceptual plateau for the contemporary car was initially established with the introduction of the Ford model 'T' (Fig. 27) and that the majority of cars still incorporate

- a body with windows and doors,
- four wheels, one at each corner,
- a front engine with rear wheel drive and brakes on each wheel,
- front and rear lights,
- pneumatic tyres, and
- front and rear bumpers (fenders) ([5], pp. 158–160).

Figure 27: Ford model 'T', 1924. Published with permission from Chris Dowlen.

This analysis is, however, somewhat naïve and many of the product characteristics listed are somewhat spurious. The 'conceptual plateau' above was not really established with the model T, and, as we argued earlier, the model T was in no way innovative as a product. It was based on an existing tradition of car design, the so-called 'système Panhard', established by the French firm of that name. This embodied the basic concept of: front engine and rear wheel drive, as listed by Pugh. Even at the time Pugh was writing (1990) the majority of carmakers had changed to the front wheel drive system that prevails to this day. BMW, Lexus and Jaguar (with the exception of the X-type) are the only mainstream carmakers who still use rear-wheel drive. However, many so-called sports utility vehicles (SUVs) and pick-up trucks (which together with multi-purpose vehicles (MPVs) now represent over half the US market) have retained the separate chassis found in the model T and its contemporaries, although even here the pre-Budd paradigm is now under threat. We have also established that Ford did not introduce the 'manufacturing paradigm' still prevalent today. Although he made a significant contribution, the key element of the modern car-manufacturing paradigm is due to Budd [48]. Nevertheless, the car has been designed around production methods and the manufacturing paradigm remains more or less unchanged since the introduction of 'Buddism' and the overall concept is thus described as *static* rather than *dynamic*.

Developments in the subsystems such as engine technologies and materials selection have occurred but they tend to be incremental, as are modifications in vehicle aesthetics and styling. Thus contemporary vehicles look very different to those produced at the beginning of the 20th century but again change was gradual (Figs 28–30). Alfred Sloan, a manager at General Motors, developed 'Sloanism', to maximise profit through the annual introduction of superficially modified, fashionable models [104]. This practice was first introduced during the 1920s but the annual model change characteristic of US carmakers in the 1950s and 1960s has largely disappeared because it led to high costs, lowering of residual values in the market and consumer resentment. Sloan also introduced the concept of the product range within each brand and beyond that a range of brands for each manufacturer. General Motors pioneered the concept from the 1920s onwards, with Chevrolet as the entry level, the next step up being variously Pontiac and Oakland, then Buick and Oldsmobile and La Salle with Cadillac at the top of the range. Makes such as Oakland and La Salle have come and gone over time and Oldsmobile (which appealed primarily

Figure 28: Cadillac, 1960s.

Figure 29: MGA, 1956.

Figure 30: VW Beetle, 2001.

to import buyers), were recently phased out. Nevertheless, this system is now mirrored by Ford, Fiat, Volkswagen and others, who offer a range of brands in different market segments [48].

The introduction of robots automated many of the processes once carried out by a human workforce and again, the majority of vehicles are produced on an assembly line similar to that first conceived by Ford in 1908 [105]. Concern about the eventual depletion of oil reserves, pollutants, emissions and CO_2 in particular has encouraged research into and development of

more fuel efficient fossil-fuelled vehicles and alternative fuel and power sources while problems deriving from disposal of vehicles at end of life have encouraged EU legislation relating to the recycling and reuse of automotive materials and components. Although these factors are significant, evidence to date confirms that changes arising from the above and the overall trend in vehicle design and manufacture remain incremental.

7.1 Internal combustion engine vehicles

Since the oil crisis in the early 1970s vehicle manufacturers have sought to improve fuel economy by developing engine technologies, reducing aerodynamic drag and vehicle weight. Market forces and, more recently, government schemes such as the Partnership for New Generation Vehicles (PNGV) in the USA [106, 107] and the Foresight Vehicle and 'Clean Fuels' programmes in the UK [108] encouraged these developments as has a voluntary agreement between the European Commission and European Automobile Manufacturers Association (ACEA) reached in 1998 [109]. This committed manufacturers to reduce the CO_2 emissions from new passenger cars by over 25% to an average of 140 g/km by 2008 and was eventually followed by similar agreements in Korea and Japan.

PNGV stipulated that vehicles should be three times as fuel efficient as current conventional, mid-sized sedans, while improving recyclability and maintaining comparable performance, utility, safety, and cost of ownership. Ford subsequently produced the Prodigy (and later P2000), DaimlerChrysler, the Dodge ESX3 and GM, the Precept. The P2000 is an internal combustion engine vehicle constructed from aluminium, titanium, composites and other low weight materials and is therefore 40% lighter than a conventional vehicle and capable of 70 mpg. The ESX3 and Precept both exploit 'mybrid' or mild internal combustion–electric hybrid diesel technologies and are thus capable of 72 mpg and 80 mpg respectively [106]. The PNGV was cancelled by the Bush administration and replaced in January 2002 by FreedomCAR (for Cooperative Automotive Research). Energy conservation is no longer the primary issue and has been replaced by a move to hydrogen power produced from domestic renewable sources in order to reduce the country's reliance on imported oil. FreedomCAR is a public–private partnership between the US Department of Energy and the 'Big3', namely GM, Ford and DaimlerChrysler and although the new project refers to PNGV achievements, there is little if any intention of altering the typical US product characteristics in the way PNGV did. The vehicles developed under PNGV therefore remain concepts although it is possible that what was learned will influence car design in the future [110].

In April 2001 the British government also introduced graduated vehicle excise duty (VED) so that new (and generally smaller) cars with lower CO_2 emission levels are now subject to lower duty [109]. During the 1990s the percentage of smaller cars on the road increased and the Mercedes Smart Car (Fig. 31), for example, achieved cult status immediately after launch. However, it is too early to assess whether graduated VED has further influenced the choice of vehicle.

Diesel cars have been popular in mainland Europe for many years and have become increasingly common in the UK because distance travelled per litre is higher than that of petrol. CO_2 emissions are lower than from petrol and petrochemical research has led to reduced sulphur emissions [109], both of which are environmentally beneficial. However, diesel still produces damaging particulates and one detrimental emission is therefore exchanged for another. Nevertheless, research into various aspects of diesel and related engine technology is ongoing and lower fuel consumption could also encourage an increase in ownership of diesel cars ([54], pp. 28–29) [111].

The significance and impact of vehicle aerodynamics were discussed at length in *Design and Nature*, Volume 1 (The evolution of land-based locomotion: the relationship between form and

Figure 31: MCC Smart City Coupe, 2000.

aerodynamics for animals and vehicles with particular reference to solar powered cars), which demonstrated how a low coefficient of drag (C_d) contributes to energy efficiency. Most manufacturers have sought to reduce drag [112], one example being the Volkswagen Group who achieved a 30% reduction in their most popular models between 1980 and 1990. This is equivalent to a 10% reduction in fuel consumption further reductions of which will be concurrent with further reductions in drag [113].

In May 2002 Volkswagen revealed their latest concept car: a two-seater car capable of travelling 100 km on 1 l of fuel. This car is constructed from composite materials including carbon fibre, aluminium and magnesium and at 290 kg, weighs 430 kg less than the Mercedes Smart Car. At 3.5 m it is only 1 m longer than the Smart Car despite the fact that the driver and passenger sit in tandem (one behind the other). The single cylinder mid-mounted engine is a high-efficiency, direct-injection naturally aspirated diesel engine. These features in conjunction with low road resistance (comprised of aerodynamic drag and road/tyre friction) contribute to the fuel economy. The streamlined overall body form and reduced frontal area (A) mean that this car has a far lower coefficient of drag (C_dA) than other cars on the road. The tandem seating means that the car is 1.25 m wide and 1 m high. This makes for a frontal area of 1.25 m^2 which is half that of an average family saloon car at approximately 2.5 m^2. The composite material wheels are only 400 mm in diameter, which with low rolling resistance tyres also contribute to fuel economy [114].

In addition to the above developments changes in materials in mass-market vehicles have reduced body weight and thereby fuel consumption. Since 1980, for example, the use of non-ferrous metals such as aluminium has increased and the use of thermoplastics and thermoset plastics has risen to approximately 10% of overall vehicle weight. Fundamental vehicle structure and manufacturing processes remain unchanged in the majority of cars despite the utilisation of lighter-weight materials although there is evidence of more radical approaches in some alternatively powered vehicles.

7.2 'Alternative' and emerging fuels and technologies

7.2.1 Clean Fuels

The British government describes liquefied petroleum gas (LPG), compressed natural gas (CNG), methanol, and ethanol as 'clean fuels' because CO_2 emissions are marginally lower than those

produced through petrol combustion. LPG is a by-product from petrol refining and, like CNG, is plentiful and with minor modification, these fuels can be used in petrol engines. Like diesel, they have been widely used in mainland Europe for many years, but lack of a fuel supply infrastructure prohibited use in the UK until recently. The 'clean fuel' and PowerShift programmes were launched in 2000 to encourage use of LPG and CNG in particular and supply has risen to more than 1,000 outlets although the majority of the 40,000 users are fleet vehicles. Nevertheless, as more LPG vehicles are produced, thus eliminating conversion costs, lower tax may yet encourage greater use by private owners [108].

7.2.2 Biodiesel

Biodiesel can be used in standard diesel engines and is processed from oilseed rape or recycled from used vegetable oil. Although biodiesel produces CO_2 and other emissions when combusted, it is also described as a 'clean fuel' because it is argued that the crops grown for fuel production act as carbon sinks and absorb CO_2. Again this fuel is comparatively readily available in mainland Europe but will be available in the UK from August 2002 where use could increase as a result of a 20p per litre reduction in fuel tax [115, 116].

7.2.3 Electric vehicles

Invented at the same time as the internal combustion engine, electric vehicles were initially as popular as their internal combustion counterparts and particularly among women drivers because they did not require cranking to start. They were comparable in speed but developments in internal combustion engine technologies led to travel at higher speed and that in conjunction with the development of the electric starting motor in 1912 and the limited range of electric vehicles contributed to the dominance of the internal combustion engine vehicle [117]. Nevertheless, developments in battery technologies (namely nickel metal hydride (NMH), nickel cadmium (Ni-Cd), lithium ion (Li-ion) and lithium polymer batteries) and more efficient motors and motor controllers have all increased the electric vehicle range [118, 119]. Some electric vehicles, known as conversions, are standard steel-bodied internal combustion engine vehicles with replacement electric powertrains (e.g. the Peugeot 106E (Fig. 32)) but others (including the Ford Th!nk, GM EV1 (Fig. 33) and Nissan Hypermini (Fig. 34)) embody a more radical approach and are described as 'ground-up' vehicles. These incorporate plastic body panels so that lower body weight therefore increases vehicle range.

Although limited, use of electric cars in California and particular European towns is considerably higher than in the UK where purchase is restricted to commercial operators. Average daily mileage in the UK is less than 16 km and so with an average range of 80 km urban and use of electric vehicles as second cars is perfectly feasible and if manufacturers open the private market in the UK electric vehicle ownership could also increase.

7.2.4 Internal combustion–electric hybrid vehicles

Introduced to the UK in 2000, the Honda Insight (Fig. 35) and Toyota Prius (Fig. 36) are both Internal combustion–electric hybrid 'dual-fuel' vehicles where an electric motor and batteries are combined with an internal combustion engine in order to reduce fuel consumption without limiting range. The two standard hybrid systems, series and parallel, operate slightly differently in that a series hybrid system utilises the heat engine to power a generator whereas in a parallel hybrid, power from the heat engine and generator is delivered directly to the drive train. Unlike a series hybrid, the electric drive in a parallel hybrid vehicle can therefore function autonomously or in conjunction with the heat engine [120].

Figure 32: Peugeot 106E.

Figure 33: GM EV1.

Figure 34: Nissan Hypermini.

Figure 35: Honda Insight.

Figure 36: Toyota Prius.

The two-seat Honda Insight and the four-seat Toyota Prius both switch between series and parallel hybrid mode. With lightweight plastic body panels, an overall weight of 773 kg (1700 lbs) and low drag coefficient of 0.25, the Insight combined (urban and motorway) fuel consumption is 83.1 mpg (3.4 l/100 km) [121]. The metal-bodied first generation Prius is more traditionally styled and constructed with a body weight of 1200 kg (2640 lbs) and a drag coefficient of 0.29. Consequently average combined fuel consumption is 57.6 mpg (4.9 l/100 km) [122] but this is considerably better than average UK fuel consumption at 30 mpg.

Like many pure electric vehicles, the Prius exploits regenerative braking to further charge the batteries. Again like electric vehicles, when stationary, the internal combustion engine does not run in either vehicle, thus reducing emissions and saving fuel although in certain circumstances, the hybrid's electric motor may continue to run. The Insight and the Prius are thus defined as super ultra-low emission vehicles under California Air Resources Board rules, with CO_2 emissions at 80 gm/km and 114 gm/km respectively. These levels are well below the permitted maximum of 140 gm/km for all new cars that will be introduced as an EU mandate in 2008 [109].

Some experts claim that, because hybrid technology does not limit range, it is the optimum means of reducing emissions until or unless battery and other electric vehicle technologies can be seen to compete with that of internal combustion engine vehicles. However, the dual system makes these vehicles more expensive than internal combustion engine equivalents and design costs alone are estimated to be between $100 and $300 million [123].

Both Honda and Toyota see these particular models as retail prototypes and both were 'loss-leaders' in that they are sold for less than the cost of manufacture [124, 125] but investment in the Prius has paid off and Toyota have now 'broken even'. Ford recently dropped the planned introduction of a hybrid version of the Explorer SUV in 2005, because, at 27 mpg, fuel economy was not as good as expected [123]. They still plan to introduce a smaller hybrid SUV, the Escape, with an urban fuel consumption of 40 mpg in 2003 [123, 126]. At a time when automotive industry profits have decreased, making manufacturers even more cautious than when the market is buoyant, it is possible that the high design cost of hybrids could deter some manufacturers from investing in hybrid vehicles and that only those with existing plant will continue production thereof.

7.2.5 Hydrogen fuel cell vehicles

First invented in 1839, fuel cells were developed by NASA for the Gemini and Apollo space programmes in the 1960s. Using gaseous or liquid hydrogen, methanol, ethanol or natural gas, electricity is produced by combining hydrogen and oxygen to produce water and is thus the reverse of hydrolysis. Fuel cells are comparable to batteries in that they store electricity, but, as opposed to being discarded like primary batteries, they are slowly and continually recharged while in use like secondary batteries [119]. The majority of current fuel cell applications are stationary, but the proton exchange membrane fuel cell (manufactured mainly by the Canadian company Ballard) has already been demonstrated both in buses and in several concept cars including the DaimlerChrysler Necar 5 (new electric car), the Jeep Commander 2 and the Honda FCX-V3.

Storage of liquid or gaseous hydrogen can be problematic because it is somewhat unstable when compressed, whether subjected to heat or cold. Research relating to the storage of hydrogen as NMH is ongoing but use of methanol as a hydrogen carrier appears to be the most prevalent solution for vehicle propulsion at present and has been adopted by Honda and DaimlerChrysler in all the above models.

The large physical volume of early fuel cells prohibited their use in anything other than concept cars because they occupied a considerable amount of interior space. The size of fuel cells has now been reduced and is reputed to be comparable to a standard internal combustion drive system. Had this not happened, fuel cell dimensions would have influenced overall vehicle size and in addition to prohibiting their use in smaller urban-type cars, would have adversely influenced fuel consumption. Similarly the cost of fuel cells was, until recently, 10 times higher than that of high efficiency diesel engines [127]. Costs have decreased but it is more than likely that, like the Insight hybrid vehicle, the Necar and others will be 'loss-leaders' in market research programmes. A further and significant reduction in fuel cell cost will is imperative if they are to become feasible power sources for personal transportation. Nevertheless, although prototypes for pilot programmes, in 2003 Honda intends to introduce 300 FCX-V3 or similar cars to the US and Japanese markets and DaimlerChrysler will introduce city buses in 2002 followed by a version of the Necar in 2004 [121].

Fuel cell systems are approximately twice as efficient as internal combustion engines, but although fuel cell vehicles emit only water, production of hydrogen from methanol or natural gas produces emissions although these 'up stream' emissions are lower than those produced by use of fossil fuelled internal combustion engine vehicles. However, as with pure electric vehicles, these vehicles could be truly zero emission if electricity generated from renewable sources is used for hydrogen production [128]. Because problems relating to the use of gaseous and liquid hydrogen are as yet unresolved, the use of sustainably produced hydrogen is unlikely in the immediate future.

As yet there is no fuel supply infrastructure, although DaimlerChrysler claim that oil companies are backing the development of a methanol supply infrastructure [121] which could include the decentralisation and installation of an on-site hydrogen production plant [128]. It is reasonable to suppose that not only will oil companies want to supply fuels in addition to oil derivatives, but that they will also support the most economic means of fuel supply via the existing filling station network in order to minimise investment costs.

7.3 EU policy on ELVs

Until recently there was no definite policy pertaining to vehicle disposal but recent EU legislation under the Priority Waste Streams Programme demands that this waste must be managed correctly. Thus, by 1 January 2007 a minimum of 85% of the average weight of a vehicle must be reused or recovered, rising to a target of 95% by 2015 including the recovery of energy through waste incineration. In addition to the banning of hazardous substances in cars from 2003, from July 2002, the EU directive will force manufacturers to pay for the disposal of cars manufactured after April 2001 and the subsequent disposal of all cars from 2007 [102].

