



**WE TRIP THE LIGHT
FANTASTIC**

VISIONS OF THE FUTURE

Chemistry and Life Science

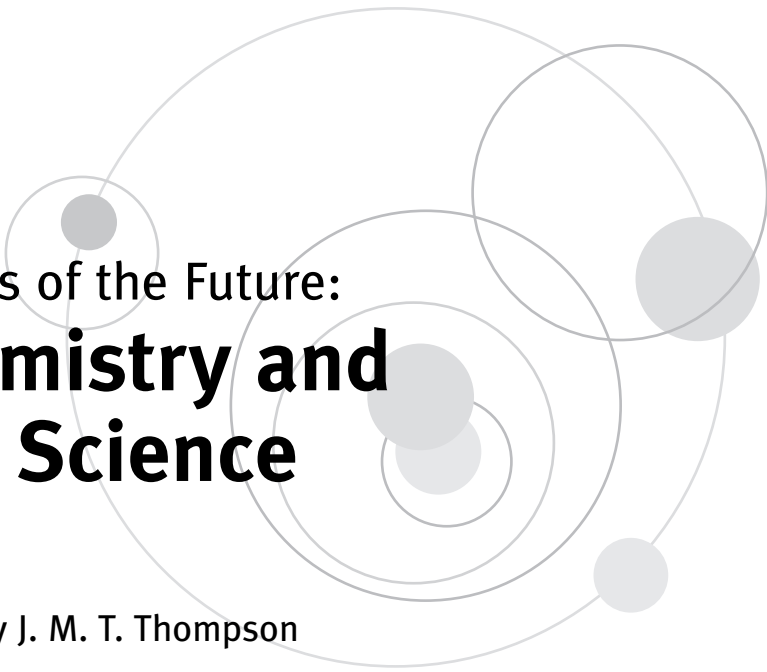
J. Michael T. Thompson

CAMBRIDGE

Visions of the Future: **Chemistry and Life Science**

Leading young scientists, many holding prestigious Royal Society Research Fellowships, describe their research and give their visions of the future. The articles, which have been re-written in a popular and well-illustrated style, are derived from scholarly and authoritative papers published in a special Millennium Issue of the Royal Society's *Philosophical Transactions* (used by Newton; this is the world's longest-running scientific journal). The topics, which were carefully selected by the journal's editor, Professor J. M. T. Thompson FRS, include studies of atoms and molecules in motion; new processes and materials; nature's secrets of biological growth and form; progress in understanding the human body and mind. The book conveys the excitement and enthusiasm of the young authors for their work in chemistry and life science. Two companion books cover astronomy and earth science, and physics and electronics. All are definitive reviews for anyone with a general interest in the future directions of science.

Michael Thompson is currently Editor of the Royal Society's *Philosophical Transactions* (Series A). He graduated from Cambridge with first class honours in Mechanical Sciences in 1958, and obtained his PhD in 1962 and his ScD in 1977. He was a Fulbright researcher in aeronautics at Stanford University, and joined University College London (UCL) in 1964. He has published four books on instabilities, bifurcations, catastrophe theory and chaos, and was appointed professor at UCL in 1977. Michael Thompson was elected FRS in 1985 and was awarded the Ewing Medal of the Institution of Civil Engineers. He was a senior SERC fellow and served on the IMA Council. In 1991 he was appointed director of the Centre for Nonlinear Dynamics.



Visions of the Future:
**Chemistry and
Life Science**

Edited by J. M. T. Thompson



CAMBRIDGE
UNIVERSITY PRESS

published by the press syndicate of the university of cambridge
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

cambridge university press
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011-4211, USA
10 Stamford Road, Oakleigh, VIC 3166, Australia
Ruiz de Alarcón 13, 28014 Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

<http://www.cambridge.org>

© The Royal Society 2001

This book is in copyright. Subject to statutory exception
and to the provisions of relevant collective licensing agreements,
no reproduction of any part may take place without
the written permission of Cambridge University Press.

First published 2001

Printed in the United Kingdom at the University Press, Cambridge

Typeface Trump Mediaeval 9/13 pt. *System* QuarkXPress™ [se]

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

Library of Congress Cataloguing in Publication data

Visions of the future : chemistry and life science / edited by J. M. T. Thompson
p. cm

Includes index.

ISBN 0 521 80539 2 (pb.)

1. Chemistry. 2. Life sciences. I. Thompson, J. M. T.

QD39.V53 2001

540-dc21 00-053007

ISBN 0 521 80539 2 paperback

Contents

Preface by J. M. T. Thompson

page vii

Atoms and molecules in motion

- | | | |
|---|--|----|
| 1 | Laser snapshots of molecular motions | 1 |
| | Gareth Roberts | |
| 2 | Enzymology takes a quantum leap forward | 21 |
| | Michael J. Sutcliffe and Nigel S. Scrutton | |

New processes and materials

- | | | |
|---|--|----|
| 3 | World champion chemists: people versus computers | 43 |
| | Jonathan M. Goodman | |
| 4 | Chemistry on the inside: green chemistry in mesoporous materials | 59 |
| | Duncan J. Macquarrie | |
| 5 | Diamond thin films: a twenty-first century material | 75 |
| | Paul W. May | |

Biological growth and form

- | | | |
|---|---|-----|
| 6 | The secret of Nature's microscopic patterns | 95 |
| | Alan R. Hemsley and Peter C. Griffiths | |
| 7 | Skeletal structure: synthesis of mechanics and cell biology | 113 |
| | Marjolein C. H. van der Meulen and Patrick J. Prendergast | |

Understanding the human body

- 8 The making of the virtual heart 127
Peter Kohl, Denis Noble, Raimond L. Winslow and Peter Hunter
- 9 Exploring human organs with computers 151
Paul J. Kolston

Understanding the human mind

- 10 Reverse engineering the human mind 171
Vincent Walsh
- Contributor biographies* 183
- Index* 197

Preface

Writing here in a popular and well illustrated style, leading young scientists describe their research and give their visions of future developments. The book conveys the excitement and enthusiasm of the young authors. It offers definitive reviews for people with a general interest in the future directions of science, ranging from researchers to scientifically minded school children.

All the contributions are popular presentations based on scholarly and authoritative papers that the authors published in three special Millennium Issues of the Royal Society's *Philosophical Transactions*. This has the prestige of being the world's longest running scientific journal. Founded in 1665, it has been publishing cutting-edge science for one third of a millennium. It was used by Isaac Newton to launch his scientific career in 1672 with his first paper 'New Theory about Light and Colours'. Under Newton's Presidency, from 1703 to his death in 1727, the reputation of the Royal Society was firmly established among the scholars of Europe, and today it is the UK's academy of science. Many of the authors are supported financially by the Society under its prestigious Research Fellowships scheme.

Series A of the *Philosophical Transactions* is devoted to the whole of physical science, and as its Editor I made a careful selection of material to cover subjects that are growing rapidly, and likely to be of long-term interest and significance. Each contribution describes some recent cutting-edge research, as well as putting it in its wider context, and looking forward to future developments. The collection gives a unique snapshot of the state of physical science at the turn of the millennium, while CVs and photographs of the authors give a personal perspective.

The three Millennium Issues of the journal have been distilled into three corresponding books by Cambridge University Press. These cover

astronomy and earth science (covering creation of the universe according to the big bang theory, human exploration of the solar system, Earth's deep interior, global warming and climate change), physics and electronics (covering quantum and gravitational physics, electronics, advanced computing and telecommunications), and chemistry and life science (covering the topics described below).

Topics in the present book on chemistry and life science include studies of atoms and molecules in motion, the development of new processes and materials, nature's secrets of biological growth and form, physical techniques in biology, progress in understanding the human body and mind, and the computer modelling of the human heart.

J. M. T. Thompson



1

Laser snapshots of molecular motions

Gareth Roberts

*Department of Chemistry, University of Cambridge, Lensfield Road,
Cambridge CB2 1EW, UK*

1.1 Introduction

The molecular motions that drive the conversion of energy and matter in physics, chemistry and biology take place over an amazingly rapid time, measured in millionths of a billionth of a second (a femtosecond (fs)). On time scales of this duration, atoms and molecules come together, exchange energy and transfer atoms in the very act of transforming one material into another. To map out such processes as they happen necessitates the application of laser pulses with durations of tens, or at most hundreds, of femtoseconds to take ‘snapshots’ of the changes in real time.

This chapter discusses the application of femtosecond lasers to the study of the dynamics of molecular motion, and attempts to portray how a synergic combination of theory and experiment enables the interaction of matter with extremely short bursts of light, and the ultrafast processes that subsequently occur, to be understood in terms of fundamental quantum theory. This is illustrated through consideration of a hierarchy of laser-induced events in molecules in the gas phase and in clusters. A speculative conclusion forecasts developments in new laser techniques, highlighting how the exploitation of ever shorter laser pulses would permit the study and possible manipulation of the nuclear and electronic dynamics in molecules.

1.2 The interaction of intense femtosecond laser light with molecules

The interaction of femtosecond laser light with atoms and molecules is completely different to that involving longer laser pulses. This arises from the ultrashort duration of femtosecond laser pulses, which is faster than the characteristic dynamical time scales of atomic motion, and their ultra-high intensity, which initiates a whole range of unprecedented phenomena. What exactly happens when an intense, ultrafast laser beam irradiates a sample of molecules depends crucially on the intensity of the laser, which determines the number of photons supplied to an individual molecule and can contort the allowed energy levels of the molecule; also important is the frequency of the laser, which, together with the intensity, affords optical access to different molecular energy states. The detailed physics of the light–matter interaction will of course also depend on the structure of the irradiated molecule, but whatever its identity, certain general features of the excitation of atoms and molecules by ultrafast laser photons have emerged from pioneering studies by research groups throughout the world.

First to respond to the laser field are the lighter electrons, which do so on a time scale of attoseconds (a thousandth of a femtosecond): depending upon the intensity of the incident light, the one or more photons absorbed by the molecule either promote an electron to a high-lying energy state of the molecule, or the electron is removed from the molecule altogether, leaving a positively charged ion; at very high intensities multiple electron excitation and ionisation through various mechanisms can occur. Over a far longer time scale of tens or hundreds of femtoseconds, the positions of the atomic nuclei within the molecule rearrange to accommodate the new electrostatic interactions suddenly generated as a result of the new electronic state occupancy prepared by the ultrafast laser pulse: the nuclear motions may involve vibrations and rotations of the molecule, or the molecule may fall apart if the nascent forces acting on the atoms are too great to maintain the initial structural configuration. In addition, at high incident intensities, the electric field associated with the laser beam distorts the electrostatic forces that bind the electrons and nuclei in a molecule to such an extent that the characteristic energy levels of the molecule are modified during the ultrashort duration of the laser pulse.

Each of the above phenomena is the subject of intensive research pro-

grammes in its own right. Figure 1.1 offers a simplified portrayal of some of these events, showing the ionisation of an electron from the warped potential energy structure an atom by an intense laser pulse, the path subsequently followed by the electron in response to the oscillating electric field of the laser pulse, and the emission of a high-frequency harmonic photon which occurs when the electron scatters off the ion core (high-harmonic emission can be exploited to generate attosecond laser pulses, discussed in Section 1.4.1). A similar series of events, with due alteration of the details, occurs in molecules exposed to intense laser light.

From careful measurements of such process, it is possible to develop quantitative models to describe the molecular dynamical response to impulsive laser excitation. These enable the fundamental interaction of intense, ultrafast laser light with molecules to be understood from first

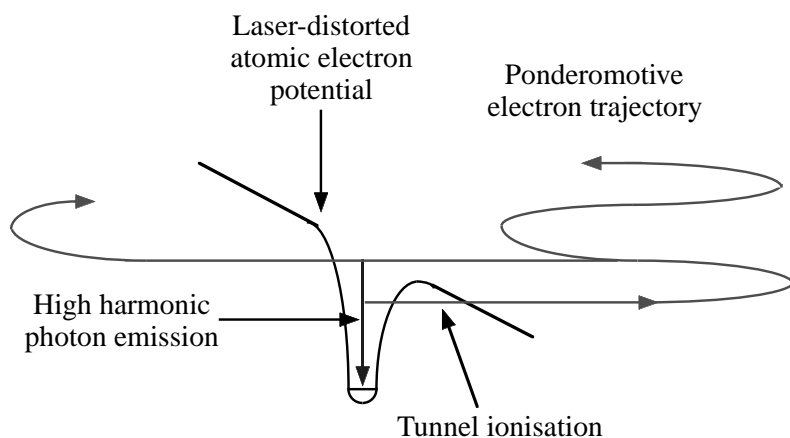


Figure 1.1. A sequence of events following the interaction of an intense, ultrafast laser pulse with an atom. The potential energy structure of the electron, which would otherwise be symmetric either side of a minimum, thereby confining the electron to the vicinity of the atomic nucleus, is distorted by the incident laser radiation. The electron first escapes (ionises) from the atom by tunnelling through the barrier on the side of lower potential energy and then executes an oscillatory trajectory determined by its kinetic (or ponderomotive) energy in the electric field of the laser pulse. If the electron follows a trajectory that brings it back close to the nucleus of the ionised atom, emission of a high-frequency photon can occur as the negatively charged electron is accelerated by the positively charged ion. This high-frequency photon is found to be an exact harmonic (overtone) of the laser frequency originally absorbed by the atom.

principles in terms of the wave description of matter and light due to quantum theory.

Following a description of femtosecond lasers, the remainder of this chapter concentrates on the nuclear dynamics of molecules exposed to ultrafast laser radiation rather than electronic effects, in order to try to understand how molecules fragment and collide on a femtosecond time scale. Of special interest in molecular physics are the critical, intermediate stages of the overall time evolution, where the rapidly changing forces within ephemeral molecular configurations govern the flow of energy and matter.

1.3 Femtosecond lasers

To carry out a spectroscopy, that is the structural and dynamical determination, of elementary processes in real time at a molecular level necessitates the application of laser pulses with durations of tens, or at most hundreds, of femtoseconds to resolve in time the molecular motions. Sub-100fs laser pulses were realised for the first time from a colliding-pulse mode-locked dye laser in the early 1980s at AT&T Bell Laboratories by Shank and coworkers: by 1987 these researchers had succeeded in producing record-breaking pulses as short as 6fs by optical pulse compression of the output of mode-locked dye laser. In the decade since 1987 there has only been a slight improvement in the minimum possible pulse width, but there have been truly major developments in the ease of generating and characterising ultrashort laser pulses.

The major technical driving force behind this progress was the discovery by Sibbett and coworkers in 1990 of a new category of ultrafast laser operation in solid-state materials, the most important of which is sapphire impregnated with titanium (others are the chromium-doped colquiriite minerals). These devices rely upon the intensity dependence of the refractive index of the gain medium to generate powerful, ultrashort laser pulses in a single 'locked' mode: a photograph of a commercial titanium:sapphire laser is shown in Figure 1.2.

Titanium:sapphire lasers typically deliver pulses with durations between 4.5 and 100 fs, and can achieve a peak power of some 0.8 watts, but this is not high enough to obtain adequate signal-to-noise ratio in experiments where the number of molecules that absorb light is low. To overcome this limitation, the peak power of a femtosecond laser can be dra-

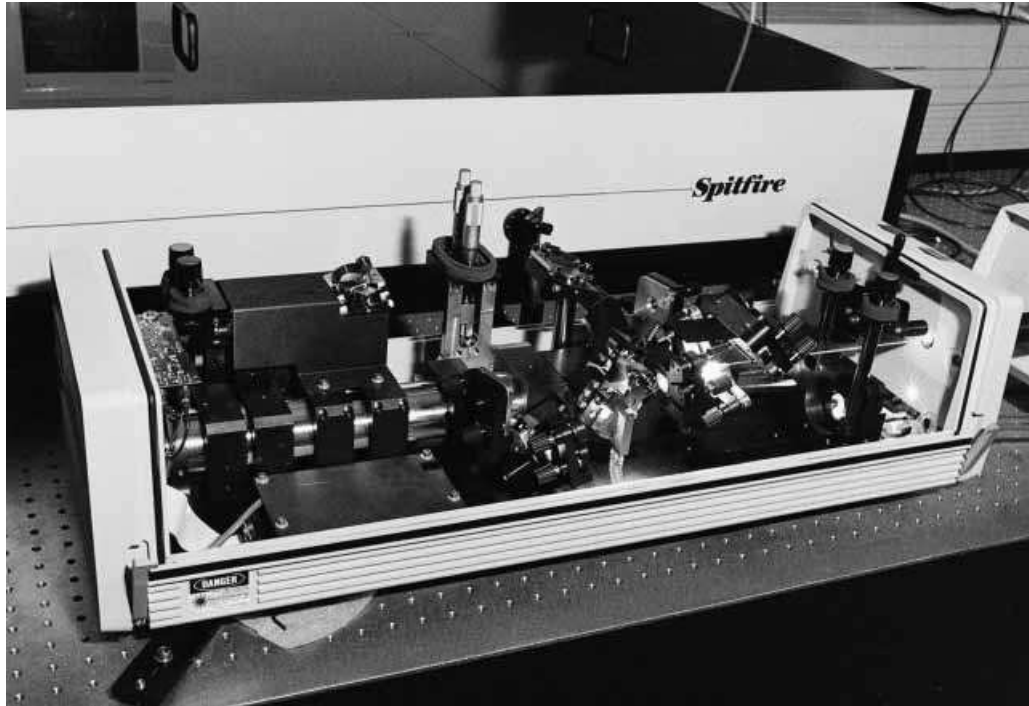


Figure 1.2. Photograph of a *Tsunami* titanium:sapphire laser manufactured by Spectra-Physics Lasers Inc. (Mountain View, California). The lasing transition in Ti:sapphire is between vibrational levels of different electronic states of the Ti^{3+} ion. Mode-locking of the laser is induced by an acousto-optic modulator, which results in the propagation of pulses with high peak powers and femtosecond durations in a single, 'locked' mode, or standing wave pattern. The energy source required to drive a Ti:sapphire laser is provided either by a diode or an argon-ion laser, both of which lase at the green wavelengths where Ti^{3+} is strongly absorbing. When the population of the Ti^{3+} excited state exceeds that of the ground state, laser radiation is emitted at red and near infrared wavelengths between 670 and 1070 nm.

matically increased by the process of chirped-pulse amplification. In this technique, the weakly intense ultrafast pulses are first stretched in time to between 100 and 1000 ps (a picosecond (ps) is 1000 fs), then amplified by about a million times in one or more further Ti:sapphire laser crystals, and finally recompressed to femtosecond durations. A typical peak power achievable with an amplified Ti:sapphire laser today is a hundred billion watts for a laser beam area of one square centimetre (the highest is just over a thousand million billion watts per square centimetre), which contrasts with an incident power of about 0.001 watts received through the iris of a human eye looking directly into the sun. For further details concerning the physics which underpins the operation of ultrafast lasers and their amplification, the interested reader is referred elsewhere for information (see Further reading).