7.3.1 The implications of EU policy

As could be expected this well-intended initiative is already provoking comment from motor manufacturers because these responsibilities will add approximately £250 to the price of a new car. In a recent consultation paper the Department of Trade and Industry stated that the total cost to business could be between £161m and £346m a year between 2006 and 2015 rising to £209m to £438m thereafter. The Society of Motor Manufacturers and Traders (SMMT) states that the cost of recycling should not just be the responsibility of manufacturers but also that of owners, insurance companies and financiers lending money for car purchases. SMMT also believes that the Treasury should contribute some of the value added tax gathered from new car sales towards recycling but as yet these financial issues remain unresolved [129]. Recycling facilities are limited at present and require investment for plant construction but whether they will be constructed in time to meet the legislation in the UK is unknown and the record for recycling to date is poor.

Recycling in parts of mainland Europe is reputedly better than in the UK at present where 80% of 28 million tonnes of domestic waste is put into landfill sites although EU law will force a cut to 33% by 2016. Members of the public have, however, already expressed concern about the impact of waste recycling and incineration plants near their communities [130, 131] and more protests may follow. In 2001, European legislation on waste electrical and electronic equipment was implemented. This requires the substitution of various heavy metals and brominated flame retardants in new electrical and electronic equipment by 2008 and until then products must be disposed of appropriately [132]. It is now illegal to deposit fridges in landfill but there are not yet adequate recycling facilities in the UK and although 900,000 fridges have been collected, they have been stockpiled until they can either be sent to Germany and the Netherlands for recycling or a new plant is opened in the UK [133]. Various bodies consequently fear that similar delays will affect automotive recycling.

7.3.2 Current automotive recycling

It has been suggested that although automotive manufacturers will be responsible for ELVs, the EU directive is not absolute and so certain components could be stockpiled on site. A review of current recycling issues and facilities indicates that there is no complete recycling infrastructure as yet and that changes in materials selection are probable.

7.3.2.1 Metals Recycling and disposal of metals varies according to type with 95 to 98% being recycled; these metals comprise of ferrous and non-ferrous metals including aluminium, copper and magnesium. Platinum, rhodium and palladium will also be recycled in greater quantities as more catalytic converters reach end of life and more ceramic casings will be recovered and powdered for refining [134–136].

7.3.2.2 Plastics Use of automotive plastics has risen to between 10 and 15% of total car components during the past 20 years, the majority of which are polyethylene, polypropylene (which accounts for approximately 40% of thermoplastics used), while polyurethane and polyvinylchloride (PVC) account for 12% of thermoplastics used. Variations in filler content and other additives including colorants mean that most cars comprise of up to 25 different types of both thermoplastics and thermoset plastics such as composites. PVC recycling is particularly problematic partly because it emits dioxins when incinerated and includes phthalate plasticisers, which are thought to be endocrine disrupters, and so there are currently no large-scale recycling schemes in operation. However, in addition to a proposed EU directive relating to the sustainable disposal of PVC [107], car manufacturers are currently investigating alternatives to PVC [137].

At present recycling of automotive thermoplastics is limited due in part to the way in which various types have been combined and because separation into type is deemed not to be cost effective. In some newer models, however, plastics recycling is now a design consideration although unlike metals (which can be reused numerous times) reuse of plastics (like aluminium) is limited because they deteriorate and lose their inherent qualities when recycled. Nevertheless, they can be used in non-structural elements as illustrated in Fiat's Cascade Recycling process: first use as a structural component like a bumper (Fig. 37), second use as a non-structural component like ventilation ducting (Fig. 38) and third use as floor covering (Fig. 39) [138]. EU legislation means that other companies are adopting similar strategies and the quantity of thermoplastics recycling should therefore increase. Composite and thermoset materials are more difficult to recycle because like eggs, when heated their chemical state changes. There is little composites recycling at present although they can be ground and used as filler but research into composites recycling is on-going.

7.3.2.3 Fluids The majority of fluids for disposal are lubricating oils accounting for 480,000 tonnes annually. Much waste oil collected for recovery in the UK is processed (by removing excess water and filtering out particulates) and burnt as fuel in heavy industry and power stations. However, tighter emission limits and fuel quality controls resulting from environmental legislation could mean a reduction in the amount of waste oil thus used. The preferred option for lubricating oils is re-refining for reuse as a base lubricant, although this does not currently occur on a large scale in the UK.

7.3.2.4 Batteries The recovery of lead-acid batteries in the UK is well established, the majority being collected by garages and local authorities. In 1998/1999 over 90% of the 144,000 tonnes consigned in England and Wales were recycled. However, many batteries are neither recovered nor recycled and may even be shredded within ELVs.

7.3.2.5 Restraint systems In 1993 secondary restraint systems such as airbags became standard components in all new UK vehicles and so some vehicles are arriving at car breakers' yards with undetonated airbags. Because they do not contain high value materials, reclamation is not a cost-effective option. However, vehicle manufacturers are being encouraged to make the dismantling thereof safer and more efficient [103].

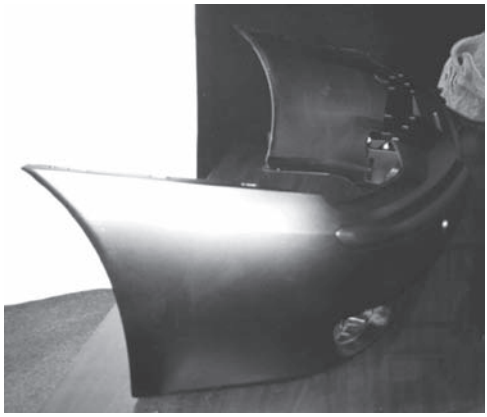


Figure 37: Bumper (fender).

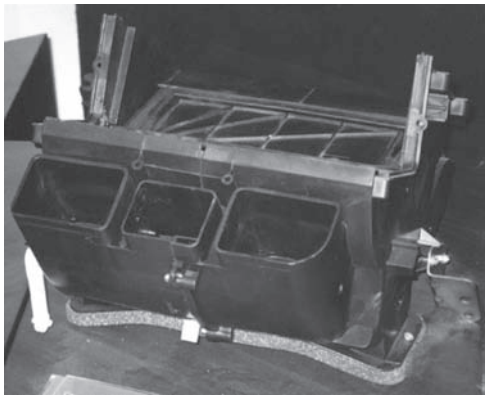


Figure 38: Ventilation ducting.

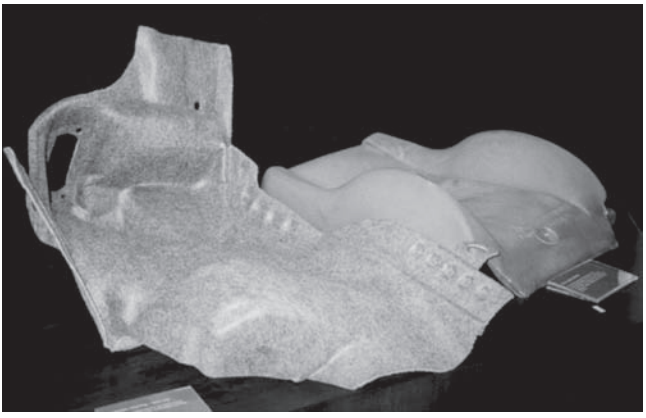


Figure 39: Floor covering.

7.3.2.6 Tyres Approximately 100,000 worn tyres are removed from vehicles every day in the UK, accounting for a total of 39.5 million tyres (467,650 tonnes) in 1998, 70% of which were recovered and reused. A significant proportion of the remaining 30% was buried as landfill where they will remain for decades because they do not degrade. However, from 2003, under the EU Landfill Directive, tyres will be progressively banned from landfill sites and so the Used Tyre Working Group is aiming for 100% recovery. Recycling and reuse include the production of retread tyres, granulation for use in children's playgrounds, running track surfaces and as carpet underlay and landfill engineering projects such as motorway embankments and a variety of marine applications. In 1998 energy recovery accounted for 18% of reuse and used tyres were incinerated to produce electricity for use by industry and local communities while two cement kilns now use tyres for fuel and other kiln operators are investigating their potential use [139, 140].

7.4 Summary

It is evident that the automobile design concept is static and that change in mass-market, internal combustion engine cars has been and continues to be incremental. At present the most radical design developments are found in concept vehicles such as those produced through the PNGV. The most innovative commercially available vehicles (e.g. the Ford Th!nk and the Honda Insight) use alternative power technologies and lightweight body materials to increase efficiency. They are 'ground-up' designs and therefore more akin to 'total design' solutions.

As stated in Sections 7.1 and 7.3.2.2 the composition of materials used for vehicle manufacture is already changing in order to reduce fuel consumption and the way in which cars are designed and constructed has developed to facilitate ease of recycling. However, the EU directive acknowledges that the guidelines on energy recovery could inhibit any increase in use of both thermo and thermo-set plastics [102]. Similarly thermo-set plastics and other lightweight materials including composites cannot be as readily recycled as thermoplastics which will in turn influence the design of lighter-weight vehicles.

Recycling processes also use energy, the amount of which will vary according to initial vehicle construction methods and material type. These factors must all be considered as part of the product lifecycle and design process, as must the production and disposal of toxic substances during both manufacture and dismantling. In short the increasing demands of environmental considerations mean that these form a significant constraint on overall design, and we now discuss the potential for change in automotive design and manufacture.

8 Potential changes in the automotive industry

The above information all indicates that current methods of designing and manufacturing mass-produced cars are contrary to the premise that all natural life forms minimise energy and material needs as a matter of survival. It is apparent that ownership and use of the private car is going to increase but rather than asking *what if* the car survives in its present form, we should ask *can* the car survive in its present form to which the answer should be *no*. It is also apparent that the need for change is urgent and that if the automotive paradigm cannot be changed immediately, it must change very soon.

We have seen several concept vehicles, the majority of which are more fuel efficient than current mass-market models but unless there is a significant change in manufacturing processes and paradigm, they will remain only as concepts. However, these are not the only examples of

radical thinking and we now discuss other potential solutions from both inside and outside the automotive industry.

8.1 The 'customised' car

In her book *No Logo* Naomi Klein discusses the many cultural and counter-cultural issues related to globalisation and associated product branding. There have already been anti-globalisation protests at G7 summits and Klein predicts that, based on current 'underground' reaction to multi-nationalism, corporate products and homogeneity will be rejected if not for political reasons then because consumers seek greater individuality [141]. Ironically some companies have already begun to exploit the potential for what has become known as *mass customisation*. For example, some Levi jeans' shops offer a 'tailoring' service with computerised measuring booths while an Internet-based luggage company allow customers to 'design' their own products. For the latter materials, colours, number of pockets and other details are selected from a range of options to produce customised bags. Similarly mobile phones can be personalised in various ways with clip-on covers, screen logos and ring tones either downloaded from the Internet or 'composed' on the phone itself.

As stated in Section 5.3 'Buddism' encouraged and is best suited to the manufacture of (ideally) millions of identical vehicles but there is evidence of car customisation both inside and outside the industry. We have already noted that Ford preferred black paint because it dried more quickly than lighter colours. Other pigment-related problems such as running or dripping have been eliminated so choice of colour has increased significantly. Industry's response to the increasing similarity between mass-market models (Figs 40–42) is the launch of 'limited editions': these models are cosmetically enhanced with graphics and/or special paint colours again to meet market demand for individuality. Alternatively owners can purchase add-on accessories (ranging from steering wheel covers to spoilers and wheel arches) at high-street retailers but the most extreme examples represent a sub-culture in which mass-market cars (known as 'custom cars') are modified by enthusiasts to the extent that original models are no longer recognisable. Such modifications involve significant manual input and are thus very costly and, because 'Buddist' car production does not lead to cars that are easy or cheap to modify, often exceeds the value of the car.

Markets, and in particular those in the industrialised countries, demand increasing differentiation and increasingly visual differentiation. The incumbent Buddist production system centred round steel unitary construction cars cannot easily accommodate this. Consequently as the demand for differentiation increases and the resulting market fragmentation leads to lower per-model volumes, breakeven points can no longer be achieved for a growing number of models and the Buddist paradigm will eventually be stretched beyond breaking point. The much-hailed *mass customisation* will therefore be accommodated within the narrow confines of 'Buddism', unless there is a paradigm change in mass car production.

The *mass-customisation* of vehicles can be extended beyond the application of cosmetic detail and over time will prove beneficial to consumers, the automotive industry and the environment. Increased design flexibility and 'tailored' vehicles will not only satisfy market demand for greater individuality but will also reduce materials and possibly energy consumption as vehicles are modified to suit owners' changing needs and desire for new models. This is of course dependent on adoption of new manufacturing paradigms which, by lowering the 'break-even' point, means that manufacturers are not restricted to the production of such large numbers of identical vehicles, thus permitting greater freedom of choice. One of the more radical concepts developed inside the automotive industry by GM is an example of *modular platform engineering* called AUTOnomy.

Figure 40: Nissan Micra.

Figure 41: Hyundai Amica.

Figure 42: Toyota Yaris.

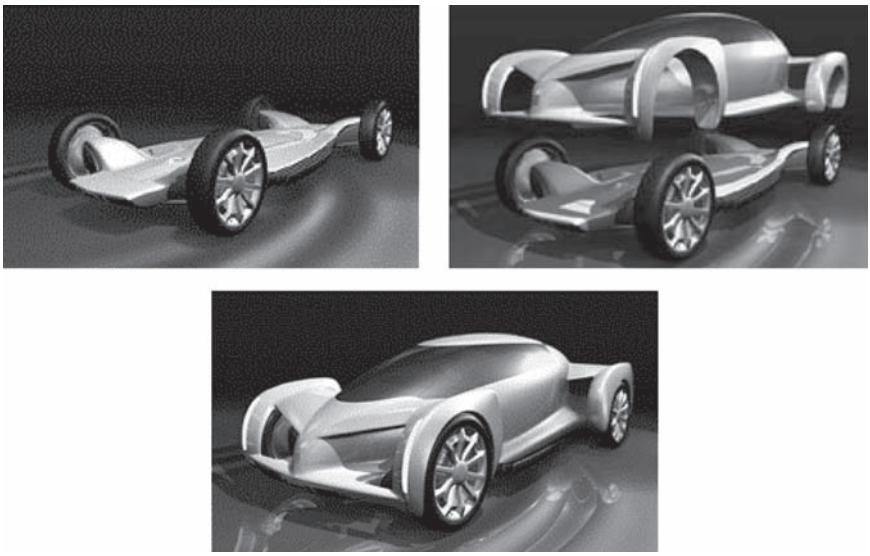


Figure 43: GM AUTOnomy. Published with permission from General Motors.

8.2 AUTOnomy – reinventing the chassis

If we consider the existing car making paradigm, which we have called 'Buddism', AUTOnomy can be regarded as an extreme example of a 'platform' approach. In a platform strategy, the key elements of the monocoque and a limited set of powertrain and chassis components are shared among a number of models to reduce costs. This technique tries to combine a reduction in mechanical diversity with an increase in perceived diversity.

AUTOnomy (Fig. 43) is a concept vehicle and was exhibited by General Motors at the 2002 Detroit International Auto Show. Few commentators outside GM appear to appreciate the full significance of this concept; rather they focus on the fuel cell powertrain, which is of course novel. This vehicle consists of a flat base unit about 15 cm thick, which contains the entire fuel-cell powertrain and holds the wheels and suspension and, known by the GM engineers as the 'skateboard', is essentially an autonomous mobile structure. Unlike conventional internal combustion systems, fuel cell technology and the electric powertrain allow for very flexible packaging while use of various lightweight alternative materials in its body and its chassis mean that it differs significantly from the Buddist solution. This 'skateboard' has one central 'docking point', which serves as a communications interface between the body and the powertrain module.

AUTOnomy is far removed from the steel monocoque and is based around a modular system. In some respects it reintroduces the separate chassis and body structures used by Henry Ford on the model T but fully exploits the separate body and chassis. Providing that the docking points are standardised, AUTOnomy allows for a wide range of bodies to interface with the standardised 'skateboard' powertrain unit. The current concept model is fitted with a two-seater sports car body although the standardised mass-produced 'skateboards' will accommodate saloon, estate, MPV, coupé, roadster and more radical body styles either made by independent small specialists or by mainstream manufacturers.

On the one hand this spells a potential return to the luxury cars of the past, which were supplied as a chassis onto which a body from a coachbuilder of choice was fitted. Within the modern idiom

it can on the other hand be compared with exchangeable covers for mobile phones. Storage space would probably prevent most customers from owning large numbers of bodies; nevertheless, many people might consider owning one or two spare bodies in addition to their daily driver. For daily commuting (if that still exists when such vehicles become mainstream) a sports car body could be used while an estate or MPV option could be fitted at the weekend. Alternatively, dealers or independent body rental firms could provide this service creating a new sub-sector. Bodies and 'skateboards' could be updated independently as technology needs, fashions and options change. More efficient powertrain items could be fitted to the 'skateboard' without affecting the bodies, provided the interface could accommodate such an upgrade. Much of such upgrading, including tuning to the needs of individual customer, could be carried out through software reconfiguration.

GM's AUTOnomy is the first purpose designed fuel cell vehicle. It also introduces drive by wire (or 'driver control unit') where all controls actuated by the driver in the body module communicate with the power module via the electric/electronic docking points in a dedicated control unit. The driver control unit works through a new generation 42 V electrical system supplied by Swedish firm SKF, and means that conventional pedals, steering wheel and steering column are no longer required. GM also argues that low tooling costs and the flexibility of dedicated body types for differing market segments could help to bring automobility to developing countries and hence to the 88% of the world's population who currently do not enjoy its benefits.