For studies in molecular physics, several characteristics of ultrafast laser pulses are of crucial importance. A fundamental consequence of the short duration of femtosecond laser pulses is that they are not truly monochromatic. This is usually considered one of the defining characteristics of laser radiation, but it is only true for laser radiation with pulse durations of a nanosecond (0.000 000 001 s, or a million femtoseconds) or longer. Because the duration of a femtosecond pulse is so precisely known, the time-energy uncertainty principle of quantum mechanics imposes an inherent imprecision in its frequency, or colour. Femtosecond pulses must also be coherent, that is the peaks of the waves at different frequencies must come into periodic alignment to construct the overall pulse shape and intensity. The result is that femtosecond laser pulses are built from a range of frequencies: the shorter the pulse, the greater the number of frequencies that it supports, and *vice versa*.

The second requirement for investigations in ultrafast photophysics is one of wide wavelength coverage. The capacity for wavelength tuning is an essential ingredient in studies of molecular dynamics due to the different energy gaps that separate the quantum levels of molecules: vibrational resonances are excited with infrared light for example, whilst electronic states that correspond to different arrangements of the molecular electrons are reached by light in the visible and ultraviolet spectrum. The high output power of chirped-pulse amplified femtosecond lasers renders them ideal for synchronous pumping of optical parametric devices, whereby photons of light at one frequency are converted through their self-interactions in non-centrosymmetric media into photons at different frequencies. Today, the

application of such schemes offers continuous tunability from the near ultraviolet, through the visible, into the infrared regions of the spectrum.

An important point is that these advances have been complemented by the concomitant development of innovative pulse-characterisation procedures such that all the features of femtosecond optical pulses – their energy, shape, duration and phase – can be subject to quantitative *in situ* scrutiny during the course of experiments. Taken together, these resources enable femtosecond lasers to be applied to a whole range of ultrafast processes, from the various stages of plasma formation and nuclear fusion, through molecular fragmentation and collision processes to the crucial, individual events of photosynthesis.

1.4 Femtosecond spectroscopy of molecular dynamics

1.4.1 Ultrafast molecular fragmentation

To determine molecular motions in real time necessitates the application of a time-ordered sequence of (at least) two ultrafast laser pulses to a molecular sample: the first pulse provides the starting trigger to initiate a particular process, the break-up of a molecule, for example; whilst the second pulse, time-delayed with respect to the first, probes the molecular evolution as a function of time. For isolated molecules in the gas phase, this approach was pioneered by the 1999 Nobel Laureate, A. H. Zewail of the California Institute of Technology. The nature of what is involved is most readily appreciated through an application, illustrated here for the photofragmentation of iodine bromide (IBr).

The forces between atoms in a molecule are most conveniently represented by a surface of potential energy plotted as a function of the interatomic dimensions measured in ångströms (\AA) (10\AA are equivalent to a millionth of a millimetre). For the IBr molecule in the gas phase, the electronic ground state in which the molecule resides at equilibrium is characterized by a bound potential energy curve, labelled V_0 in Figure 1.3. The dissociative process is governed by two, interacting potential energy curves V_1 and V_1' for different excited states, which enable the molecule to break up along a coordinate leading to ground-state atoms (I+Br) or along a higher energy route which leads to excited bromine (I+Br*). Typical separation velocities are in the range $1500\text{--}2500\text{ms}^{-1}$. The same figure illustrates how femtosecond laser pulses configured in a pump-probe sequence can be applied to monitor the time-evolution of the photodissociation.

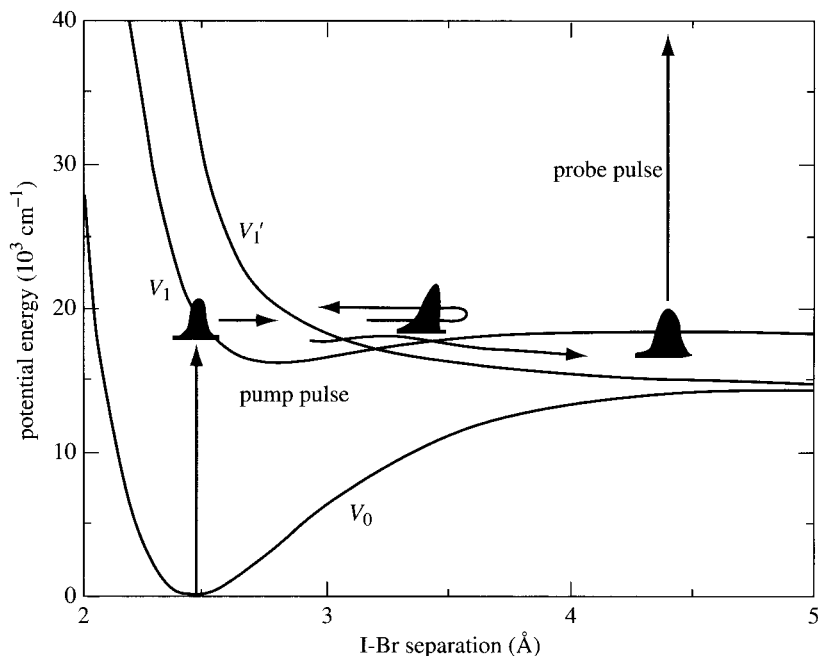


Figure 1.3. Real-time femtosecond spectroscopy of molecules can be described in terms of optical transitions excited by ultrafast laser pulses between potential energy curves which indicate how different energy states of a molecule vary with interatomic distances. The example shown here is for the dissociation of iodine bromide (IBr). An initial pump laser excites a vertical transition from the potential curve of the lowest (ground) electronic state V_0 to an excited state V_1 . The fragmentation of IBr to form I + Br is described by quantum theory in terms of a wavepacket which either oscillates between the extremes of V_1 or crosses over onto the steeply repulsive potential V_1' leading to dissociation, as indicated by the two arrows. These motions are monitored in the time domain by simultaneous absorption of two probe-pulse photons which, in this case, ionise the dissociating molecule.

An initial, ultrafast ‘pump’ pulse promotes IBr to the potential energy curve V_1 , where the electrostatic nuclear and electronic forces within the incipient excited IBr* molecule act to force the I and Br atoms apart. V_1 contains a minimum, however, so as the atoms begin to separate the molecule remains trapped in the excited state unless it can cross over onto the repulsive potential V_1' , which intersects the bound curve at an extended

I–Br bond length. Quantum theory does in fact allow such a curve-crossing to occur, with a probability that depends on, amongst other things, the velocity of the escaping atoms, the exact shape of the intersecting potentials at their crossing point, and the spacing of vibrational quantum levels available to the excited molecule in its quasi-bound state.

From a theoretical perspective, the object that is initially created in the excited state is a coherent superposition of all the wavefunctions encompassed by the broad frequency spread of the laser. Because the laser pulse is so short in comparison with the characteristic nuclear dynamical time scales of the motion, each excited wavefunction is prepared with a definite phase relation with respect to all the others in the superposition. It is this initial coherence and its rate of dissipation which determine all spectroscopic and collisional properties of the molecule as it evolves over a femtosecond time scale. For IBr, the nascent superposition state, or wavepacket, spreads and executes either periodic vibrational motion as it oscillates between the inner and outer turning points of the bound potential, or dissociates to form separated atoms, as indicated by the trajectories shown in Figure 1.3.

The time evolution of the wavepacket over the intersecting potentials V_1 and V'_1 is monitored by its interaction with a second ultrashort ‘probe’ pulse, which in this case supplies two ultraviolet photons to ionise the molecule by removal of an outer electron. The key experimental requirement in this and all other pump-probe measurements is the ability to deliver the two ultrafast laser pulses to the sample separately spaced by a controllable and accurately known difference in time. This is achieved in the laboratory by routing one of the pulses via an interferometric translation stage which can vary the path length between pump and probe pulses prior the sample with a precision of a fraction of a micrometre (μm) ($1 \mu\text{m}$ distance equates to about 3.33 fs in time). The experiment consists of measuring in a mass spectrometer the number of ionised IBr^{++} molecules excited by pump and probe pulses as function of the delay time between the two (see Figure 1.3), since this is directly proportional to the probability of locating the extended $[\text{I} \dots \text{Br}]$ molecule over different coordinates of the potential energy curves V_1 and V'_1 ; the probe pulse can be thought of as projecting onto the potentials a detection ‘window’, the width of which is determined by the spectral breadth, and hence duration, of the pulse, through which the dynamics of the dissociating molecule can be observed.

Figures 1.4(a) and (b) show examples of the ionisation signals that are

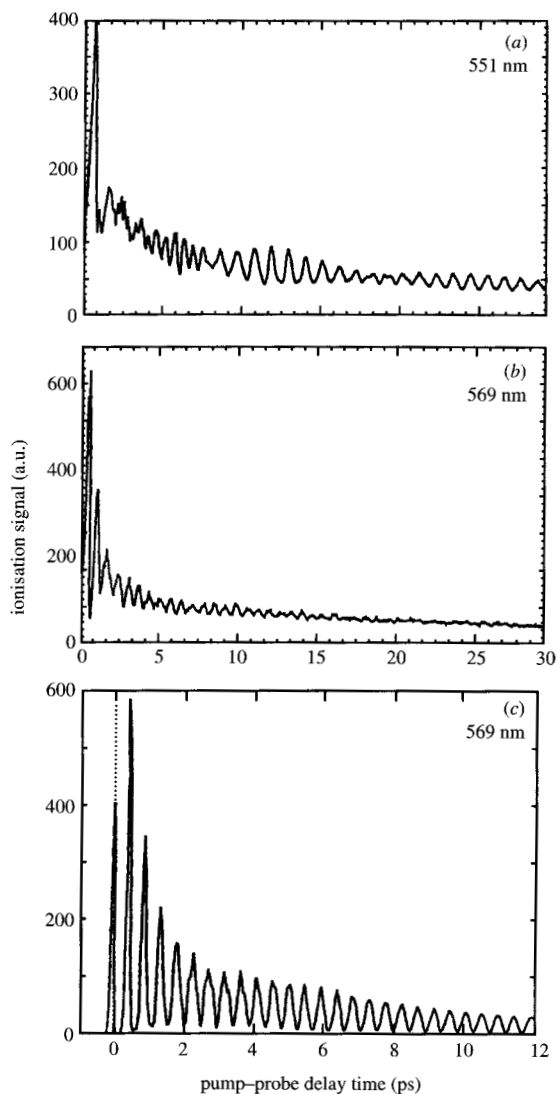


Figure 1.4. Experimental and theoretical femtosecond spectroscopy of IBr dissociation. Experimental ionisation signals as a function of pump-probe time delay for different pump wavelengths given in (a) and (b) show how the time required for decay of the initially excited molecule varies dramatically according to the initial vibrational energy that is deposited in the molecule by the pump laser. The calculated ionisation trace shown in (c) mimics the experimental result shown in (b).

recorded as a function of pump-probe time delay: the decrease in signal intensity with increasing pump-probe time delay monitors the loss of initial IBr^* to form separated I and Br over the potential V'_1 ; and the oscillations superimposed upon the decay reflect the quantized nature of vibrational motion of the quasi-bound $[\text{I} \cdots \text{Br}]$ molecules at intermediate configurations within the bound V_1 curve.

A series of measurements in which the pump wavelength is varied reveal that at some energies the oscillations predominate for times beyond 10ps, whilst at others the decay of population by curve-crossing wins out within 400fs or so. The time resolution of the experiment is in this example is determined by the convolution of the two laser pulse widths, here about 125fs.

These attributes can be accounted for by theoretical calculations of the motion of the wavepacket over the repulsive potential, which aim to determine the time-resolved ionisation signal from fundamental theory. These are performed by solving the time-dependent Schrödinger equation for the dissociation, which expresses the temporal development of the quantum wavefunction prepared by the laser pulse subject to all the forces that act on the nuclei as it progresses from starting to final states. Figure 1.4(c) displays a calculated pump-probe ionisation trace that corresponds to the same initial conditions of Figure 1.4(b). A mathematical analysis of these data using the technique of Fourier transformation reveals how quantised vibrational motion of the molecule along the dissociation coordinate is transformed into kinetic energy of separation as the I and Br atoms fly apart.

1.4.2 Ultrafast molecular collisions

Unfortunately, femtosecond laser pulses are not so readily predisposed to study collisions between atoms and molecules by the pump-probe approach. The reason is that, typically, the time between collisions in the gas phase is on the order of nanoseconds. So, with laser pulses of sub-100fs duration, there is only about one chance in ten thousand of an ultrashort laser pulse interacting with the colliding molecules at the instant when the transfer of atoms is taking place; in other words, it is not possible to perform an accurate determination of the zero of time.

An ingenious method to circumvent this problem was first devised by Zewail and colleagues, who took advantage of the vibrational and rotational cooling properties and collision-free conditions of the supersonic

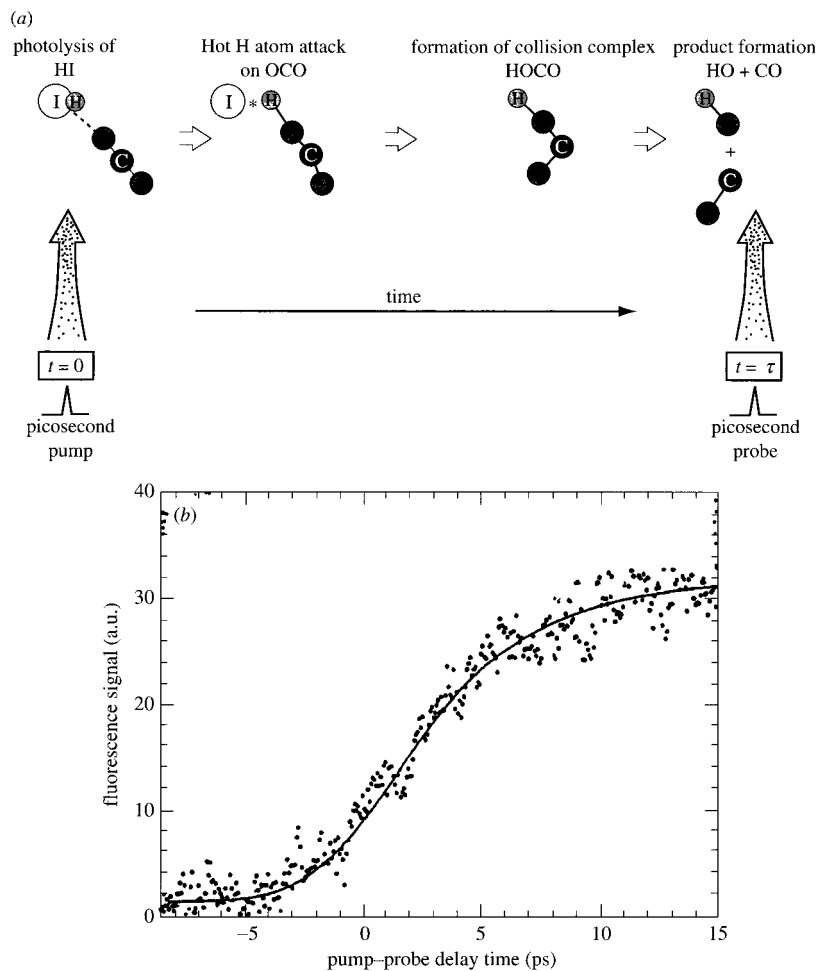


Figure 1.5. Femtosecond spectroscopy of bimolecular collisions. The cartoon shown in (a) illustrates how pump and probe pulses initiate and monitor the progress of $\text{H} + \text{CO}_2 \rightarrow [\text{HO} \dots \text{CO}] \rightarrow \text{OH} + \text{CO}$ collisions. The build-up of OH product is recorded via the intensity of fluorescence excited by the probe laser as a function of pump-probe time delay, as presented in (b). Potential energy curves governing the collision between excited Na^* atoms and H_2 are given in (c); these show how the $\text{Na}^* + \text{H}_2$ collision can proceed along two possible exit channels, leading either to formation of $\text{NaH} + \text{H}$ or to $\text{Na} + \text{H}_2$ by collisional energy exchange.