AUTOnomy is a significant attempt to start from first principles. Rick Wagoner (Chief Executive Officer at GM) claims that GM started by asking the question:

> What if we were inventing the automobile today rather than a century ago? What might we do differently?

He was indeed correct when he went on to say that

> AUTOnomy is more than just a new concept car; it's potentially the start of a revolution in how automobiles are designed, built and used.

All fixed points in the design are located within the 'skateboard' and, assuming that there are connections to the chassis, the body can be configured in any way. The traditional limitations of engine bay, bulkhead, steering column, pedals, suspension no longer exist thus permitting the body designer almost unlimited freedom. Being centred around people and their needs rather than the needs of the production system that makes the car (as is the case under the existing Buddist–Fordist system), AUTOnomy represents a return to true automotive design while the basic 'skateboard' structure would last . . . *'much longer than a conventional vehicle'* [142].

8.3 The hypercar and 'whole systems thinking'

Developed at the Rocky Mountain Institute (RMI) in Colorado, the 'Hypercar' (Fig. 44) is one of the most radical concepts to evolve outside the automotive industry. Amory and L. Hunter Lovins established the RMI in 1982 as

> an entrepreneurial, non-profit organisation that fosters the efficient and restorative use of resources to create a more secure, prosperous, and life-sustaining world [143].

In addition to education programmes, the research and consulting team works with corporations, governments, communities and citizens to help them to solve problems, gain competitive advantage, increase profits, and create wealth through a more productive use of resources. Areas of expertise include the impact of businesses, communities, energy, buildings and land,

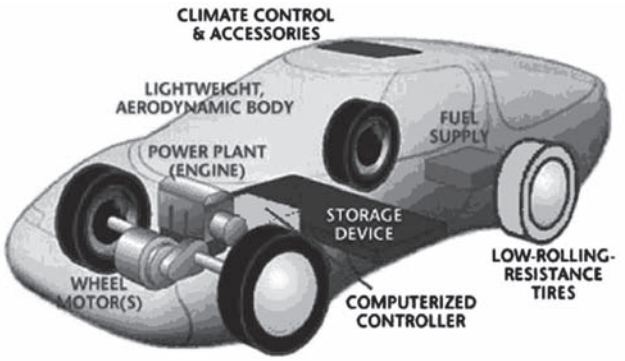


Figure 44: RMI Hypercar. Published with permission from Hypercar Inc.

climate, water and transportation and several RMI efficiency innovations have subsequently won awards.

The RMI credo is founded in systems thinking and the identification of interactions, while interdisciplinary knowledge of advanced technologies and techniques and end-use/least-cost analysis is used to develop integrative design solutions, one of which is the Hypercar.

Initially described as a 'supercar' Amory Lovins began to explore the synergistic benefits of combining ultra-light, ultra-aerodynamic construction and hybrid-electric drive in 1991. The team refined the concept, which was launched publicly in 1993 at the International Symposium on Automotive Technology and Automation and for which RMI received the Nissan Prize. Established in 1994, the Hypercar Center is dedicated to support the rapid commercialisation of ultra-light hybrid vehicles. Between 1994 and 1996 the Center concentrated on computer modelling and the publication of technical papers to prove that the Hypercar could meet all market requirements and that the innumerable technical challenges could be overcome. During the same period, the Center advised carmakers about the Hypercar concept and by mid-1997 it was engaged in various forms of discussion and collaboration with approximately 30 current or intending auto-manufacturers. Increasing interest from manufacturers then prompted the Center to launch Hypercar Inc. in 1999 in order to further develop the Hypercar concept through the application of innovative design and development practices, new automotive business concepts and the establishment of collaborative business and technology partnerships.

In the past many technical innovations which had been developed outside the automotive industry were purchased and purposefully shelved by individual motor manufacturers. The Hypercar Center adopted a radical strategy to avoid this situation and placed the majority of its intellectual property (published as *Hypercars: Materials, Manufacturing, and Policy Implications*) in the public domain. While raising capital for the Center this strategy also encouraged competition among carmakers and their utilisation of elements of Hypercar technology.

The Hypercar remains an ultra-light, aerodynamically efficient, hybrid-electric vehicle and further, hydrogen fuel cell technology ensures that the only emission is water. Use of a frameless monocoque body reduces the number of components required (currently about 15,000 in a standard internal combustion engine vehicle) while use of lightweight materials including carbon fibre and aluminium means that the vehicle is two to three times lighter than an equivalent steel-bodied model. Hydrogen fuel cells are also potentially easier to fabricate than internal combustion engines (which have about 1,000 parts). The vehicle uses 92% less iron and steel, 33% less aluminium, 60% less rubber and 80% less platinum than a traditional vehicle and production tooling costs are reduced by 90%. Similarly painting (which currently accounts for about 25% of production

costs and is one of the main pollutants produced during manufacture) will be eliminated as colour is embedded during moulding. The four principal manufacturing parameters (namely time from concept to street, investment in production, time and space needed for production, and the number of parts in the autobody) are subsequently reduced by a factor of 10. Other upstream pollutants and emissions are also minimised because the Hypercar requires no more than 10% of the amount of consumable fluids (antifreeze, brake and transmission fluids and oil) used by a standard vehicle.

As stated above the RMI applies systems thinking and interactions to develop design solutions. The Hypercar concept is, as an example of 'whole systems thinking', greater than the sum of its parts. We have discussed the history and evolution of the contemporary motor car and learned that change has been and is incremental. Lovins and his colleagues, however, returned to first principles and so the Hypercar is more than just a fuel-efficient transport solution. Like James Grier Miller's Living Systems Theory diagram (Fig. 2) that describes a hierarchy of systems beginning with *cells* and culminating with *supranational systems*, the Hypercar has been developed as a component within larger systems.

The RMI argues that the current automotive manufacturing paradigm is 'disintegrated' because the causes of and connections between problems are not defined. The increasing complexity of cars has produced increasingly specialist designers and engineers whose job is to improve and optimise a given component or subsystem. Consequently the modern mass-market car has evolved incrementally through small improvements in individual components and little change in the overall concept. The optimisation of isolated components frequently undermines the whole system because of a lack of integration and synergy, which, in the case of the car, has resulted in over-complex, oversize and inefficient designs.

Conversely the 'whole system' approach analyses the entire system (rather than individual components) and every element is considered simultaneously to reveal mutually beneficial interactions. Individually none of the defining features of the Hypercar offer enormous economic or technical benefits but when integrated, the resultant vehicle is at least three times more efficient, is less polluting, cheaper to run, and may even be cheaper to manufacture than traditional models.

The Hypercar has also been designed as a component within a larger system: prompted by the high cost of fuel cells (which is one of the factors that currently prohibits their use in cars), Lovins *et al.* propose means of both minimising costs and limiting pollution. They argue that cost will be reduced as production increases and advocate use of fuel cells in buildings to increase market share. The energy from the hydrogen will be maximised with 60% used for electricity generation; the remainder is then used to heat water to 77°C and to run heating, cooling and dehumidification systems. Excess electricity will be sold to the grid and the profit will pay for the methanol reformer, which is used to extract hydrogen for water. In-vehicle fuel cells could be purchased but leasing would be preferable and spread the cost. In order to cover leasing or purchase costs, when parked (which, for a typical car, is approximately 23 hours a day) a fuel cell vehicle would also be used as a generator. The car would be plugged into the electricity grid and the sustainably generated electricity would then be sold and both help to cover vehicle costs while reducing the need for supply from coal-fired and other polluting generators.

The final example of paradigm shift relates to scale of manufacturing plant, which is important because the US economy is, like so many others, dependent on the automotive and allied industries. Rather than decreasing the workforce, the RMI claim that Hypercar manufacture will employ as many if not more people although they will be employed in smaller, local and decentralised plants. As previously stated fuel cell costs are high as are the cost of composite body materials. In 1995 the cost of carbon fibre was 20 times higher per kilo than steel but this dropped to 12 times by 2000 with the prospect of further reductions. Some of these initial costs will be countered by cheaper tooling but vehicle purchase cost could also be limited by selling directly to the public. In the

Figure 45: 'Revolution' concept car designed by Hypercar Inc. using 'Hypercar' principles. Published with permission from Hypercar Inc.

USA vehicle production cost currently accounts for 50% of purchase price and so the Hypercar will be sold via the Internet in order to minimise overheads and thus purchase cost.

Hypercar Inc. claims that when produced, this will be a highly desirable vehicle. Congestion will thus inevitably exacerbate and the many social problems deriving from extensive car use, but the number of accidents should decrease. Crash testing has shown that innovative ultra-light vehicles are at least as safe as the standard cars due in part to structural design and dissipation of crash energy because composite materials absorb about five times the energy of steel. Lighter vehicles also means that injury to pedestrians and cyclists should be lower although the inclusion of rear view cameras and other safety features is also expected to reduce the overall number of accidents. Nevertheless, Lovins *et al.* advocate that car use should be discouraged by considering land use before mobility, developing genuinely competitive alternatives to private transport and making drivers pay the true cost of parking and driving ([22], pp. 22–47; [144–148]). However, in a country and culture in which the car is central, this seems more difficult to achieve than the straightforward introduction of an alternative to the steel-bodied, internal combustion-engined behemoth.

To date the Hypercar remains a concept on paper although Hypercar Inc. designed another vehicle based on the whole systems approach. Although not yet manufactured, a full-size model of the Revolution has been developed as shown in Fig. 45.

8.4 Summary

We have reviewed two radical alternative design and manufacturing paradigms, each of which will be technically and economically viable within the foreseeable future when a fuel supply infrastructure has been established. At present the greatest barrier to change appears to be manufacturers' lack of will. The introduction of mass-customisation and design flexibility through platform-based concepts such as GM AUTOnomy are both economically and environmentally beneficial. While these approaches meet current consumer demand for more individual and personalised products, given the prevailing attitudes of today's teenagers (and therefore tomorrow's car buyers), the introduction of design flexibility is essential to meet their future demands. Although adequate water and food supplies are of a higher priority GM claims that AUTOnomy would

contribute to automobility in the 'developing' world. As an exemplar of 'whole systems thinking' the Hypercar concept adheres more closely to living systems than other automotive paradigms. While significantly different to existing design and manufacturing paradigms, the Hypercar Center acknowledges that their model bypasses social lifecycle costs [149–152] and that there is a need for further development of composite and other recycling technologies. Therefore although full life cycle assessment is considered, it is not yet integral to their design process. Given that levels of car ownership will increase until 2050 and beyond it is now appropriate to discover if, how and when 'total design' is likely to be adopted as the preferred automotive design method and to learn about current practice.

9 LCA and automotive manufacture

During the past decade in particular there has been an increase awareness about and research related to aspects of vehicle lifecycle, much of which deals with propulsion systems and fuels and is known as the *well-to-wheel* or *mobility chain*. Analysis of the mobility chain includes primary energy consumption, greenhouse gas emissions, exhaust (tailpipe) emissions and pollutants through fuel extraction/energy generation and delivery of supply (upstream impacts), fuel use and disposal of hardware if appropriate. Both indirect and direct costs resulting from these activities are also considered. Other research relates to specific components and the *product chain* but as yet there is little life cycle assessment work relating to the integration and interrelationships between these chains in the form described in Fig. 46. We now investigate some examples of current practice beginning with a discussion about what should be included in a complete vehicle life cycle assessment.

9.1 Early eco-rating models

Although there is general agreement that cars are harmful to our environment, making a precise assessment of the environmental impact of a particular car is not easy. An ideal model developed

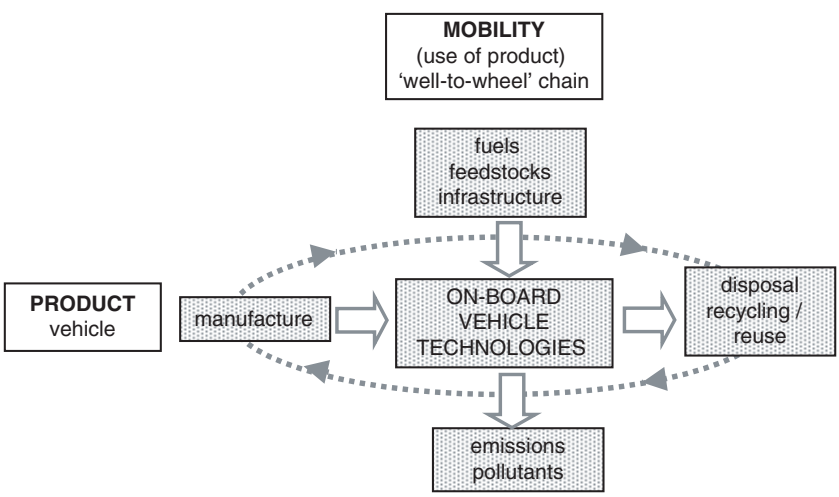


Figure 46: Vehicle LCA. Andrews, D. and Simpson, A.G. (adapted from Ogden, J., a diagram presented at the Workshop on Well-to-Wheels Comparisons for Alternative Fuelled Vehicles Nice France, May 14, 2001).

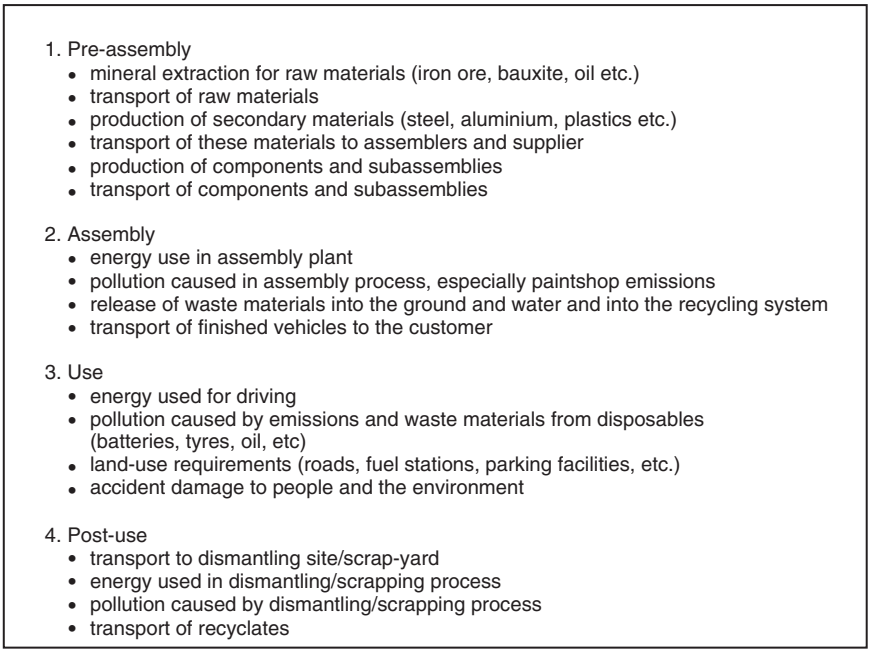


Figure 47: Car life cycle environmental impact. (Nieuwenhuis, P., Centre for Automotive Industry Research, Cardiff University).

by Nieuwenhuis and Wells in 1997 (Fig. 47) considers all materials, energy used and production of pollutants ([48], p. 140). The complexity and diversity of cars means that use of this model and conventional eco-ratings (such as those used for white goods) present insurmountable difficulties.

Nevertheless, a number of environmental ratings systems for cars have been used, notably in Germany [153, 154], Sweden's rototest (as used by *What Car?* magazine) [155], but also in the UK [156, 157]. These are all primarily aimed at private consumers. However, they omit a number of key indicators, such as product durability. In addition, the European Commission is introducing its own eco-rating for cars. This is currently based on CO_2 emissions, although a more comprehensive system will be introduced in due course [158].

9.2 The Centre for Automotive Industry Research (CAIR) Environmental Segmentation System (ESS)

Nieuwenhuis and Wells [159] presented a simple rating system aimed at corporate fleet buyers, who are facing the increasing pressure relating to their company's environmental performance and image. In the UK market, fleet or corporate sales represent between 50% and 70% of all new car sales, depending on the definition used. In 1997, for example, 1,018,419 cars were registered to businesses. This system used a number of proxies to represent a simplified lifecycle approach by using readily available data. It has since been used by a number of companies to 'green' their fleets. A number of approaches were tried, but for the sake of simplicity and availability of data it was decided to use a simple proxy formula representing the impact made by a vehicle on the environment. This formula is proving surprisingly robust and has been welcomed by

Table 1: The CAIR Environmental Segmentation System: examples. (Nieuwenhuis, P., Centre for Automotive Industry Research, Cardiff University).

Model	L (m) $\times$ $w(m)$ $\times$ wt (tonnes)	ESS
MCC Smart City Coupe	$2.50 \times 1.51 \times 0.72$	2.72
Old Mini	$3.05 \times 1.41 \times 0.685$	2.94
Lotus Elise	$3.73 \times 1.70 \times 0.67$	4.25
Mercedes A140	$3.575 \times 1.71 \times 1.061$	6.48
BMW 3-series	$4.435 \times 1.70 \times 1.235$	9.31
Audi A4	$4.48 \times 1.735 \times 1.225$	9.52
GM EV1 by Saturn	$4.31 \times 1.765 \times 1.301$	9.89
Volvo V70	$4.71 \times 1.76 \times 1.42$	11.75
BMW 5-series	$4.775 \times 1.80 \times 1.41$	12.12
Audi A8	$5.03 \times 1.88 \times 1.46$	13.8
Lamborghini Diablo	$4.46 \times 2.04 \times 1.57$	14.27
Lexus LS 400	$4.995 \times 1.83 \times 1.68$	15.35
Land Rover Discovery	$4.52 \times 1.81 \times 1.92$	15.70
Mercedes S-class	$5.11 \times 1.88 \times 1.89$	18.15
Lincoln Town Car	$5.56 \times 1.945 \times 1.83$	19.79
Rolls-Royce Silver Spirit	$5.295 \times 1.915 \times 2.43$	24.64

industry experts. It relates to a vehicle's size and weight as follows:

$$\text{Vehicle length (metres)} \times \text{width (metres)} \times \text{weight (tonnes)} = \text{ESS}, \tag{1}$$

where ESS represents Environmental Segmentation System. As a proxy for the environmental impact of a vehicle it is a fair measure as it relates the vehicle's weight to its 'footprint' and thus relates literally to its impact on the earth, and on other environmental factors such as use of resources. Table 1 gives some examples of the performance of a range of cars in terms of this first measure [159].