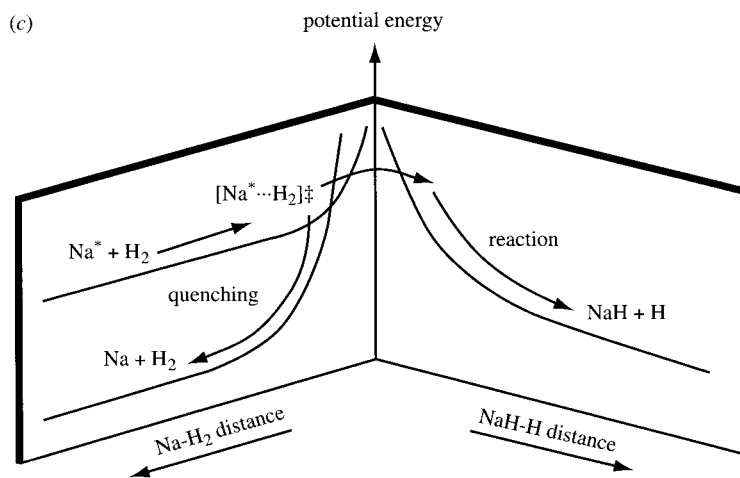


Figure 1.5. (cont.)

expansion of a jet of gas in a high-vacuum chamber – known as a molecular beam – to couple molecules closely together in a weakly bound, unreactive complex prior to femtosecond laser-initiation of the collisional trajectory. These workers chose to study collisions between hydrogen atoms H and carbon dioxide CO₂ in which the starting materials were prepared in a complex of hydrogen iodide and CO₂. A cartoon representation of the experiment performed by Zewail's group is given in Figure 1.5(a) and one of their many results is shown in Figure 1.5(b).

The wavelength of an ultrafast pump pulse is selected to trigger the reaction by cleaving the H-I bond; this liberates the H atom which translates towards CO₂ and, over the course of about 10ps, subsequently generates hydroxyl OH and carbon monoxide CO. Product formation is monitored by the detection of fluorescence from OH induced by a time-delayed probe laser pulse. In this example, the collision takes a long time to complete because H and CO₂ initially combine to form a metastable [HO . . . CO] intermediate, which survives long enough to redistribute energy amongst its available degrees of freedom until such time as sufficient vibrational motion with the correct energy and phase is localised within the HO-CO mode. When this point is reached, the force of repulsion between OH and CO exceeds the attractive interactions between the two components and the diatomic moieties spin away from each other.

The use of molecular beams to lock reactants together within femtosecond striking distance is not the only way to perform ultrafast spectroscopy of bimolecular reactions. Another is to initiate the final approach trajectory of collision between a metastable atom or molecule in a high-pressure atmosphere of a second partner, thereby reducing the time required for the collisional encounter to below a picosecond. This approach is illustrated in Figure 1.5(c) for collisions between excited sodium atoms Na^* and molecular hydrogen, in which the outermost electron of the sodium is first promoted to an energised state by an ultrafast laser pulse. The $\text{Na}^* + \text{H}_2$ system serves as a paradigm for transfer of matter and energy in atom-molecule scattering since, as shown in Figure 1.5(c), the atom and molecule can either form $\text{NaH} + \text{H}$ by swapping a hydrogen atom or can transfer the initial excitation energy of the sodium atom to the intact H_2 molecule, resulting in the emergence of a deactivated sodium atom and vibrationally excited hydrogen. The trajectory of the scattering event again proceeds via one or more curve crossings between potential energy surfaces, representing the different forces between the atom and molecule at different stages of the collisional evolution.

Current research at the Max-Planck-Institut für Quantenoptik in Garching, Germany, is concentrating on the mechanism of collisional deactivation via electronic-to-vibrational energy transfer, in which the temporal progress from initial to final states is monitored by the simultaneous absorption of three 20fs probe photons and re-emission of a fourth by the $[\text{Na}^* \dots \text{H}_2]$ intermediate configuration as it forms and breaks apart. This type of coherent scattering spectroscopy is extremely sensitive and enables the appearance of deactivated sodium atoms to be probed as a function of time as they emerge from the curve crossing. Experimental measurements are supported by theoretical calculations of the cross sections for light scattering in real time, from which the wavepacket motion over the intersecting potential energy curves can be deduced. These reveal that the $[\text{Na}^* \dots \text{H}_2]$ species formed during the initial approach stage persists for durations up to 120fs before it fragments, during which time the excitation energy carried by the Na^* atom is funnelled into the H_2 coordinate by repeated multidimensional transfer of population between the colliding partners. The collision is said to be 'sticky', as the $\text{Na}^* + \text{H}_2$ collide, bounce off one another and exchange energy and population over a time scale that is very long compared to the period of H_2 vibrations (about 8fs).

1.4.3 Many-body effects on ultrafast dynamics

Over recent years, advances in high-vacuum technology and mass spectrometry have enabled experimentalists to prepare clusters of selected size and composition in the gas phase. A cluster is a smallish globule, comprising up to about 1000 atoms or molecules held together by weak attractive forces, that is supremely well-suited for the study of ultrafast phenomena in which many-body effects dominate the collisional outcome. The most important of these concern the fate of the energy initially deposited in the cluster by the laser pulse as a result of intra- and intermolecular energy redistribution, coherence loss of the nascent wavepacket and molecular fragmentation, and how these effects evolve with increasing degrees of freedom. A popular choice for investigation has been the dissociation of molecular halogens attached to one or more rare gas atoms.

A recent experimental study by the group of Neumark at UC Berkeley, USA, on the dissociation of the negatively charged diiodide (I_2^-) ion in the presence of zero, six or 20 argon atoms exemplifies marvellously the way in which the issues listed above can be successfully addressed by femtosecond spectroscopy. In these experiments, the dissociation of size-selected $I_2^- \cdot Ar_n$ clusters was triggered using 100 fs pulses from a Ti:sapphire laser and monitored by a second ultrafast pulse which detaches the excess electron from the negatively charged molecule. Measurements of the kinetic energy distribution of the photoejected electrons, called a photoelectron spectrum, as a function of pump-probe delay time turn out to be an extremely sensitive probe of the rapidly changing local environment of the detached electron, in that they reveal how the forces between the iodine atoms and between the I_2 molecule and its immediate surroundings evolve during the dissociative separation of the halogen atoms. The experiments show that, whereas in the absence of argon atoms the break-up of diiodide to I and I^- evolves over a time scale of 250 fs, it is effectively stopped and returned to near its starting position when 20 argon atoms form a shell around the dissociating molecule; subsequent to the caging process, vibrational cooling of the I_2^- molecule thereby regenerated takes an amazingly long 200 ps to complete!

Experiments such as these provide an incomparable level of detail on the temporal ordering of elementary processes in a multidimensional collisional environment. To understand the dynamical evolution of many-body systems in terms of the changing forces that act on the interacting

atoms requires sophisticated computer simulations to map out the motions of the individual atoms and to elucidate the structures of the transient molecular configurations that control the flow of energy between atoms and molecules over a femtosecond time scale. For clusters containing, say, a diatomic molecule bound to one or two atoms, with computational facilities available today it is possible to carry out calculations in which the dissociative evolution along every degree of freedom is treated by quantum dynamics theory.

An example of this type of calculation is shown in Figure 1.6, which portrays a snapshot of the wavepacket motion of iodine bromide attached to Ar initiated by a 100fs laser pulse. The early-time (≤ 150 fs) motions of the complex, which is almost T-shaped, comprise a simultaneous lengthening of the I–Br distance and a slower transfer of vibrational energy from the intramolecular mode to the IBr–Ar coordinate. Just as was found for the isolated IBr molecule (Section 1.4.1), a fraction of the wavepacket amplitude along the I–Br direction proceeds to dissociation by curve crossing whilst the remainder becomes trapped in the quasi-bound potential well. By 840fs, bursts of vibrational energy transfer to the atom–molecule degree of freedom give rise to a stream of population which eventually leads to expulsion of argon from the complex. To connect this dynamical picture with information available from experiments, calculations of the vibrational spectra of the cluster as a function of time after the femtosecond pump pulse show that relaxation of the nascent IBr vibrational content is at first sequential but at times longer than about 500fs becomes quasi-continuous as a result of a complex interplay between intermode vibrational energy redistribution and molecular dissociation.

1.5 What else and what next? A speculative prognosis

Ultrafast laser spectroscopy is very much a science that, by its very nature, is driven by improvements in laser and optical technology. Dangerous though it is to make forecasts of scientific advances, what is clear at the time of writing (early 2000) is that at the cutting edge of this research field is the progress towards even faster laser pulses and the ability to design femtosecond laser pulses of a specified shape for optical control of individual molecular motions.