This ESS figure is then used as an input measure for a formula that incorporates the vehicle's performance against the regulatory and agreed requirements for various emissions as well as enabling any other measure to be incorporated. The system illustrates the benefits of weight reduction in larger cars (e.g. the Audi A8 competes with the Mercedes S-Class) but also tackles the price anomaly that would put the Lotus Elise and Land Rover Discovery in the same segment. In each case we use the lightest variant listed; i.e. with the lowest level of specification. It must be remembered that optional extras can significantly increase the weight of a car.

Although the issues are complex it is increasingly important to make available some sort of eco-rating for cars. Both private car users and fleet buyers are beginning to expect such information while EU and individual country legislation and taxation is increasingly reliant on environmental criteria; an example is the UK company car taxation system introduced in April 2002 and based on CO_2 emissions. It is, however, important to move away from the narrow toxic emissions-only view of the environmental impact of cars, as previously argued in *The Green Car Guide* and *Motor Vehicles in the Environment* [54, 160]. The urgent necessity is to consider how to make motorised mobility sustainable, and we will now discuss this.

9.3 Towards sustainable automobility

Sustainability is the most fundamental of environmental concepts and ultimately defines any practice in which we cannot indulge indefinitely without lasting environmental damage or impact as 'unsustainable'. We will investigate what this means for the automotive sector below. However, practices may also be sustainable for shorter periods. Thus in one sense, oil use is sustainable for the next 10 years, less so for the next 150 years and is unsustainable in a pure sense, in that we cannot continue to use oil indefinitely.

But what does sustainability mean in practice? An environmentally sustainable motor industry would not use finite resources and would not cause pollution that could not be easily absorbed by nature. At first this appears to be an impossible task, but it is actually technically possible to operate in this way. The first requirement would be a 'closed loop' economy as discussed in Section 2.4. Given the secondary materials currently available to world economies, with judicious recycling a car could be made without extracting additional raw materials. Also any energy used in this process would need to come from sustainable and renewable sources and not cause pollution that could not be readily absorbed. This would of course also apply to the transport of these secondary materials.

Use of renewable energy sources would be part of the answer but although this technology exists, it is not yet widespread enough to make an impact. It may never in fact meet our current requirements, so in addition to a reduction in the overall level of consumption, a 'closed loop' sustainable system would also imply a dramatic cut in energy use. This would involve a correspondingly dramatic reduction in the numbers of new cars produced. Again this is all technically possible; although the means by which this can be achieved together with examples of best practice as analysed by von Weizsäcker and Lovins [161] are not yet practised on the requisite scale. Despite the apparent fanciful nature of these concepts, they are becoming mainstream among environmentalists as well as some regulators and in the longer term will be unavoidable. Thus these future concepts must be considered when devising any longer term strategy at the present time.

For several years, the new more comprehensive environmental sustainability concept was largely confined to the environmental and academic communities but an award-winning paper in the *Harvard Business Review* by Stuart Hart [162] brought it to the attention of the wider business community. Hart asserts that sustainability should not be confused with mere pollution prevention or waste reduction, but requires a fundamentally different mind set. He writes that: '. . . *in meeting our needs, we are destroying the ability of future generations to meet theirs*'.

Hart foresees the development of completely new technologies together with completely new types of businesses in order to meet emerging sustainability needs. He predicts that in the developed economies the demand for virgin materials will decline as reuse and recycling become more common and that over the next decade or so sustainable development will become one of the biggest opportunities in the history of commerce. He postulates that businesses will have to decide whether they are part of the problem or part of the solution. Hart does not ignore the car sector and states that '*Although the auto industry has made progress, it falls far short of sustainability*'. He also extends the concept of producers' responsibility of producers further than ever before when he asserts that

> Companies can and must change the way customers think by creating preferences for products and services consistent with sustainability.

He concludes by saying that although changes in policy and consumer behaviour are essential, business can no longer hide behind these 'fig leaves' and must actively work to change consumer behaviour through education.

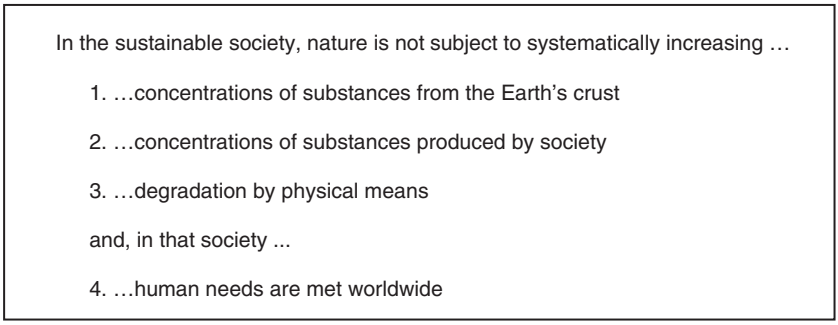


Figure 48: The four system conditions. (Nieuwenhuis, P., Centre for Automotive Industry Research, Cardiff University).

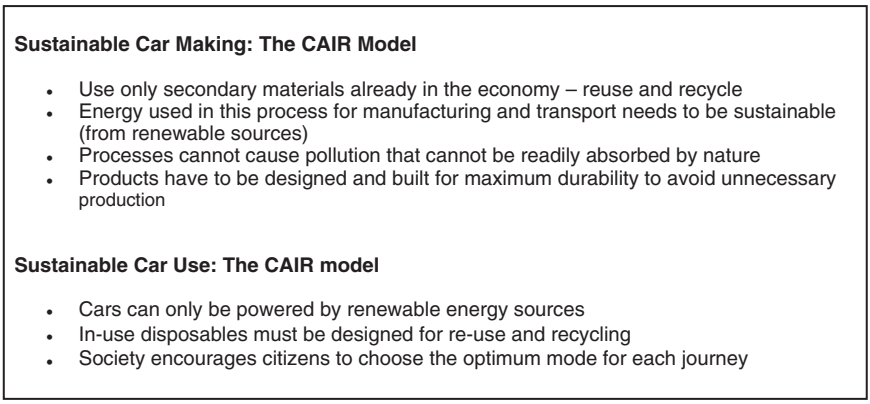


Figure 49: A model for sustainable automobility. (Nieuwenhuis, P., Centre for Automotive Industry Research, Cardiff University).

Inspired by The Natural Step (developers of which include Nattrass and Altomare) a model for sustainable car making and sustainable automobility was developed at the Centre for Automotive Industry Research. The Natural Step is a practical model for business sustainability originally developed in Sweden and used by a number of companies. It recognises a number of basic principles, or 'system conditions' as given in Fig. 48 [163].

In Nieuwenhuis [164] these general conditions were adapted for automotive use and are shown in Fig. 49:

Clearly these simple statements describe an approach that is very far removed from current practice. Nevertheless, having established this set of desirable outcomes and a vision of a sustainable automotive future, we must ask how this could be achieved. One key element is the so-called closed loop economy which we discussed briefly in Section 2.5.

9.4 Achieving a closed-loop economy

We have already stated that in a 'closed-loop economy', firstly, no new raw materials are added and only the existing pool of secondary materials is used, reused and recycled, and, secondly, any

energy used in reusing and recycling has to come from renewable non-fossil sources. In addition, no net increase in emissions is allowed. This means that emissions must be readily absorbed, for example, by growing crops.

In practice many developing countries are much closer to a closed-loop economy than most of the developed world which, it must be admitted, is far removed from it. Nonetheless, there are some areas where economic reasons have encouraged a move in this direction. As we know, paper and glass already use some recycled material. Similarly, steel bicycle frames, for example, tend to be made from recycled steel sourced from minimills, rather than the predominantly ore-derived steel used for car bodies [165]. The incidence of such examples is increasing and in many cases the appropriate technology is available. The main problem lies with economics considerations, which do not yet encourage reuse and recycling. The taxation of raw materials would be one way of moving in the right direction and this and other policy measures are being discussed by environmentalists and environmental economists around the world. The vehicle-recycling sector in particular will also have to play a key role in any serious moves towards a closed-loop economy. Although apparently utopian at present, in the longer term such a move seems inevitable because most current calculations show that many key raw materials will run out during the present century.

There are other serious considerations: although there are already enough secondary materials in the world economy to sustain a (limited) automotive industry on this basis, the ability even to experiment with new alternative materials will be restricted unless diverted from other uses. As previously stated use of renewable energy has been growing rapidly in recent years, particularly in countries like Denmark, but availability is still very limited. Although there is scope for a dramatic increase in this area, a concurrent dramatic cut in energy use through increased efficiency would greatly reduce overall requirements and thus make a closed-loop economy more achievable.

9.5 'Total design' and the automobile

This agenda has major implications for all types of design and that of the automobile in particular. In order to bring these technologies closer to natural sustainable processes, designers will need to change their mindset. Working in harmony with natural processes will become a basic requirement and will obviously affect both the *mobility chain* (choice of fuel and powertrain technologies) and the *product chain.* The latter comprises materials selection, finishes and vehicle form, (which will influence production processes and energy requirements for manufacture) and vehicle performance characteristics and efficiency. Simultaneous consideration of both chains is of paramount importance because of the influence each has on the other, while customer appeal and utility are vital for business success. It will only be possible for the automotive industry to progress and to begin to resolve the many problems deriving from current design and manufacturing paradigms when LCA becomes a key element within the design process and design methods.

10 Conclusion

This chapter began with a discussion about the key characteristics of natural systems, one of which is that they are self-regulating. While they may be described as 'open' (in that they respond to change in their surroundings and adapt and evolve in order to survive), they are also 'closed loops' so that when life ends, for example, residual material forms nutrients and therefore energy for elements within the same or other systems. We then learned that records of man's inspiration from the natural world date from the Ice Age and that the concept of *mimesis* and 'producing something in the fashion of nature' dates back to Classical Greece. Subsequent academic study of the science

of biomimetics and formal recognition of its value to the design and engineering professions is, however, comparatively recent and really developed during the 20th century. An examination of biomimesis and living systems and their relevance to the automotive industry followed. This revealed that, like living systems, in order to create a 'closed loop' economy, the use and reuse of automotive materials must involve intermediate processes and the entire economy rather than the individual sector.

We then explained the design process and reviewed several current examples of design methods (including that developed by Stuart Pugh known as 'total design') and found them to be incomplete and described as linear activities. Having explained life cycle assessment, we proposed that there is an urgent need to update these models and to integrate life cycle assessment within design methods to make design a holistic activity; only then will Pugh's model genuinely warrant the name 'total design'.

It was then appropriate to discover why the motor car evolved in the way that it did and to compare it with the development of the radio and the personal stereo (or Walkman). We discovered that both the radio and Walkman were 'new inventions' and are thus described as *dynamic* products. The designers and engineers who developed these products were unconstrained by historical precedents and were therefore able to develop products around user needs. In contrast the immediate antecedents of the motor car were horse-drawn vehicles and, in addition to exploiting bicycle technologies, the earliest models were designed and constructed using traditional coach-building processes and materials. These cars were thus known as 'horseless carriages'. The motor car gradually developed a specific product identity but was still constructed using 'craft' processes. Henry Ford is usually credited with democratising the car and certainly increased availability by adopting Cadillac's idea of using interchangeable parts and by introducing the assembly line production process. However, Ford's fundamental error was to design this process around craft techniques and to then design subsequent models around the production process. Many of the ensuing problems related to body construction and were resolved by Edward Budd and Joe Ledwinka when they developed a welded steel body process (that is here called 'Buddism') and replaced the separate chassis and body with a monocoque body. Since then change in the automotive industry has been incremental. Although contemporary cars are now more fuel efficient, produce lower emissions and look very different to those produced during the 1920s and 1930s, the automotive industry still adheres to a Fordist–Buddist paradigm and the motor car is therefore described as a *static* product. Moreover cars are still designed around an out-dated manufacturing process rather than around the users; consequently as drivers and passengers, we have no option other than to use vehicles that at best only partly meet our changing needs.

Nevertheless, the car has become one of the most popular and ubiquitous products in the entire world and there are currently more than 600 million on the road globally; by 2030 there will be 1 billion and the global market is not expected to stabilise until at least 2050. Car ownership and use have both contributed to and been encouraged by societal, cultural and economic change and the car fulfils a variety of psychological and practical roles. All evidence shows that, despite the numerous negative outcomes from extensive car use, the automobile or a similar form of personal transport is here to stay. However, this expected permanence is not without problems. Pollution and emissions are produced during car use and manufacture, both of which also consume finite resources; materials and energy while current disposal methods create waste, landfill and other pollutants. There is therefore an unquestionable and urgent need for change in both automotive design and manufacturing paradigms.

Research into 'alternative' fuels is ongoing as is continuing improvement in fossil-fuelled internal combustion engine technologies. However, the majority of commercially available 'alternatively powered' vehicles are adaptations of the industry standard while the most radically

designed vehicles remain as concepts. On first reading, the EU directive on ELVs may appear to be a positive change and will undoubtedly encourage the redesign of vehicles to facilitate recycling and reuse. However, unless we learn from living systems and ensure that the 'closed loop' includes the entire economy rather than just the individual sector, opportunities to develop new design and manufacturing paradigms could well be limited and even though components, materials and technologies are recycled, many current issues will still remain unresolved.

Both GM AUTOnomy and the RMI 'Hypercar' concepts represent an encouraging and dramatic paradigm shift and a return to first principles. In the Hypercar concept the vehicle is designed around the user rather than the production process. RMI adopted a synergistic approach to the problem describing the Hypercar as an example of whole-system thinking. As an example of platform engineering on the other hand, the GM AUTOnomy offers a level of flexibility unseen in the mainstream automotive industry until now. Although both concepts embody economic manufacture and zero-emission energy efficient propulsion systems, neither project includes LCA and so it is not yet possible to assess the overall level of sustainability of either vehicle. Nevertheless, they are currently the most advanced automotive concepts and could be indicative of the future of personal transportation solutions.

In the final section of this chapter we stated the need to fully integrate the *mobility* and *product chains* because of the impact each has on the other. We acknowledge that while it is difficult to develop and apply a comprehensive vehicle LCA system there is a need for a more comprehensive eco-rating system because many current examples omit a number of key indicators, such as product durability. The CAIR Environmental Segmentation System is acknowledged as one of the more extensive models although this was developed in response to industry's existing needs rather than future requirements and resultant changes will therefore be incremental. Based on The Natural Step model for business sustainability, the CAIR Sustainable Car Making and Car Use models are, however, far more visionary than others and embrace all aspects of vehicle design and operation. Although they appear to be somewhat fanciful at present, it is inevitable that they or similar models will be utilised in the long term because many natural resources will be depleted during this century.

We conclude therefore that like living systems, the automotive industry must be prepared to adapt and evolve in order to survive. Change will include a return to first principles and the development of new design and manufacturing paradigms, which will be determined in part by the integration of life cycle assessment within the automotive design process and methods. The eventual integration of LCA is inevitable but until this happens there will always be unanswered questions about the optimum vehicle design solution. However, rather than waiting until this practice becomes mandatory, it would be economically, socially and environmentally beneficial for a holistic design process (involving genuinely 'total design' methods) to be adopted by the automotive industry immediately.

References

[1] Pearsall, J. & Hanks, P., (eds). *New Oxford English Dictionary*, Oxford University Press: Oxford, 1998.

[2] Cook, R., *Animal Cognition & Learning*, Tufts University: Medford, MA, retrieved August 2002 from http://www.pigeon.psy.tufts.edu/psych26/birds.html

[3] Goodall, J., *The Chimpanzees of Gombe: Patterns of Behavior*, Belknap Press of Harvard University Press: Cambridge, MA, 1986.