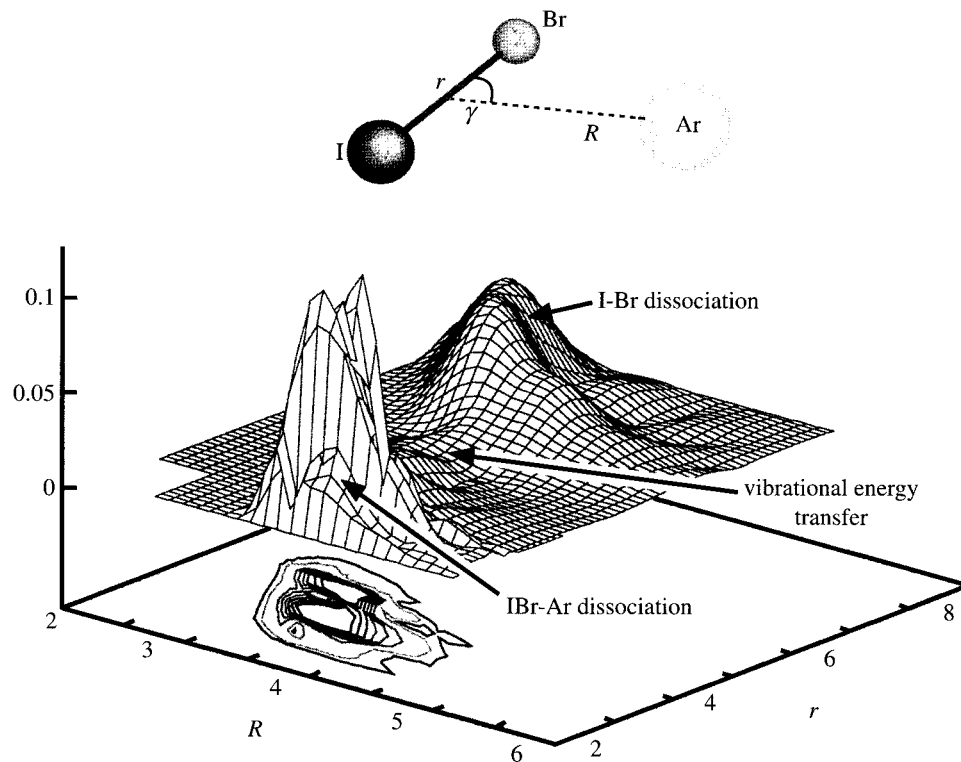


Figure 1.6. Quantum theory of IBr-Ar dissociation, showing a snapshot of the wavepacket states at 840fs after excitation of the I-Br mode by a 100fs laser pulse. The wavepacket maximum reveals predominant fragmentation of the IBr molecule along the r coordinate at short IBr-Ar distances (R coordinate), whilst a tail of amplitude stretches to longer R coordinates, indicating transfer of energy from the I-Br vibration to the IBr-Ar dimension, which propels the argon atom away from the intact IBr molecule.

1.5.1 Attosecond laser pulses

Of course, even when the world's fastest laser pulses are available, there is always a feeling that what is really required is pulses that are faster still! Laser pulses with durations in the attosecond regime would open up the possibility of observing the motions of electrons in atoms and molecules on their natural time scale and would enable phenomena such as atomic and molecular ionisation (Section 1.2) and the dynamics of electron orbits about nuclei to be captured in real time.

There are several proposals actively being pursued around the world to generate laser pulses that are significantly shorter than the shortest available today (the current world record is 4.5 fs). The physics of each scheme is well understood and the technology required to implement them in existence; what is tricky is that the proposals are not so easy to apply in the laboratory. To reach the attosecond regime, laser pulses must be composed of very many different frequencies, as required by the time–energy uncertainty principle, and they must be coherent. A usable source of attosecond pulses must also be intense enough to result in experimentally detectable changes in light absorption or emission, and they must be separated in time by at least one millionth of a second so that the changes they induce can be recorded by modern electronic circuitry.

One scheme which has generated considerable optimism is that suggested by Corkum and colleagues at the National Research Council in Ottawa, Canada, which takes advantage of the high harmonic frequencies simultaneously generated when an intense femtosecond laser pulse ionises a gas of helium or neon in a narrow waveguide to construct the broad spectrum of colours necessary to support attosecond laser emission. These harmonics are just like the overtones of a musical note: they are generated by oscillations of the electrons liberated by ionisation in the laser field and are formed coherently, that is with their amplitudes in phase with one another. Figure 1.1 presents a schematic illustration of the mechanism by which a high-harmonic photon is emitted in an atom. At the present time researchers have succeeded in generating up to the 297th harmonic in helium of the original 800 nm light from a 25 fs titanium:sapphire laser by this approach, yielding a harmonic spectrum which extends into the X-ray region as far as 2.7 nm, and current research focusses on exploiting this broadband emission to construct a usable attosecond laser. In addition to providing a possible source of attosecond light, high-order harmonic gen-

eration also offers the chance to develop coherent, ultrafast X-ray laser devices.

1.5.2 Coherent control of molecular dynamics

When it was invented in 1960, the laser was considered by many to be the ideal tool for controlling the dynamics of molecular dissociation and collisions at the molecular level. The reasoning was that by choosing the frequency of a monochromatic (long pulse duration) laser to match exactly that of a local vibrational mode between atoms in a polyatomic molecule, it ought to be possible to deposit sufficient energy in the mode in question to bring about a massively enhanced collision probability, and thereby generate a selected set of target states. With the benefit of hindsight, it is clear that the approach failed to take into account the immediate and rapid loss of mode specificity due to intramolecular redistribution of energy over a femtosecond time scale, as described above.

Eight years ago it was suggested by US researchers that in order to arrive at a particular molecular destination state, the electric field associated with an ultrafast laser pulse could be specially designed to guide a molecule during a collision at different points along its trajectory in such a way that the amplitudes of all possible pathways added up coherently just along one, specific pathway at successive times after the initial photoabsorption event. To calculate the optimal pulse shapes required by this scheme dictates the use of a so-called 'evolutionary' or 'genetic' computer algorithm to optimise, by natural selection, the electric field pattern of the laser applied to the colliding molecule at successive stages, or 'generations', during its dynamical progress from the original progenitor state until the sought-after daughter state is maximally attained.

In order that this proposal can be made to work, what is required is a device which can make rapid changes to the temporal pattern of the electric field associated with a femtosecond laser pulse. The recent development of liquid-crystal spatial light modulators to act as pulse shapers fulfils this task, and may open the gateway to a plethora of experimental realisations of coherent control of molecular dynamics. There has been much theoretical work on the types of laser pulse shapes required to bring about specific molecular goals. In the laboratory, successful optical control of molecular events has been demonstrated for strategic positioning of wavepackets, enhancement of molecular ionisation probabilities, and optimisation of different photodissociation pathways. With the advent of

femtosecond laser technology, the potential for control of molecular collision dynamics with laser beams is becoming a reality.

1.6 Further reading

Kapteyn, H. & Murnane, M. 1999 *Phys. World*, 31.

Roberts, G. 2000 *Phil. Trans. R. Soc. Lond. A* **358**, 345.

Rullière, C. (ed.) 1998 *Femtosecond laser pulses: principles and experiments*. Berlin: Springer.

Suter, D. 1997 *The physics of laser-atom interactions*. Cambridge: Cambridge University Press.

Zewail, A. H. 1996 *Femtochemistry*, Vols 1 & 2. Singapore: World Scientific.



2

Enzymology takes a quantum leap forward

Michael J. Sutcliffe¹ and Nigel S. Scrutton²

¹ Department of Chemistry, University of Leicester, Leicester LE1 7RH, UK

² Department of Biochemistry, University of Leicester, Leicester LE1 7RH, UK

2.1 Introduction

Enzymes facilitate life via a plethora of reactions in living organisms. Not only do they sustain life – they are also involved in a myriad of processes that affect our everyday lives. These include applications in medicine, household detergents, fine chemical synthesis, the food industry, bioelectronics and the degradation of chemical waste. Since the discovery of enzymes just over a century ago, we have witnessed an explosion in our understanding of enzyme catalysis, leading to a more detailed appreciation of how they work. Over many years, much effort has been expended in the quest to create enzymes for specific biotechnological roles. Prior to the early 1980s, the only methods available for changing enzyme structure were those of chemical modification of functional groups (so-called ‘forced evolution’). The genetic engineering revolution has provided tools for dissecting enzyme structure and enabling design of novel function. Chemical methods have now been surpassed by knowledge-based (i.e. rational) site-directed mutagenesis and the grafting of biological function into existing enzyme molecules (so-called ‘retrofitting’). More recently, gene-shuffling techniques have been used to generate novel enzymes. Rational redesign of enzymes is a logical approach to producing better enzymes. However, with a few notable exceptions, rational approaches have been generally unsuccessful, reiterating our poor level of understanding of how enzymes work. This has led to a more ‘shot-gun’ approach to redesign, involving

random mutagenesis – producing modest success, but dependent on being able to ‘pull out’ an improved enzyme by ‘fishing’ in a very large collection of randomly modified enzymes. However, development of a suitable test (i.e. producing the correct ‘bait’) to identify an improved enzyme is intrinsically very difficult. Therefore the rational approach, although generally unsuccessful, cannot be ignored.

Enzymes are large biological molecules – usually proteins – that speed up chemical reactions. Molecules that speed up chemical reactions, but are unchanged afterwards, are known as catalysts. The substances that enzymes act on are known as substrates. Enzymes exhibit remarkable specificity for their substrate molecules, and can approach ‘catalytic perfection’. A popular approach to modelling catalysis has been to visualise an energy barrier that must be surmounted to proceed from reactants to products (Figure 2.1). The greater the height of this energy barrier, the slower the rate of reaction. Enzymes (like other catalysts) reduce the energy required to pass over this barrier, thereby increasing reaction rate. The structure of the reactant at the top of the barrier is energetically unstable, and is known as the ‘transition state’. The energy required to pass over the barrier is the so-called ‘activation energy’ – the barrier is surmounted by thermal excitation of the substrate. This classical over-the-barrier treatment – known as transition state theory – has been used to picture enzyme-catalysed reactions over the past 50 years. However, recent developments indicate that this ‘textbook’ illustration is fundamentally flawed (at least in some circumstances).

The transition state theory considers only the particle-like properties of matter. However, matter (especially those particles with smaller mass) can also be considered as having wave-like properties – this is known as the wave–particle duality of matter. For enzyme-catalysed reactions, an alternative picture to transition state theory has emerged from considering the wave–particle duality of matter. All matter exhibits both particle- and wave-like properties. Large ‘pieces’ of matter, like tennis balls, exhibit predominantly particle-like properties. Very small ‘pieces’ of matter, like photons (of which light is composed), whilst showing some particle-like properties exhibit mainly wave-like properties. One important feature of the wave-like properties of matter is that it can pass through regions that would be inaccessible if it were treated as a particle, i.e. the wave-like properties mean that matter can pass through regions where there is zero probability of finding it. This can be visualised, for example, by considering the

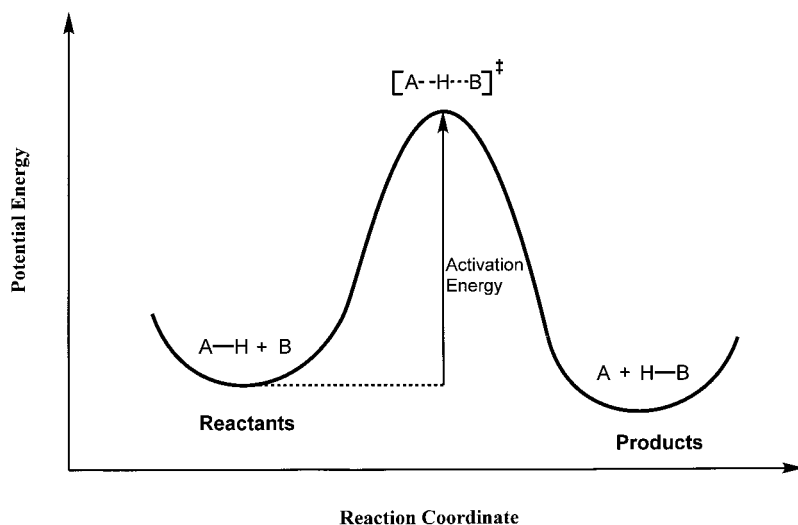


Figure 2.1. A popular approach to modelling catalysis has been to visualise an energy barrier that must be surmounted to proceed from reactants to products. This process is shown schematically. For the reaction to proceed, reactants ($A-H + B$) must pass over the potential energy barrier to the product ($A + H-B$) side via the so-called transition state (denoted by $[A \cdots H \cdots B]^\ddagger$) at the top of the energy barrier. This transition state is energetically unstable. The greater the height of this energy barrier, the slower the rate of reaction. Enzymes (like other catalysts) reduce the energy required to pass over this barrier, thereby increasing reaction rate. This classical over-the-barrier treatment – known as ‘transition state theory’ – has been used to picture enzyme-catalysed reactions over the past 50 years. However, recent developments indicate that this representation is, at least in some circumstances, fundamentally flawed and should instead be considered in terms of quantum tunnelling *through* the barrier.

vibration of a violin string – some parts of the string are stationary (known as nodes) and yet the vibration passes through these nodes (Figure 2.2). Thus, the pathway from reactants to products in an enzyme-catalysed reaction may not need to pass over the barrier, as in transition state theory with particle-like behaviour, but could pass through the barrier. This passing through the barrier (quantum tunnelling; Figure 2.3) can be likened to passing from one valley to an adjacent valley via a tunnel, rather than having to climb over the mountain between. As the analogy suggests, this can lower significantly the energy required to proceed from reactants to

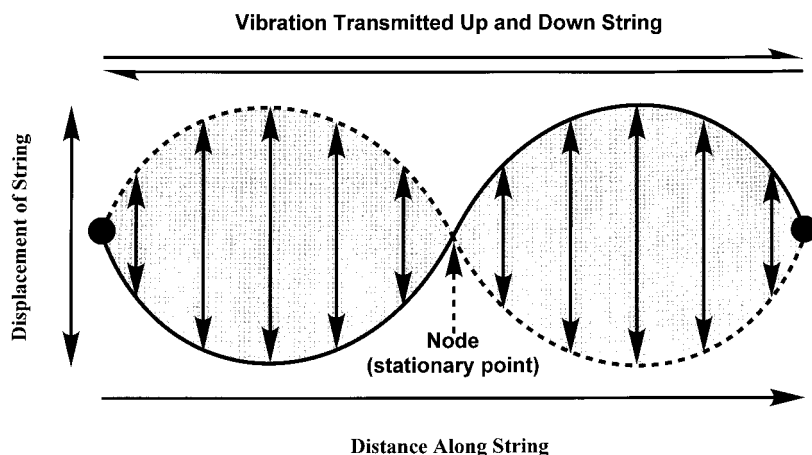


Figure 2.2. Illustration of the wave-like property of matter by analogy with the vibrations on a violin string. The solid and dashed lines illustrate the extremities of the vibration. Although there is a node (a position where the string is stationary) in the centre of the string, the vibration is transmitted through this node – this is analogous to passing through a region of zero probability as in quantum tunnelling.

products. Thus, quantum tunnelling may play an important role in driving enzyme-catalysed reactions, especially for the transfer of small nuclei such as hydrogen.

Indeed, quantum tunnelling is the established mechanism for enzyme-mediated transfer of the much smaller electron. Proteins are electrical insulators; nevertheless, electrons can travel large distances on the atomic scale (up to around 3×10^{-9} m) through them. This apparent paradox – of an electron passing through an electrical insulator – can be understood in terms of the wave-like properties of the electron. Thus, the electron can pass via quantum tunnelling through regions from which it would be excluded by its particle-like nature.

In contrast to electron transfer via quantum tunnelling, evidence for hydrogen tunnelling in enzyme molecules is extremely limited. This arises conceptually because the mass of the hydrogen is approximately 1840 times greater than that of the electron. The probability of tunnelling decreases with increasing mass, which reduces significantly the probability of hydrogen versus electron tunnelling. Nevertheless, for those enzyme-catalysed reactions with a large activation energy – requiring a

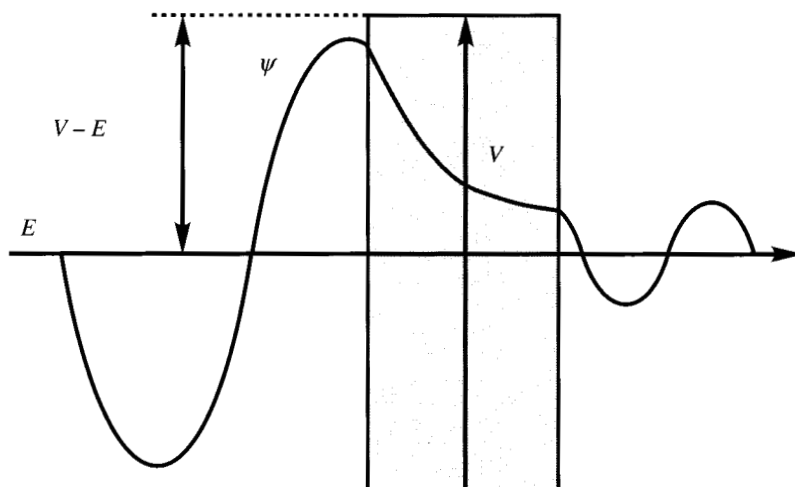


Figure 2.3. Tunnelling of a wave with kinetic energy E through a rectangular potential energy barrier, height V . The narrower the barrier, the smaller the mass of the particle and the smaller the difference between V and E , the greater the tunnelling probability. If the amplitude of the wave has not reached zero at the far side of the barrier, it will stop decaying and resume the oscillation it had on entering the barrier (but with smaller amplitude).

large amount of energy to pass from reactants to products – quantum tunnelling is an attractive means of transferring hydrogen from reactant to product. Until recently, quantum tunnelling was thought to be significant only at very low (so-called ‘cryogenic’) temperatures. However, deviations from classical transition state theory behaviour have been seen recently, implying that hydrogen tunnelling may be significant at physiological temperatures. These results have, in the main, been modelled as hybrid ‘over’ (transition state theory) and ‘through’ (quantum tunnelling) barrier transfer reactions, i.e. quantum correction models of transition state theory.

Our own studies have revealed for the first time that quantum tunnelling can be the *sole* means by which an enzyme catalyses hydrogen transfer during C–H (carbon–hydrogen) bond breakage. The reaction pathway does not pass up the energy barrier prior to tunnelling – as with the quantum correction models of transition state theory – but tunnels through the barrier from the starting (or so-called ‘ground’) state.