[4] Feucht, D.L., Design in Nature and the Nature of Design, *Origins & Design* 19:2 37, Access Research Network, Origins & Design Archives, 1999, retrieved August 2002 from http://www.arn.org/docs/odesign/od192/designinnature192.htm
[5] Pugh, S., *Total Design, Integrated Methods for Successful Product Engineering*, Addison-Wesley: Wokingham, p. 5, 1990.
[6] Cross, N., British Standards Design Management Systems BS7000. *Engineering Design Methods, Strategies for Product Design*, 3rd edn, John Wiley: Chichester, p. 201, 2000.
[7] Parent, E., *The Living Systems Theory of James Grier Miller*, 40th Anniversary Conference of the International Society for the Systems Sciences (ISSS), 1996 [electronic version], retrieved August 2002 from http://www.newciv.org/ISSS_Primer/asem14ep.html
[8] Heylighen, F., *Physiological Homeostasis*, Principia Cybernetica Project (PCP), retrieved August 2002 http://pespmc1.vub.ac.be/ASC/HOMEOSTASIS.html
[9] *Physiology of Organ Systems*, Department of Physiology, School of Medicine, University of Minnesota at Duluth, Minnesota, 1999, Retrieved August 2002 from http://www.d.umn.edu/medweb/phsl/physiology/5601.html
[10] Barnes-Svarney (ed.), The New York Public Library Science Reference Desk, Stonesong Press: New York, p. 221, 1995.
[11] Section of Plant Biology. *Reproduction of Large Organisms (Multicellular, Eukaryotic)*, University of California at Davis, 2002, retrieved August 2002 from http://www-plb.ucdavis.edu/courses/f98/bis1a/LifeCycl.htm
[12] de Magalhães, J.P., *A Unified View of Human Aging*, Gerontology Information, 2001 [electronic version], retrieved August 2002 from http://www.senescence.info/sense.htm
[13] Platonis Opera, Vol. 1, Euthyphro, Apologia Socratis, Crito, Phaedo, Cratylus, Theaetetus, Sophistes, Politicus, eds. E.A. Duke, W.F. Hicken, W.S.M. Nicoll, D.B. Robinson & J.C.G. Strachan, Clarendon Press: Oxford, 1995.
[14] Lovelock, J.E., *Gaia: A New Look at Life on Earth*, Oxford University Press: Oxford, 1979. New edition with updated Preface, 1987.
[15] Kauffman, S.S., *Investigations*, Oxford University Press, Oxford, pp. 215–216, 229, 2000.
[16] Margulis, L., *Symbiotic Planet: A New Look At Evolution*, Weidenfeld & Nicolson: London, 1998.
[17] Charlton, N., *Guide to Philosophy and the Environment*, retrieved August 2002 from http://www.lancs.ac.uk/users/philosophy/mave/guide/gaiath~1.htm
[18] Houghton, J., *Global Warming, The complete Briefing*, Lion Publishing: Oxford, pp. 124–125, 1994.
[19] Crowe, J., *An Investigation into the Evolutionary Roots of Form and Order in the Built Environment*, MIT Press: Cambridge, Massachusetts, pp. 20, 142, 1995.
[20] Rykwert, J., *The Dancing Column*, MIT Press: Cambridge, Massachusetts, pp. 46, 123–138, 1995.
[21] Benyus, J.M., *Bio-mimicry Explained*, Redesigning Resources Conference, 1999, [electronic version], retrieved August 2002 from http://www.redesigningresources.org/conference/con_menu.cfm
[22] Hawken, P., Lovins, A. & Lovins, H.L., *Natural Capitalism, Creating the Next Industrial Revolution*, Little Brown & Co: Boston, 1999.
[23] Buckminster Fuller Institute, *Who Is Buckminster Fuller?* 2001, retrieved August 2002 from http://www.bfi.org/introduction_to_bmf.htm
[24] Benyus, J.M., *Biomimicry: Innovation Inspired by Nature*, William Morrow, New York, 1997, retrieved August 2002 from http://www.biomimicry.org/case_studies_materials.htm and http://www.biomimicry.org/case_studies_processes.html

[25] Macquet, M. & Sweet, S., The myth of the closed-loop recycling system – towards a broader perspective on re-cycling, *10th Conference of the International Greening of Industry Network*, June 23–26. Gothenburg, Sweden, 2002.
[26] Jones, T., Reverse logistics – bringing the product back. Taking it into the future. *Strategic Supply Chain Alignment: Best Practice in Supply Chain Management*, ed. J. Gattorna, Aldershot: Gower, 1998.
[27] Archer, L.B., *Systematic Method for Designers*, reprinted from Design, The Design Council: London, 1966.
[28] Pahl, G. & Beitz, W., *Engineering Design*, The Design Council: London, 1984.
[29] World Commission for Environment and Development (WCED), *Our Common Future*, Oxford University Press: Oxford, 1987.
[30] Holmberg, J., (ed.). *Making Development Sustainable: Redefining Institutions, Policy, and Economics*, Island Press: Washington, DC, 1992.
[31] Diesendorf, M., Renewable energy and sustainability, ANZSES lecture, Sydney, May 2002, retrieved August 2002 from http://www.sustainabilitycentre.com.au
[32] Demmers, M. & Lewis, H., *Life Cycle Assessment: How Relevant is it to Australia?* Centre for Design, RMIT, 1996, http://www.cfd.rmit.edu.au/lca/LCAframe1.html
[33] Berkhout, F., *Life Cycle Assessment and Industrial Innovation*, ESRC Global Environmental Change Programme, 1996, retrieved August 2002 from http://www.sussex.ac.uk/Units/gec/pubs/briefing/brief-14.htm
[34] WTEC, *Life Cycle Analysis ISO 14040-14043*, Workshop on Environmentally Benign Manufacturing (EBM) Technologies, 2000, [electronic version], retrieved August 2002 from http://itri.loyola.edu/ebm/views/bras/sld012.htm
[35] The Nobel Foundation, *Guglielmo Marconi – Biography*, from Nobel Lectures, Physics 1901–1921, 2002, [electronic version], retrieved August 2002 from http://www.nobel.se/physics/laureates/1909/marconi-bio.html
[36] Andrews, S.D., *Sunrise to Sunset, A Brief History of British Radio Design*, unpublished MA thesis, Royal College of Art: London, 1986.
[37] Forty, A., *Objects of Desire, Design and Society 1750–1980*, Thames and Hudson: London, pp. 200–206, 1986.
[38] Schoenherr, S.E., *Recording Technology History*, History at the University of San Diego, 1999–2002, retrieved August 2002 from http://history.acusd.edu/gen/recording/notes.html
[39] du Gay, P., Hall, S., James, L., Mackay, H. & Negus, K., Doing Cultural Studies: The Story of the Sony Walkman, The Open University: Milton Keynes, 1997.
[40] Cross, N., *Engineering Design Methods, Strategies for Product Design*, 3rd edn, John Wiley: Chichester, 2000.
[41] Eckermann, E., *World History of the Automobile*, updated English translation of *Vom Dampfwagen zum Auto* (1989), Society of Automotive Engineers: Warrendale, PA, 2001.
[42] Eckermann, E., *Technikgeschichte im Deutschen Museum*: Automobile, Munich: C.H. Beck & Deutsches Museum, 1989.
[43] Whitt, F.R. & Wilson, D.G., *Bicycling Science*, 2nd edn, The MIT Press: Cambridge, MA, 1982.
[44] Lanchester, G., *A History of the Lanchester Motor Company*, transcript of a paper read before the Veteran Car Club 21st February; reprinted in *Lanchester Cars 1895-1965*, Freeman, A., Long, B. & Hood, Ch (eds., 1990) 24-29, Academy Books: London, 1948.

[45] Platt, M., 'Lanchester and the Motor Vehicle' transcript of a lecture delivered to the ImechE on 4th September; reprinted in *Lanchester Cars 1895-1965*, Freeman, A., Long, B. & Hood, Ch (eds., 1990) 30-44, Academy Books: London, 1968.
[46] Bird, A., '*Frederick Lanchester'*, eds. R. Barker & A. Harding, *Automobile Design: Great Designers and their Work*, 55-80, Newton Abbot: David & Charles, 1970.
[47] Grayson, S., '*The All-Steel World of Edward Budd'*, *Automobile Quarterly*, 4th quarter, **16(4)**, pp. 352–367, 1978.
[48] Nieuwenhuis, P. & Wells, P., *The Death of Motoring; Car Making and Automobility in the 21st Century*, John Wiley & Sons: Chichester, 1997.
[49] Rhys, D.G., New technology and the economics of the motor industry. *Proceedings of the International Conference on Future Developments in Technology: the Year 2000*, Waldorf Hotel, London, 4–6 April 1984.
[50] Haglund, R., *'Tell-all Tome; GM-Published History Includes Company's Woes'*, review of, W. Pelfrey (2000), General Motors: Priorities and Focus – Yesterday, Today and Tomorrow, in SAH Journal-The Newsletter of the Society of Automotive Historians, Issue Nr 191, March–April, p. 8, 2001.
[51] Naylor, G., *The Bauhaus reassessed, Sources and Design Theory*, The Herbert Press: London, 1985.
[52] Clutton-Brock, J., *Horse Power*, British Museum (Natural History): London, pp. 68–73, 12, 1992.
[53] Sperling, D., *Future Drive: Electric Vehicles and Sustainable Transportation*, Island Press: Washington DC, p. 1, 1995.
[54] Nieuwenhuis, P., Cope, P. & Armstrong, J., *The Green Car Guide*, Green Print: London, p. 1, 1992.
[55] Faiz, A., *et al.*, *Automotive Air Pollution: Issues and Options for Developing Countries*, Infrastructure and Urban Development Department, The World Bank, August 1990.
[56] Dargay, J. & Gately, D., Incomes effect on car and vehicle ownership worldwide 1960–2015. *Transportation Research A*, **3**, pp. 101–138, 1999.
[57] Clegg, M.W. & Dumoulin, H., The barriers to technological change: the case of transportation. *International Journal of Vehicle Design*, **13(5/6)**, pp. 443–448, 1992.
[58] Perkin, H., *The Age of the Automobile*, Quartet Books: London, p. 39, 1976.
[59] SMMT (Society of Motor Manufacturers and Traders Ltd.), *World Automotive Statistics*, Motor industry of Great Britain, London, 1999.
[60] DTI (Department of Trade and Industry), *The Energy Report – Market Reforms and Business Innovation*, The Stationery Office: London, 2000.
[61] DETR, Transport Statistics, *Focus on Personal Travel: 2001 edn*, The Stationery Office: London, [electronic version], retrieved August 2002 from http://www.transtat.dtlr.gov.uk/tables/2001/fperson/fpers01.htm, 2001.
[62] SMMT, *Motor Industry PR Campaign: Five Year Strategic Direction*, 21:07:98, Motor industry of Great Britain: London, 1998.
[63] 1998 Lex Report on Motoring 'Driving the Future', Lex Service PLC, research by MORI.
[64] Pettifer, J. & Turner, N., *Automania*, Guild Publishers: London, p. 124, 1984.
[65] Salomon, I. & Mokhtarian, P.L., Coping with congestion: understanding the gap between policy assumptions and behaviour. *Transportation Research Part D, Transport and the Environment*, **2(2)**, pp. 107–123, 1997.
[66] Jeremiah, D., Filling up – the British experience 1896–1940. *Journal of Design History*, **8(2)**, University of Plymouth, 1995.

[67] Reid, A., Murgatroyd, L. & Noble, B., *Variations in Travelling Time – the Social Context of Travel*, ONS DETR Transport Trends: GSS/DETR 0115520880, March 1999.
[68] Nye, D.E., *Consuming Power: A Social History of American Energies*, MIT Press: Massachusetts, pp. 236–237, 1998.
[69] Bayley, S., *Sex, Drink and Fast Cars*, Faber: London, 1986.
[70] Marsh, P. & Collett, P., *Driving Passion, The Psychology of the Car*, Jonathan Cape: London, p. 32, 1986.
[71] Black, S., *Man and Motor Cars: An Ergonomic Study*, Secker & Warburg: London, p. 65, 1966.
[72] Dowds, L. & Ahrecht, D., Fear of crime. *British Social Attitudes 12th Report*, eds. R. Jowens *et al.*, Dartmouth Publishing Co Ltd: Aldenshot.
[73] Glaister, S., Financing engineering or engineering finance, Inaugural lecture, Imperial College of Science Technology and Medicine, London, 22 February 2000.
[74] 1994 Lex Report on Motoring, *The Consumer View*, Lex Service PLC, research by MORI, London, p. 17, 1994.
[75] Baird, N., *The Estate We're In: Who's Driving Car Culture?* Indigo Paperback: London, 1998.
[76] Reid, A., Murgatroyd, L. & Noble, B., *Variations in Travelling Time – the Social Context of Travel*, ONS DETR Transport Trends: GSS/DETR, 1999.
[77] Mitchell, C.G.S. & Lawson, S.D., How the Car Has Changed our Travel and our Lives, Transport Trends: GSS/DETR, The Stationery Office: London, 1999.
[78] CfIT (Commission for Integrated Transport), *Road Traffic Reduction (National Targets) Act 1998: Tackling Congestion and Pollution – The Government's First Report*, 1999, DETR, The Stationery Office: London, [electronic version], retrieved August 2002 from http://www.roads.dtlr.gov.uk/roadnetwork/rtra98/report1/fore.htm
[79] Rathbone, D.B. & Huckabee, J.C., *Controlling Road Rage A Literature Review and Pilot Study Prepared for the The AAA Foundation for Traffic Safety*, The InterTrans Group 9:06 99, 1999.
[80] DETR, *Transport 2010 – The Ten Year Plan*, The Stationery Office: London, 2000.
[81] DETR, *Road Accidents Great Britain 1998*, The Stationery Office: London, 1999.
[82] DETR, *Road Accidents Great Britain 1999*, The Stationery Office: London, 2000.
[83] WHO, The Third World Health Organisation Ministerial Conference on Environment and Health, *London, Health Costs due to Road Traffic-related Pollution; An Impact Assessment Project in Austria, France and Switzerland*, WHO, Geneva, 1999.
[84] WHO (World Health Organisation), *World Health Report 1999: Making a Difference*, WHO, Geneva, 1999.
[85] RCEP (The Royal Commission on Environmental Pollution) 22nd report, *Energy – The Changing Climate*, The Stationery Office: London, 2000.
[86] NAEI (UK National Atmospheric Emissions Inventory), DETR with the London Research Centre in collaboration with RSK Environment (1999), *The UK Emission Factors Database*, AEA Technology, NOAA (US Federal National Oceanic and Atmospheric Administration) (2000), El Niño Page version, retrieved August 2002 from http://www.elnino.noaa.gov/
[87] HSPH (The Harvard School of Public Health) on behalf the World Health Organisation and the World Bank, *The Global Burden of Disease*, 2000, retrieved August 2002 from http://www.hsph.harvard.edu/organizations/bdu/gbdsum/gbdsum6.pdf
[88] DETR, *Transport Statistics of Great Britain 2001*, 27th edn, [electronic version], retrieved August 2002 from http://www.transtat.detr.gov.uk

[89] Foley, G., *The Energy Question*, Penguin Books, 1992.
[90] *The Economist*, The future of energy, October 7, 1995.
[91] Riva, J.P., Oil distribution and production potential. *Oil & Gas Journal*, pp. 58–61, January 18, 1988.
[92] MacKenzie, J.J., *Oil as a finite resource: when is global production likely to peak?* World Resources Institute 1996, updated 1999, www.lgc.org/wri/climate/jm_oil_001.html
[93] DTI, *The Energy Report, 2000*, [electronic version], retrieved August 2002 from http://www.dti.gov.uk/EPA/digest00/contents00.htm
[94] Porritt, J., *Engineering: The Key to a Sustainable Future*, Institution of Incorporated Engineers, 1999.
[95] Automotive Intelligence [electronic version], retrieved August 2002 from http://www.autointell.com/management/sales_stats-2000.htm
[96] Jacques, A., *Environmental Implications of the Automobile*, 1992, *Road Vehicle Fuel Economy*, TRRI/HMSO, 1992.
[97] Oak Ridge National Laboratory, operated by UT-Battelle, LLC, under contract for the US Department of energy, *The Environmental Impact of the Car*, a Greenpeace Report, London, 1992.
[98] Ashby, M.F. & Johnson, K.W., *Classification and Choice in Product Design*, Cambridge University Engineering Department, Technical Report CUED/C-EDC/TR108, 2001.
[99] Pollution Probe, *The Costs of the Car: A Preliminary Study of the Environmental and Social Costs Associated with Private Car Use in Ontario*, Toronto, 1991.
[100] Energy Information Association, [electronic version], retrieved August 2002 from http://www.eia.doe.gov/
[101] Gow, D., *The Guardian*, Monday, September 3, 2001.
[102] Directive 2000/53/EC of the European Parliament and of the Council of 18 September 2000 on end-of life vehicles, Brussels, 2000.
[103] Waste Watch Wasteline, *Car Recycling*, 2002, [electronic version], retrieved August 2002 from http://www.wastewatch.org.uk/informtn/carrec.htm
[104] Clarke, S., Managing design: the art and colour section at general motors 1927–41. *Journal of Design History*, **12(1)**, Oxford University Press: Oxford, p. 66, 1999.
[105] Cowan, R.S., *A Social History of American Technology*, OUP: Oxford, p. 228, 1997.
[106] PNGV (Partnership for New Generation Vehicles), [electronic version], retrieved August 2002 from http://www.uscar.org/pngv/index.htm
[107] US Department of Energy, *Energy Secretary Abraham Launches FreedomCAR, Replaces PNGV*, [electronic version], retrieved August 2002 from www.energy.gov/HQPress/releases02/janpr/pr02001.htm, 2001.
[108] TransportAction PowerShift, *Clean Fuel Vehicles Market Report*, 2001.
[109] DETR, *Graduated Vehicle Excise Duty Reforms*, [electronic version], retrieved August 2002 from http://www.dvla.gov.uk/newved.htm, 2001.
[110] Retrieved August 2002 from http://www.cartech.doe.gov/freedomcar
[111] BP p.l.c., *Low Sulphur Diesel – Frequently Asked Questions*, [electronic version], retrieved August 2002 from http://www.bp.com.au/products/fuels/low_sulphur/faq.asp?menuid=ec, 1999–2002.
[112] Hucho, W-H., *Aerodynamics of Road Vehicles, From Fluid Mechanics to Vehicle Engineering*, Butterworths: London, pp. 34–46, 1987.
[113] Seifert, U. & Walzer, P., *Automobile Technology of the Future*, Society of Automotive Engineers: Warrendale, PA, pp. 27–30, 1992.

[114] Wheels 24, *VW unveils 1 litre/100 km car*, News 24.com, 2002, [electronic version], retrieved August 2002 from www.news24.com/News24/Wheels24/News/0,3999, 2-15-47_1168834,00.html
[115] DEFRA (Department of Environment Food and Rural Affairs, The Renewables Obligation Order 2002, 2002 No. 914 Electricity, England and Wales, [electronic version], retrieved August 2002 from http://www.defra.gov.uk/, Crown Copyright 2002.
[116] BBC Farming Today, [electronic version], retrieved August 2002 from http://www.bbc.co.uk/radio4/news/farmingtoday/index.shtml, 2002.
[117] Schiffer, M.B., *Taking Charge, The Electric Automobile in America*, Smithsonian Institution Press: Washington, 1994.
[118] Rand, D.A.J., Woods, R. & Dell, R.M., *Batteries for Electric Vehicles*, Research Press Studies, 1998.
[119] Riley, R.Q., *Alternative Cars in the 21st Century: A New Personal Transportation Paradigm*, Society of Automotive Engineers: Warrendale, PA, pp. 225–246, 1994.
[120] *Honda Insight*, Marketing Brochure, Honda (UK) Cars, Slough, 2000.
[121] DaimlerChrysler, *Background: Fuel Cell Technology: Independent and Versatile*, 2002, [electronic version], retrieved August 2002 from http://www.daimlerchrysler.com/index_e.htm?/news/top/2000/t01107b_e.htm
[122] *Toyota Prius*, Marketing Brochure, Toyota (GB) PLC, Redhill, 2000.
[123] Levin, D., *Ford Finds that it's Not Easy Being Green*, Detroit News Auto-insider, [electronic version], retrieved August 2002 from http://detnews.com/2001/insiders/0112/06/-360522.htm
[124] Interview with Honda UK Marketing Executive, 27:4:01
[125] Interview with Nick Wilson, Toyota UK Marketing Executive 27:4:01
[126] Ford Motor Company, The Escape, Hybrid Electric Vehicle, [electronic version], retrieved August 2002 from http://www.hybridford.com/Error.asp
[127] *The Car That Will Save The World*, The Telegraph, 1999.
[128] Contadini, J.F., *Social Cost Comparison Among Fuel Cell Vehicle Alternatives*, the Methanol Institute, [electronic version], retrieved August 2002 from http://www.methanol.org/fuelcell/special/contadini_pg1.html
[129] France, K., *The end-of-life vehicle: position in the EU*, [electronic version], retrieved August 2002 from http://www.twi.co.uk/j32k/protected/band_3/kslkf001.html
[130] Hencke, D., Landfill scheme 'must come clean', *The Guardian*, 2002.
[131] Scott, K., Victims of burgeoning waste crisis, *The Guardian*, 2002.
[132] The European Commission, *Proposal for a Directive of the European Parliament and of the Council on Waste Electrical and Electronic Equipment and on the Restriction of the Use of Certain Hazardous Substances in Electrical and Electronic Equipment*, COM (2000) 347, Brussels, [electronic version], retrieved August 2002 from http://europa.eu.int/ eur-lex/en/com/pdf/2000/en_500PC0347_02.pdf
[133] Brown, P., Councils get £40m for fridge mountain, *The Guardian*, 2002.
[134] CARE (the Consortium for Automotive Recycling), *Glass Recycling: An Automotive Perspective*, 1999, [electronic version], retrieved August 2002 from http://www.caregroup.org.uk/download.shtml
[135] ACORD (Automotive Consortium on Recycling and Dismantling), *Second Annual Report: Summer 1999 (Reporting 1998 Performance).*
[136] Charles Trent Ltd, *The Disposal of End of Life Vehicles in the UK*, 2000, [electronic version], retrieved August 2002 from http://www.trents.co.uk

[137] Association of Plastics Manufacturers in Europe (APME), *Plastics: A Material of Choice for the Automotive Industry*, 1999, [electronic version], retrieved August 2002 from http://www.apme.org/literature/htm/02.htm
[138] Martorana, D. & Zeloni, L., *Ecological Use of Materials in the Car: Cooperation between Industry and Public Research Centre*, Elasis (Fiat Research Centre), Turin, [electronic version], retrieved August 2002 from http://www.elasis.it/it/rassegna-stampa/download/cnr-roma.pdf.
[139] The Environment Agency, *Tyres in the Environment*, [electronic version], retrieved August 2002 from http://www.environment-agency.gov.uk/envinfo/tyres
[140] UK tyre disposal, http://www.tyredisposal.co.uk/disposaluk.htm, 2002.
[141] Klein, N., *No Logo*, Flamingo: London, 2000.
[142] General Motors Corporation, *GM's AUTOnomy Concept Vehicle Reinvents Automobile*, [electronic version], retrieved August 2002 from http://gm.com/company/gmability/environment/products/adv_tech/autonomy1_010702.html
[143] Rocky Mountain Institute, [electronic version], retrieved August 2002 from http://www.rmi.org
[144] *How to Reduce the Environmental Impacts of Transportation?* [electronic version], retrieved August 2002 from http://www.rmi.org/sitepages/pid386.php
[145] *How Hypercars work*, [electronic version], retrieved August 2002 from http://www.rmi.org/sitepages/pid18.php
[146] *Hypercar Design and Technology, A New Perspective on Automobile Design - Whole-System Design*, [electronic version], retrieved August 2002 from http://www.rmi.org/sitepages/pid390.php
[147] *Hypercar Inc.*, [electronic version], retrieved August 2002 from http://www.hypercar.com/
[148] RMI Library: *Transportation*, [electronic version], retrieved August 2002 from http://www.rmi.org/sitepages/pid175.php
[149] Cramer, D.R. & Brylawski, M.M., *Ultra-light-hybrid vehicle design: implications for the recycling industry*, Society of Plastics Engineers Recycling Division's 3rd Annual Recycling Conference Proceedings (Chicago, IL) 7-8 November, 1996.
[150] Mascarin, A.E., Dieffenbach, J.R., Brylawski, M.M., Cramer, D.R. & Lovins, A.B., *Costing the Ultralite in Volume Production: Can Advanced Composite Bodies-In-White be Affordable?*
[151] International Body Engineering Conference and Exposition, Detroit, MI, 31 October–2 November, 1995.
[152] Brylawski, M.M. & Lovins, A.B., *Advanced Composites: The Car is at the Crossroads*, The Hypercar Center, Rocky Mountain Institute: Snowmass, Colorado, 1996.
[153] VCD *Auto-Umweltliste '96*, Bonn, Verkehrsclub Deutschland, 1996.
[154] Fischer Th 'Zum Thema Schema: das neue Testraster von auto motor und sport', *Auto Motor u Sport*, 6/1997, 30 March, 1997.
[155] What Car? 'Britain's filthiest cars, its cleanest cars ... and the real culprit; Green Car Guide', *What Car?*, pp. 60–65, 1998.
[156] ETA *Car Buyer's Guide*, Weybridge, Environmental Transport Association, 1994.
[157] ETA *Car Buyer's Guide 95*, Weybridge, Environmental Transport Association, 1995.
[158] AEA 'Government and Policy', *FT Automotive Environment Analyst*, Issue 42, 13–16 July 1998.

[159] Nieuwenhuis, P. & Wells, P., *Developing an Environmental Rating System for Cars*, Presented to the 7th International Conference of the Greening of Industry Network, 15-18 November, Rome, Italy, 1998.
[160] Nieuwenhuis, P. & Wells, P. (eds). *Motor Vehicles in the Environment*, John Wiley: Chichester, 1994.
[161] Weizsäcker, E., von Lovins, A. & Lovins, L., *Factor Four; Doubling Wealth – Halving Resource Use; The New Report to the Club of Rome*, Earthscan: London, 1997.
[162] Hart, S., Beyond greening: strategies for a sustainable world. *Harvard Business Review*, January-February, pp. 66–76, 1997.
[163] Nattrass, B. & Altomare, M., *The Natural Step for Business; Wealth, Ecology and the Evolutionary Corporation*, New Society, Gabriola Island, p. 23, 1999.
[164] Nieuwenhuis, P., *Is Sustainable Car Making Possible?* Paper presented to the 10th International Greening of Industry Conference, June 23-26, Gothenburg, Sweden, 2002.
[165] Ryan, J. & Thein Durning, A., *Stuff; The Secret Lives of Everyday Things*, Seattle: Northwest Environment Watch, 1997.

Chapter 14

Emergent behaviours in autonomous robots

B. Hutt, K. Warwick & I. Goodhew
Department of Cybernetics, University of Reading, UK.

Abstract

It is well known from the field of artificial life that complex and life-like behaviours can be generated by apparently simple systems. Such behaviours are considered as being emergent, as the simple systems that generate the behaviours do not themselves explicitly mention the behaviour generated. Using such an approach it is possible to generate complex life-like behaviours in robots by merely following a simple set of rules. Flocking in birds, herding in cattle and shoaling in fish are all common examples of emergent coordinated group behaviours found in nature. Despite the apparent complexity of these behaviours, it is possible to convincingly synthesise such behaviours in robots using only simple sensors and programming. The evolution of life itself can be considered as an emergent process with intelligence and many other interesting properties emerging through the mechanisms of evolution. Algorithms based on the mechanisms of evolution have been used with great success to evolve robot behaviours that are doubly emergent – where the evolved behaviour is itself an emergent property of the evolutionary process and the implementation of that behaviour is also emergent, as it is the product of simple rule following that does not mention the behaviour.

1 Introduction

The natural world is teeming with a vast range of different biological organisms many of them exhibiting extremely complex behaviours. It is well known that apparently complex and life-like behaviours can be generated in artificial systems by surprisingly simple systems – a good example of this is Conway's game of life [1]. In Conway's game high-level patterns and structures form from extremely simple low-level rules.

The game of life is played out on an infinitely large grid of square cells rather like a chessboard. In this system every cell in the grid is surrounded by eight other cells; with each cell on the grid in one of two states – dead or alive. It is the number of live cells surrounding each cell that determines the next state of the cell. The rules that govern the game of life are as follows – a dead cell with exactly three live neighbours becomes a live cell, a live cell with two or three live neighbours remains alive, in all other cases the cell becomes (or remains) dead. The elaborate patterns that can

be produced are really quite surprising when we consider they are purely the result of following such simple rules – the game of life is one of the simplest examples of emergent complexity.

2 Complexity from simplicity – emergent behaviour from simple rules

Amongst the earliest work to explore the complexity of behaviour that could be produced with simple artificial systems is the work of William Grey Walter.

2.1 Early reactive robots

In 1948 William Grey Walter built his first 'tortoise' robot. Walter was interested in the complexity of behaviour that minimally simple robots were capable of producing. Consequently, the 'brain' of his tortoise robot was deliberately restricted to just two functional elements (miniature radio valves).

> In order to discover what degree of complexity of behaviour and independence could be achieved with the smallest number of elements in a system providing the greatest number of interconnections [2].

The robot was based on a three wheeled (tricycle shaped) chassis enclosed by a hard shell – it had two motors one driving both rear wheels, in order to propel the robot, and the other controlling the front steering wheel. It had only two sensors; a photocell mounted on top of the steering column, at the front of the robot, pointing in the same direction as the steering wheel and a contact switch attached via the shell so that when an object pushed on the shell it closed a contact enabling collisions to be detected.

Each tortoise had two basic reflexes (or 'instincts') that enabled it to head towards sources of light and negotiate obstacles. The primary, light seeking, behaviour used the photocell to search for and drive the robot towards sources of light, however, if the light source was too strong it caused the robot to steer away from the light instead of towards it. The second behaviour, obstacle avoidance, was triggered by the contact sensor, which caused the tortoise to retreat and then sidestep, thereby allowing the robot to avoid or push aside obstacles. The robots behaved in a phototropic manner somewhat analogous to flight paths of moths – executing graceful curves as they searched the environment for sources of light and retreating when they encountered an obstacle or got too close to a light source.

Walter's next robot was even more complex, having the ability to learn a simple conditioned response. This robot possessed an additional module 'bolted on' to the back of his previous robot which created a connection between the robot's light reflex, its contact reflex and a third sensor that detected when a whistle was blown. The robot could be trained by blowing the whistle and then kicking the robot initiating the obstacle avoidance reflex.

> After five or six beatings, whenever the whistle was blown [the robot] turned and backed from an 'imagined' obstacle [3].

The level of autonomy and behavioural complexity achieved by Walter's tortoise robots is really quite surprising, especially considering the technological limitations of the day. Walter had produced a proof of concept that complex behaviours can stem from the interaction of a few ingeniously connected but simple parts [4].

> These machines are perhaps the simplest that can be said to resemble animals. Crude though they are, they give an eerie impression of purposefulness, independence and spontaneity [2].

2.2 Thought experiments with simple robots

Unfortunately it appears that the significance of Walter's work was not properly understood at the time and it was not until the mid 1980's that researchers once again started to consider the potential capabilities of simple mobile autonomous robots [5]. Through a set of thought experiments Braitenberg investigated the properties of a number of simple robot controllers capable of generating unexpectedly complex behaviours. He argued that if we did not know the simple principles by which the robots were functioning we might call such behaviours 'aggression', 'love', 'foresight' and even 'optimism'. Interestingly, Braitenberg also considered the possibility of using evolution to design robot controllers as well as the effects of Hebbian style learning.

3 Modern reactive robots

In the mid 1980's Rodney Brook's group at MIT began to examine the performance of simple layered controllers with real robots using his subsumption architecture ([6, 7]). Robots designed and programmed using such methods are capable of exploring a typical office environment in real time, reacting to and avoiding obstacles with few problems. The performance of such robots was significantly better than other mobile robots reportedly in existence at the time so that in [8] the following claim was (rather modestly) made:

> We claim as of mid-1987 our robots, using the subsumption architecture to implement complete Creatures, are the most reactive real-time mobile robots in existence. Most other mobile robots are still at the stage of 'experimental run' in static environments, or at best in completely mapped static environments. Ours, on the other hand, operate completely autonomously in complex dynamic environments at the flick of their on switches, and continue until their batteries are drained [8].

3.1 Reactive robots

Essential to these systems is the concept of emergent behaviour, i.e. the idea that a behaviour is often the by-product of simple rule following that does not specifically mention the behaviour exhibited. Basic Braitenberg vehicles of the type shown in Fig. 1 provide an extremely simple example of emergent behaviour, such vehicles are incredibly minimal but can exhibit surprisingly complex behaviour [5].

Some schematics of basic Braitenberg vehicles are shown in Fig. 1. Two different vehicles are shown. Each vehicle has two light-sensitive sensors and two motors, with which to move. The sensors are used to directly control the motors, such that if a sensor is activated by a light source its output enables power to be supplied to a connected motor. There are two obvious possibilities, either the sensors can control the motors on the same side of the vehicle as the sensor, or they can control the motors on the opposite side. When exposed to a light source the vehicle pictured on the left in Fig. 1 will try to evade it, whilst the vehicle shown on the right will tend to approach it.

There is no explicit mention of any light evading or following goal in such systems, rather the behaviour is implicit in the structure of the robot itself and the way in which its body allows it to interact with the environment. More importantly there is no representation or internal model of the world and therefore there is no separate perception, planning or action. Effectively the control system is entirely decentralised.

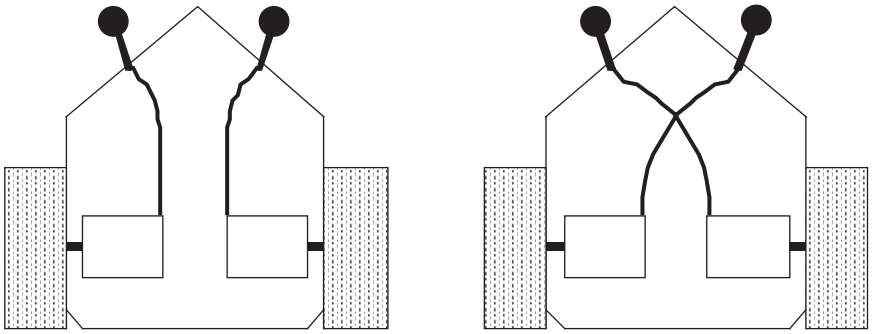


Figure 1: Braitenberg vehicle schematic. Each vehicle has two light sensors (illustrated by black circles on stalks) connected to two motors. The vehicle shown on the left will avoid light, whilst the vehicle shown on the right will tend to head towards light, the only difference between the two vehicles is that the two internal wires are swapped around.

4 More complex behaviours

The natural world contains many examples of coordinated animal behaviour, e.g. flocking in birds, herding in cattle, shoaling in fish and swarming in insects to name a few. These group behaviours are fascinating as such groups are composed of many individual animals, yet the motion of the entire group of animals is well-coordinated, smooth and fluid in nature. Such groups of animals appear to move under some form of intelligent and centralised control. However, these group behaviours are merely the result of many individuals interacting with one another solely on each individual's local perception of the environment [9].

The first published work that truly captured the motion of large groups of animals was done by Craig Reynolds within the field of computer animation. Reynolds was interested in simulating the aggregate motion of animals exhibiting coordinated group behaviours such as flocks of birds. In his 1987 paper [9], Reynolds describes a system that is capable of simulating the behaviour of a flock of birds in a realistic manner. Reynolds' system was based on an elaboration of a particle-based system, with each particle representing an individual 'bird' and the entire collection of particles representing the flock. Reynolds named his simulated birds 'boids'. The real breakthrough of Reynolds' approach was that he used a distributed behavioural system; each boid was individually simulated as a separate agent and was able to choose its own course based on local information such as the location and speed of other nearby boids in the flock.

Reynolds' flocking system was based on three simple steering behaviours, these steering behaviours are shown in Fig. 2. The first steering mechanism, *separation*, ensures that there is sufficient space between nearby boids. The separation behaviour causes each boid to steer in order to avoid crowding local flock mates. The second steering behaviour, *alignment*, ensures that boids are all heading in the same direction. The alignment behaviour causes each boid to steer towards the average heading of its local flock mates. The third steering behaviour, *cohesion*, acts to keep the flock from separating. The cohesion behaviour causes each boid to steer towards the average position of local flock mates.

In Reynolds' model a boid will only react to boids in its own local neighbourhood. This neighbourhood is shown in grey in Fig. 2. A boid's neighbourhood is defined by both the range and angle measured from the current direction of flight, any boid outside the neighbourhood is ignored (see Fig. 3). The neighbourhood acts in order to model a limited field of perception. This is perhaps

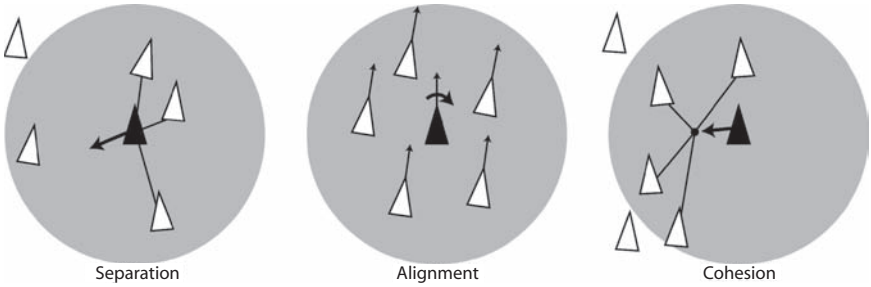


Figure 2: The three steering behaviours used in Reynolds' flocking system. The separation behaviour causes each boid to steer to avoid local flock mates. The alignment behaviour causes each boid to steer towards the average heading of its local flock mates. The cohesion behaviour causes each boid to steer towards the average position of local flock mates.

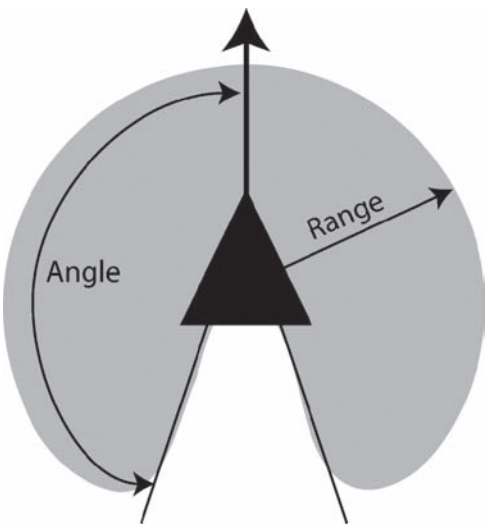


Figure 3: The neighbourhood of a boid.

somewhat akin to the field of vision of a fish swimming in murky water, with the fish able to view nearby objects ahead of it and to the side but not behind.

Many papers have been published on the simulation of animal group behaviours and a large variety of different flocking models are available [10, 11].

Reynolds' flocking model was designed for use in films and therefore used scripted motion in which the boids have a weak goal-seeking behaviour that causes them to move to predefined locations. By using predefined destinations, boids do not require the use of leaders.

When a flock of Reynolds' boids encounters an obstacle, they split up and go around both sides of the obstacle, much like a real flock of birds. Once the obstacle has been passed the boids rejoin to form the original flock. Sometimes, as in nature, some of the boids become trapped between the flock and an obstacle and almost fly into the obstacle; however, they quickly recompose themselves and rejoin the flock.

5 Hardware implementation

For any form of group activity in physical robots they require the ability to perceive the relative positions of their teammates in addition to obstacles within their environment. In order to obtain co-operation between multiple robots, some form of direct or indirect communication between the robots is also required.

In [12] a group of flocking robots are described. These robots use ultrasonic range-finding sensors to detect obstacles in the environment and to detect the range and angle to other robots in the flock by means of an infrared localisation system.

A modified version of Reynolds' flocking algorithm is used because, unlike Boids, the hardware robots have no pre-defined goal to head towards. In order to overcome this problem Kelly introduced the concept of flock leaders – the algorithm running on each robot then becomes:

1. AVOID OBSTACLES [most basic high-level behaviour].
2. If no other robots visible become a LEADER and WANDER.
3. If in a FLOCK try to maintain position.
4. If a flock can be seen in the distance, speed up and head towards it, with more priority being given to following the closest visible leader.

Both obstacle avoidance and wandering behaviours are achieved using information acquired purely from the ultrasonic range-finding sensors. The range and angle to other robots in the locality is determined solely from the infrared localisation system. In Kelly's system only the infrared emitters facing towards the rear of the robot are switched on, causing robots to follow behind a robot. As well as range and angle information being conveyed by infrared signalling an additional infrared signal is also communicated allowing robots to determine whether an individual is a leader or a follower.

In this hardware implementation, the selection of the leader has to be dynamic, since like Reynolds' boids (and flocks of animals) the flock needs to be able to split up, go around obstacles and then rejoin once past the obstacle. If the leader was to be pre-defined this would not be possible. Also, since any real-world environment is finite there are boundaries which would cause problems when the flock meets the boundary of the environment. In this case the pre-defined leader would have to fight its way through the other robots. In addition to this if the pre-defined leader was to fail then the whole flock would also fail.

In Kelly's system the leadership is dynamic, any robot can become a leader and can relinquish leadership when required and more than one leader can co-exist. In this system the flock can split up into two smaller flocks to go around both sides of an obstacle and then rejoin once past the obstacle. If the leader gets trapped between other robots, then it is now part of a flock and therefore relinquishes leadership. One of the robots on the outside of the flock then takes over leadership and the rest of the robots follow it. In order to ensure that this leader does not just turn around and rejoin the main body of the flock a short-time delay was added during which the robot is not allowed to relinquish leadership to any robots that are followers. However, even during this delay the robot will still relinquish leadership to another leader in front of it.

Kelly's flocking system is flexible with his leadership priority system depending on the number of 'visible' robots. The weighting given to following any particular leader is equal to the number of other visible robots in front of any given robot. It is interesting then to note the tendencies of different robots to become a leader and their pulling power on other robots when they are a leader. Physical differences between the robots clearly play a part in this as do the sensor characteristics. A problem occurring in a specific robot then has an important role to play in its ability to become a leader.

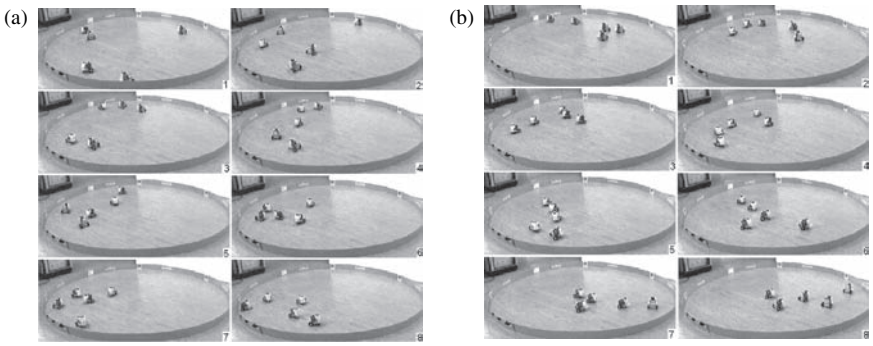


Figure 4: Kelly's work shows that an emergent flocking behaviour is possible with co-operation occurring between multiple autonomously controlled robots. With each robot only using a low-bandwidth communication system utilising data from only a few simple sensors combined with some simple rules, resulting in apparently complex group behaviours.

From the point of view of a single robot, when no other robots are visible in front, it can become a leader. As a leader the robot will wander around by moving forwards in a straight line until it encounters an obstacle, at which point it will turn away from it. Robots that are followers head towards the greatest density of visible robots, with a higher priority being given to following the leader. This attracts robots towards each other, thereby forming flocks. If a follower sees a flock in the distance it will attempt to catch up with them by increasing its speed. If any robot is in a flock it will try to match the speed and direction of its nearest neighbours much like in Reynolds' boids system.

Using Kelly's flocking algorithm the robots display an emergent flocking behaviour in the presence of boundaries with both moving and stationary obstacles. Two separate flocking sequences are shown in Fig. 4a and b.

6 Emergent behaviour through evolution

The evolution of life itself can be considered as an emergent process with intelligence and many other interesting properties emerging through the mechanisms of Darwinian evolution. In Darwin's theory, random variants continuously arise in a population, a fraction of which happen to result in some population members having a reproductive advantage. These variants, over time, tend to result in the replacement of those individuals that are less productive. Over many generations, this *natural selection* causes a population to change and adapt or evolve.

In order for evolution to operate there are just a few basic requirements. These are:

- *Variation* – within a population there must be some variations between individuals.
- *Heritability* – the traits that an individual possesses must be able to be inherited by that individual's offspring.
- *Selection* – there must be some form of selection in place that gives individuals with beneficial traits a reproductive advantage over those that have neutral or deleterious traits.

These requirements seem pretty simple yet the complexity such evolutionary systems are able to create are astounding. It is possible to mimic the processes of biological evolution in order to

evolve robot behaviours from scratch. However, any artificial evolutionary system must include the basic requirements for evolution listed above.

6.1 Artificial evolution

There are many ways in which an artificial evolutionary system could be implemented; however, by far the most common is genetic algorithms [13]. Genetic algorithms have been applied to a diverse range of different applications including machine learning, task planning, automatic programming and many more. Many of these use very different implementations to Holland's original genetic algorithm, often blurring the distinction between other evolutionary algorithms – the term genetic algorithm is therefore fairly broad and is often used to describe any population based method that uses some form of selection and reproduction with variation in order to produce new sample points within a search space.

6.2 Genetic algorithms

The classical genetic algorithm, whilst loosely inspired by biological evolution, is typically more concerned with optimisation problems than implementing a biologically plausible model of evolution. However, genetic algorithms still employ the basic principles of biological evolution in order to solve problems, namely selection with some form of reproduction with variation.

As with all evolutionary algorithms, genetic algorithms are population based methods, that is, they sample multiple points in a problem space at the same time. Each member of the population occupies only a single point in the problem space, with the entire population consisting of many such points. As with any optimisation algorithm (other than random or exhaustive searches) the current sample points are used to bias the selection of subsequent sample points, hopefully resulting in an effective search of the problem space.

A standard genetic algorithm will typically start with a population of randomly initialised population members representing a set of random points in a problem space, the quality of each individual solution to the problem being measured by some objective or *fitness function*. Those individuals representing better problem solutions, having a greater fitness, are then selected in such a way as to give a reproductive advantage over those points with lower fitness.

In a genetic algorithm good parent solutions are recombined in order to form new solutions, this is achieved through the process of crossover and mutation. In the simplest form a random point is selected and the chromosome of both parents are cut at this point, the child chromosome then inherits the first chunk of genetic material from the first parent and the second chunk of genetic material from the second parent. The genome of the offspring is then subjected to mutation – a common way of implementing this is to choose a bit at random in the genome of the offspring and invert it. The process of crossover and mutation is illustrated in Fig. 5.

The process of selection, crossover and mutation is repeated until enough offspring have been formed to replace the entire population. This set of offspring forms a new population or *generation* of individuals – the general flow of a genetic algorithm is as follows:

1. Initialise all N chromosomes with random data.
2. Decode and evaluate all N chromosomes thereby obtaining a fitness value for each individual.
3. Select two individual chromosomes from the population based on their fitness value.
4. Recombine these chromosomes using the crossover operator in order to form a new chromosome.
5. Subject this new chromosome to the mutation operator.

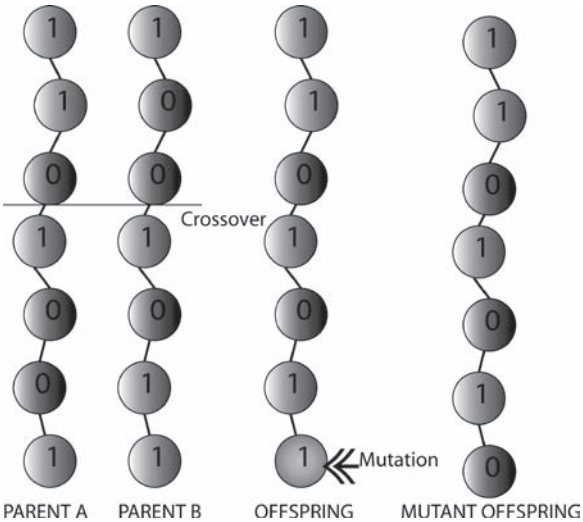


Figure 5: A mutant offspring is created from two parents via crossover and mutation.

6. Repeat steps 3–5 until N new chromosomes have been generated – use these to form the next generation.
7. Repeat steps 2–6 until either a sufficient fitness is attained or a fixed number of generations have passed.

For a good introduction to genetic algorithms see [14].

6.3 Evolutionary robotics

In the field of evolutionary robotics a genetic algorithm is used to design from scratch a controller for a robot. This controller often takes the form of an artificial neural network and the job of the genetic algorithm is to choose the relevant weights and biases for the neural network in order to optimise some behavioural fitness criterion (see Fig. 6).

The choice of fitness function is critical as it will determine the resulting behaviour – in fact when this fitness criterion is implicit then the evolved behaviour can often be considered as being emergent. An implicit function is one that does not describe the specifics of a task but rather the general goal, for example, an explicit function might reward *specific* motions of a robot such as moving forwards and turning away from obstacles. In contrast an implicit function might reward a robot for the area it covers within a fixed period of time. Such functions are far more general and less specific than explicit functions and they allow the genetic algorithm a great deal more freedom in solving a problem. Due to this there can be many potential solutions to the problem – in other words implicit functions are often highly multi-modal. With some modifications to the basic genetic algorithm [15], it is possible to evolve multiple solutions or *species* that achieve a high fitness score with each species using different techniques to achieve that fitness.

In our experiments a population of 100 simple differential drive robots equipped with ultrasonic range-finding sensors were simulated and allowed to evolve. The population was evolved for 500 generations in a square 10×10 m arena with a fitness function that rewarded robots purely for the area covered within a fixed period of time. Each robot was initialised to a random position and

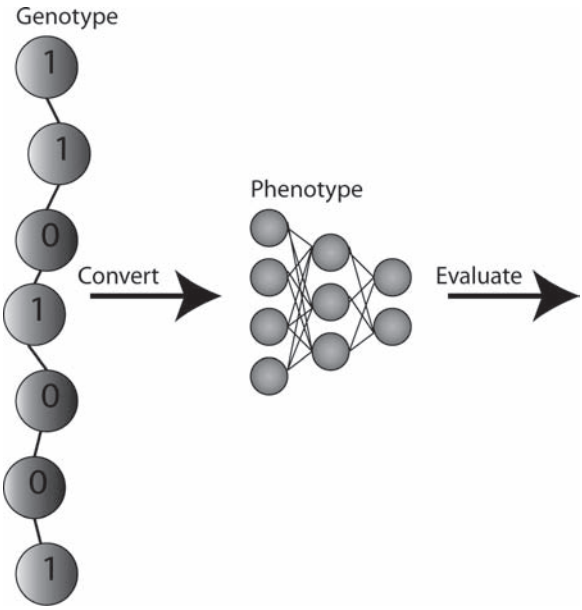


Figure 6: The genotype must first be converted into the phenotype, a neural network; it is the phenotype that is then evaluated on its ability to solve the problem.

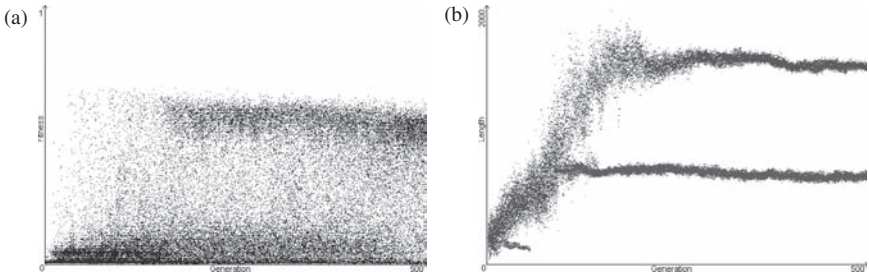


Figure 7: (a) Population fitness variation scatter plot. (b) Population genome length variation scatter plot.

orientation within the arena, left to run for 2 min. and evaluated for the area it covered in that time. Figure 7a shows the evolution of fitness whilst Fig. 7b shows the evolution of genome length. In this particular case two distinct species are formed as can be observed from the genome length plot. The area of the arena covered by the robots in various generations as evolution progressed is shown in Fig. 8.

Evolution was allowed to run for 500 generations at which point the fittest robot from each species was removed from the population and tested in isolation. The resulting motion of these robots is shown in Fig. 9 for the species with the longest genome and in Fig. 10 for the species with the shorter genome. As can be seen each species solves the problem of covering the greatest area in rather different ways. The first species solves the problem by utilising a wall-following behaviour whilst the second species solves the problem by utilising a far more chaotic

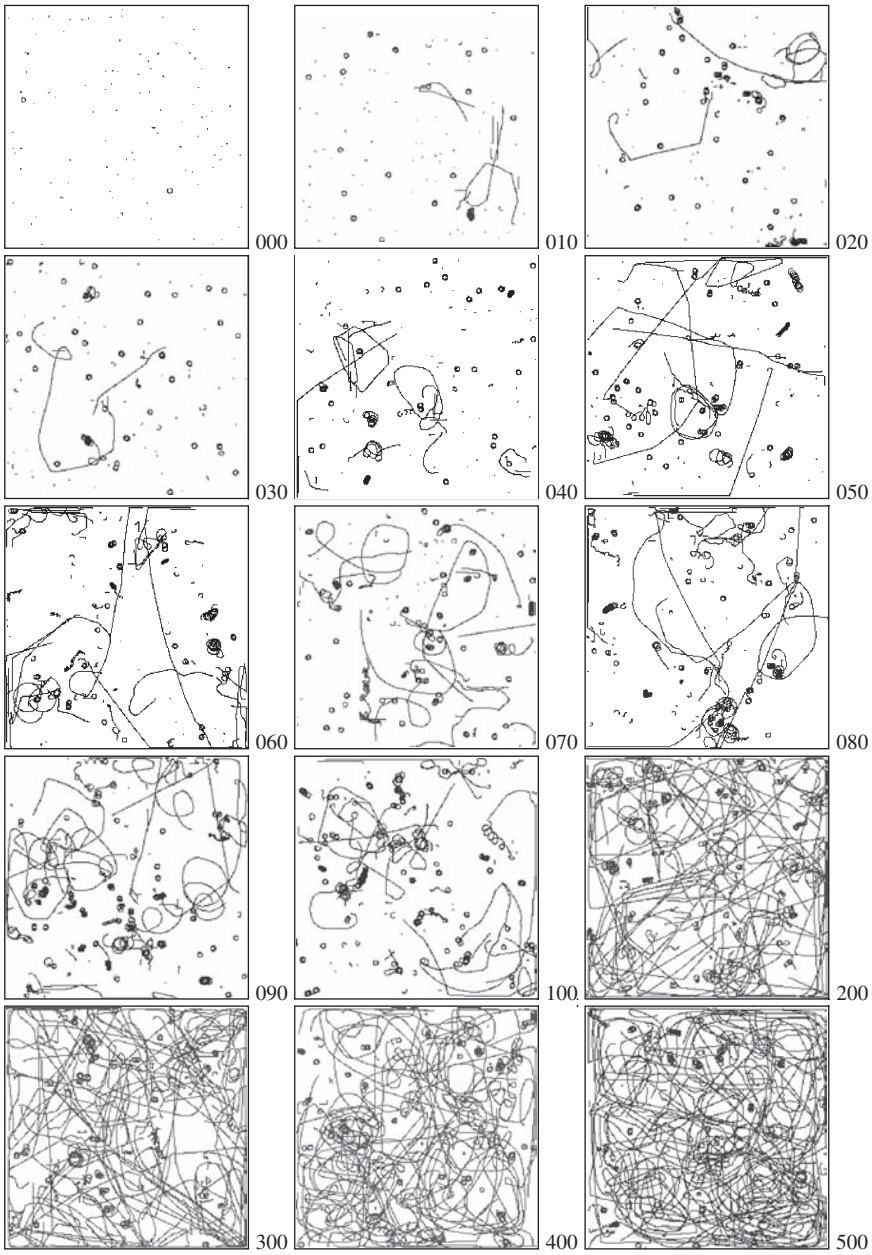


Figure 8: Arena area covered or 'trail patterns' from selected generations as the population evolves.

wandering behaviour. It is worth noting that these behaviours are both emergent as neither of the behaviours is explicitly mentioned by the fitness function.

If populations are allowed to interact then it is possible for far more complex, novel and seemingly intelligent behaviours to evolve. One example of this is the Cyclops fish simulator [16]. In this simulation there are two separate populations of simulated 'Cyclops' fish. The first

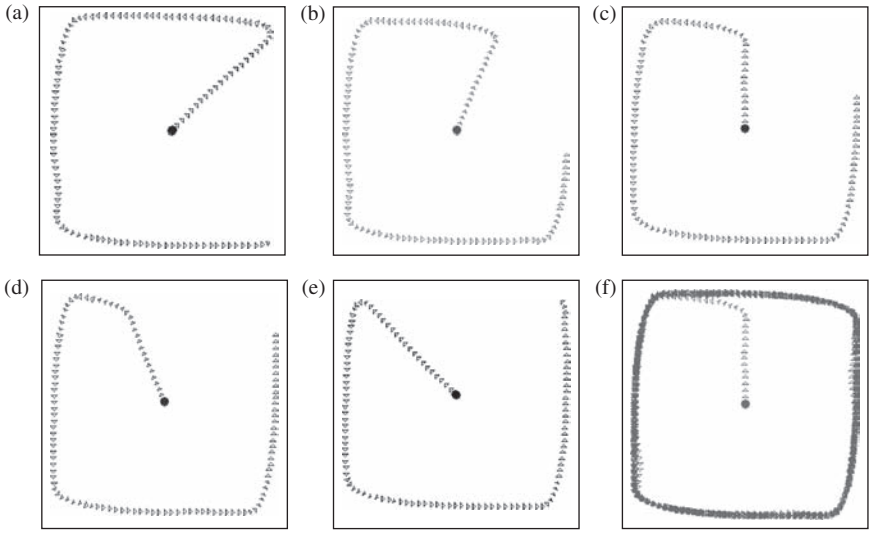


Figure 9: Individual path of the fittest member of the first species (long genome) in generation 500 plotted for various initial angles over 2 min. Angle of (a) 45°, (b) 67.5°, (c) 90°, (d) 112.5°, (e) 135° and (f) 90°, but for an extended run of 20 min.

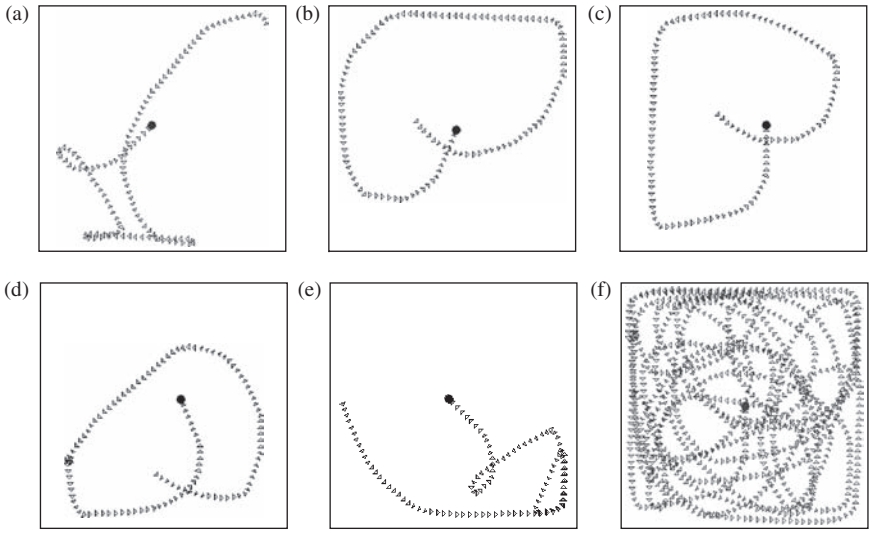


Figure 10: Individual path of the fittest member of the second species (short genome) in generation 500 plotted for various initial angles over 2 min. Angle of (a) 45°, (b) 67.5°, (c) 90°, (d) 112.5°, (e) 135° and (f) 90°, but for an extended run of 20 min.

population consists of foragers that are rewarded purely on the number of floating fish food pellets they are able to 'eat' in a specific period of time. The second population consists of predators that are rewarded purely on the number of foragers that they are able to 'eat' in the same time period. This is a co-evolutionary system in that if the predators get good at catching the foragers then there will be evolutionary pressure on the foragers to avoid the predators. Likewise if the foragers become expert at evading predation then there will be increased evolutionary pressure on the predators to try a new way to catch the foragers. Such systems can lead to an evolutionary 'arms race' and interesting behaviours.

In our own simulations the predators were able to evolve an extremely effective strategy for catching foragers within a few hundred generations. What the predators learnt to do was to seek out large clusters of food pellets and then stop, in so doing the predator is then obscured by the food pellets. The predator would then wait behind the food pellets until a forager came along, spotted the food pellets and swam straight into the predator's trap. Even in the simple environment of the fish tank simulator the evolved predator exhibits a very natural and seemingly 'cunning' solution – hide in the weeds until something comes along and then eat it. It is worth remembering that in our case these behaviours are purely the result of the two populations co-evolving, neither of the fitness functions explicitly mentions this behaviour and the resulting behaviour is truly emergent. In systems like this it is not always possible to know how a robot will behave in advance of evolving that behaviour.

6.4 Embodied evolution

Another intriguing possibility is that of embodied evolution [17] – a system in which the genetic algorithm is implemented by the interaction of the robots themselves. In such systems robots are able to exchange genetic material using a localised communication system such as a low range infrared system. In this system each robot possesses two sets of genetic material (see Fig. 11).

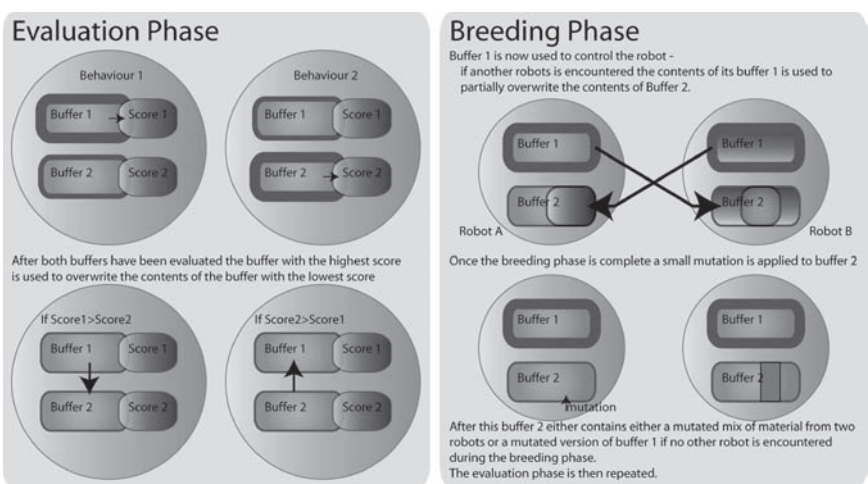


Figure 11: The embodied evolutionary algorithm. Each robot contains two genomes contained in two buffers which are evaluated in turn. If another robot is encountered, portions of their buffers can be exchanged and crossed over.

In an embodied evolutionary system, evolution takes place by means of the interaction of the robots themselves. No central computer is required in order to determine the fitness of individual robots or to hand out new sets of genomes for evaluation.

Each robot has two integral buffers, each containing a complete genome describing how to build a neural network that controls the robot. Evolution occurs in two discrete stages, the first is the evaluation phase, buffer 1 is used to create a neural network which is used to control the robot – the performance of the robot is evaluated over a fixed period of time and the buffer is assigned a score based on this performance. This evaluation process is then repeated for buffer 2 and a performance-based score is allocated. The buffer with the highest score is then used to replace the buffer with the lowest score. At this point the robot enters the second phase, the breeding phase. During this stage the neural network described by buffer 1 is used to control the robot. If during this phase another robot enters communication range a portion of both of the robots' buffer 1 is copied into buffer 2 of both robots. At the end of the breeding phase a small point mutation is then applied to buffer 2 and the evaluation phase is repeated. If no other robots are encountered during the breeding phase then the contents of buffer 2 will merely be a mutated version of buffer 1. Due to this, evolution may take place in either an asexual or sexual manner depending on the interaction of the robots themselves.

As in the previous case, if the evaluation function is implicit in nature – the behaviours evolved on the robots are not rigidly defined by the evaluation function. Instead the robots are free to explore a variety of solutions to a problem. Using such a system it is possible to evolve behaviour in real robots without the use of an overall supervisory computer. In this way the evolutionary process is completely decentralised and robots can be removed or added and evolution will still continue.

6.5 Initial results

A population of five simulated robots were allowed to evolve for 36 h (2160 min) – the robots were evaluated on their ability to reach a light source. As soon as a robot reaches and hits the light source, the light source is turned off and another light source elsewhere in the arena is turned on – Fig. 12a shows the evolution of light source hits over time whilst Fig. 12b shows the number of crossovers occurring (when the robots breed sexually).

It can be seen from Fig. 12 that at around 600 min the number of light hits per minute rapidly starts to increase, levelling off at around 1000 min. It is also at this point that the number of crossovers per minute rapidly increases enabling the light-seeking genes to gradually propagate throughout the population. This work is still in its initial stages and is not without its problems. However, it is a proof of concept of the viability of such techniques.

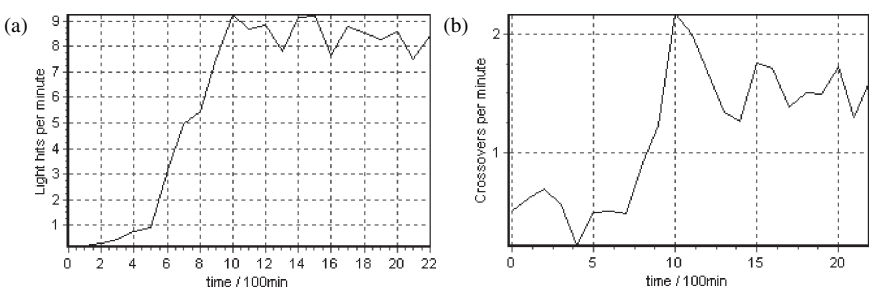


Figure 12: (a) Light hits per minute. (b) Number of crossovers per minute (mean).

7 Conclusion

The principal of emergence is widely used in the field of robotics and has already been employed extremely successfully. It is possible to produce robots that exhibit complex behaviour using only simple rules. When evolving robot behaviours with implicit fitness functions it is not always possible to know what the resultant behaviour will be without actually evolving the behaviour in the first place. The results of such an approach are indeed impressive with some behaviours seemingly being 'intelligent' or 'cunning'. The use of an embodied evolutionary algorithm allows the possibility of evolving emergent behaviours in an entirely decentralised and autonomous fashion.

One feature of behavioural emergence that can be regarded as a problem with scientific verification is the repeatability of results. By the nature of the problem part of the research interest is investigating how different emergent behaviours appear as a result of slightly different environmental conditions, variations in the selection mechanisms and differences between individuals, particularly in earlier generations. Whilst this does present a serious challenge if one research group wishes to repeat another research group's results, in that identical trial parameters are necessary, it does, at the same time, throw back important questions into the biological arena which in turn pose many questions towards our understanding of the evolution of life.

Of recent interest in the field is the concept of evolutionary convergence. What is meant by this is that certain behavioural end results and patterns emerge after several generations irrespective of initial conditions and even reproductive procedural modifications. Often the power of a selection method or fitness function can play a major role in this. It does however raise questions as to the potential in biological systems for a certain amount of determinism in the evolutionary process. With autonomous robot systems, an evolutionary scenario can be played out time and again, with the results, be they good or bad, witnessed first hand. If similar end results are encountered, the range of initial conditions in the experimental settings can be readily investigated.

References

[1] Gardner, M., The game of life, parts I–III (Chapters 20–22), *Wheels, Life, and other Mathematical Amusements*, W.H. Freeman: New York, 1983.
[2] Walter, W., An imitation of life. *Scientific American*, pp. 42–45, May 1950.
[3] Walter, W., A machine that learns. *Scientific American*, pp. 60–63, August 1951.
[4] Holland, O., Grey Walter: the pioneer of real artificial life. *Artificial Life V: Proceedings of the Fifth International Workshop on the Synthesis and Simulation of Living Systems*, ed. C. Langton, MIT Press, pp. 34–44, 1996.
[5] Braitenberg, V., *Vehicles: Experiments in Synthetic Biology,* Bradford Books: MIT Press: Cambridge, MA, 1984.
[6] Brooks, R., A robust layered control system for a mobile robot. *IEEE Journal of Robotics and Automation*, **2(1)**, pp. 14–23, 1986.
[7] Brooks, R., A robot that walks; emergent behaviour from a carefully evolved network. *IEEE International Conference on Robotics and Automation*, pp. 292–296, May 1989.
[8] Brooks, R., Intelligence without representation. *Artificial Intelligence*, **47**, pp. 139–159, 1991.
[9] Reynolds, C.W., Flocks, herds, and schools: a distributed behavioral model. *Computer Graphics*, **21(4)**, pp. 25–34, 1987.

[10] Huth, A. & Wissel, C., The simulation of the movement of fish schools. *Journal of Theoretical Biology*, **156**, pp. 365–385, 1991.
[11] Niwa, H., Self-organizing dynamic model of fish schooling. *Journal of Theoretical Biology*, **171**, pp. 123–136, 1994.
[12] Kelly, I., The development of shared experience learning in a group of mobile robots. PhD Thesis, Department of Cybernetics, The University of Reading, UK, 1997.
[13] Holland, J.H., *Adaptation in Natural and Artificial Systems*, University of Michigan Press: Ann Arbor, MI, 1975.
[14] Goldberg, D., *Genetic Algorithms in Search, Optimisation and Machine Learning*, Addison-Wesley, 1989.
[15] Hutt, B., Evolving artificial neural network controllers for robots using species based methods, PhD Thesis, Department of Cybernetics, The University of Reading, UK, 2002.
[16] Hutt, B. & Keating, D., Artificial evolution of visually guided foraging behaviour. *Proceedings of the Sixth International Conference on Artificial Life*, eds. C. Adami, R. Belew, H. Kitano & C. Taylor, MIT Press, pp. 393–397, 1998.
[17] Watson, R.A., Ficici, S.G. & Pollack, J.B., Embodied evolution. Poster presentation at The Fifth International Conference on Simulation of Adaptive Behaviour (SAB98), 1998.

Index

This page intentionally left blank

This page intentionally left blank

This page intentionally left blank

This page intentionally left blank

This page intentionally left blank

This page intentionally left blank

This page intentionally left blank

This page intentionally left blank

This page intentionally left blank