Paradoxically, reaction rates (as with transition state theory) are still highly dependent on temperature. This observation is inconsistent with a pure 'ground state' tunnelling reaction, since the probability of tunnelling (and thus rate of reaction) is a function of barrier width, but is independent of temperature. This apparent paradox is resolved by taking into account the temperature-dependent natural breathing of enzyme molecules which distorts the structure of the protein to produce the geometry required for nuclear tunnelling (achieved by reducing the width of the barrier between reactants and products, thus increasing the probability of tunnelling). In this dynamic view of enzyme catalysis, it is thus the width – and not the height (as with transition state theory) – of the energy barrier that controls the reaction rate.

The important criterion thus becomes the ability of the enzyme to distort and thereby reduce barrier width, and not stabilisation of the transition state with concomitant reduction in barrier height (activation energy). We now describe theoretical approaches to enzymatic catalysis that have led to the development of dynamic barrier (width) tunnelling theories for hydrogen transfer. Indeed, enzymatic hydrogen tunnelling can be treated conceptually in a similar way to the well-established quantum theories for electron transfer in proteins.

2.2 Enzyme catalysis in the classical world

In the classical world (and biochemistry textbooks), transition state theory has been used extensively to model enzyme catalysis. The basic premise of transition state theory is that the reaction converting reactants (e.g. A–H + B) to products (e.g. A + B–H) is treated as a two-step reaction over a static potential energy barrier (Figure 2.1). In Figure 2.1, $[A \cdots H \cdots B]^{\ddagger}$ is the transition state, which can interconvert reversibly with the reactants (A–H + B). However, formation of the products (A + B–H) from the transition state is an irreversible step.

Transition state theory has been useful in providing a rationale for the so-called 'kinetic isotope effect'. The kinetic isotope effect is used by enzymologists to probe various aspects of mechanism. Importantly, measured kinetic isotope effects have also been used to monitor if non-classical behaviour is a feature of enzyme-catalysed hydrogen transfer reactions. The kinetic isotope effect arises because of the differential reactivity of, for example, a C–H (protium), a C–D (deuterium) and a C–T (tritium) bond.

The electronic, rotational and translational properties of the H, D and T atoms are identical. However, by virtue of the larger mass of T compared with D and H, the vibrational energy of C–H > C–D > C–T. In the transition state, one vibrational degree of freedom is lost, which leads to differences between isotopes in activation energy. This leads in turn to an isotope-dependent difference in rate – the lower the mass of the isotope, the lower the activation energy and thus the faster the rate. The kinetic isotope effects therefore have different values depending on the isotopes being compared – (rate of H-transfer) : (rate of D-transfer) \approx 7:1; (rate of H-transfer) : (rate of T-transfer) \approx 15:1 at 25 °C.

For a single barrier, the classical theory places an upper limit on the observed kinetic isotope effect. However, with enzyme-catalysed reactions, the value of the kinetic isotope effect is often less than the upper limit. This can arise because of the complexity of enzyme-catalysed reactions. For example, enzymes often catalyse multi-step reactions – involving transfer over multiple barriers. In the simplest case, the highest barrier will determine the overall reaction rate. However, in the case where two (or more) barriers are of similar height, each will contribute to determining the overall rate – if transfer over the second barrier does not involve breakage of a C–H bond, it will not be an isotope-sensitive step, thus leading to a reduction in the observed kinetic isotope effect. An alternative rationale for reduced kinetic isotope effects has also been discussed in relation to the structure of the transition state. For isoenergetic reactions (i.e. the reactants and products have the same energy; the total energy change = 0), the transition state is predicted to be symmetrical and vibrations in the reactive C–H bond are lost at the top of the barrier. In this scenario, the maximum kinetic isotope effect is realised. However, when the transition state resembles much more closely the reactants (total energy change < 0) or the products (total energy change > 0), the presence of vibrational frequencies in the transition state cancel with ground state vibrational frequencies, and the kinetic isotope effect is reduced. This dependence of transition state structure on the kinetic isotope effect has become known as the ‘Westheimer effect’.

2.3 A role for protein dynamics in classical transfers

The transition state theory is likely an oversimplification when applied to enzyme catalysis – it was originally developed to account for gas phase

reactions. Solvent dynamics and the natural 'breathing' of the enzyme molecule need to be included for a more complete picture of enzymatic reactions. Kramers put forward a theory that explicitly recognises the role of solvent dynamics in catalysis. For the reaction $\text{Reactants} \rightarrow \text{Products}$, Kramers suggested that this proceeds by a process of diffusion over a potential energy barrier. The driving force for the reaction is derived from random thermally induced structural fluctuations in the protein, and these 'energise' the motion of the substrate. This kinetic motion in the substrate is subsequently dissipated because of friction with the surroundings and enables the substrate to reach a degree of strain that is consistent with it progressing to the corresponding products (along the reaction pathway) – the so-called 'transient strain' model of enzyme catalysis. By acknowledging the dynamic nature of protein molecules, Kramers' theory (but not transition state theory) for classical transfers provides us with a platform from which to develop new theories of quantum tunnelling in enzyme molecules.

2.4 Wave-particle duality and the concept of tunnelling

Tunnelling is a phenomenon that arises as a result of the wave-properties of matter. Quantum tunnelling is the penetration of a particle into a region that is excluded in classical mechanics (due to it having insufficient energy to overcome the potential energy barrier). An important feature of quantum mechanics is that details of a particle's location and motion are defined by a wavefunction. The wavefunction is a quantity which, when squared, gives the probability of finding a particle in a given region of space. Thus, a nonzero wavefunction for a given region means that there is a finite probability of the particle being found there. A nonzero wavefunction on one side of the barrier will decay inside the barrier where its kinetic energy, E , is less than the potential energy of the barrier, V (i.e. $E < V$; if $E > V$, it can pass over the barrier). On emerging at the other side of the barrier, the wavefunction amplitude is nonzero, and there is a finite probability that the particle is found on the other side of the barrier – i.e. the particle has tunneled (Figure 2.3).

Quantum tunnelling in chemical reactions can be visualised in terms of a reaction coordinate diagram (Figure 2.4). As we have seen, classical transitions are achieved by thermal activation – nuclear (i.e. atomic position) displacement along the R curve distorts the geometry so that the

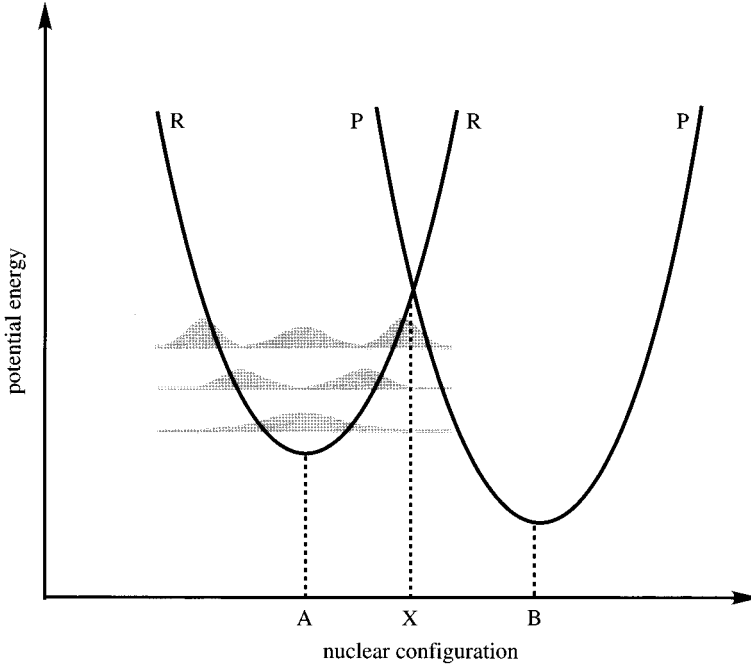


Figure 2.4. Reaction coordinate diagram for a simple chemical reaction. The reactant A is converted to product B. The R curve represents the potential energy surface of the reactant and the P curve the potential energy surface of the product. Thermal activation leads to an over-the-barrier process at transition state X. The vibrational states have been shown for the reactant A. As temperature increases, the higher energy vibrational states are occupied leading to increased penetration of the P curve below the classical transition state, and therefore increased tunnelling probability.

intersection of the R and P curves is reached (the so-called transition state). Quantum mechanics is based on the premise that energy is quantised (i.e. can have only specific, discrete values). Thus in the reaction coordinate diagram, the quantised vibrational energy states of the reactant and product can be depicted (Figure 2.3). At ambient temperatures it is almost exclusively the ground state vibrational energy levels that are populated.

Factors that enhance tunnelling are a small particle mass and a narrow potential energy barrier. In biology, electron transfer is known to occur over large distances (up to about 25×10^{-10} m). Given the mass of protium is 1840 times that of the electron, the same probability for protium

tunnelling gives a transfer distance of 0.6×10^{-10} m. This distance is similar to the length of a reaction coordinate and is thus suggestive of high tunnelling probability. The larger masses of deuterium and tritium lead to corresponding transfer distances of 0.4×10^{-10} m and 0.3×10^{-10} m, respectively, thus making kinetic isotope effect studies attractive for the detection of hydrogen tunnelling in enzymes. Tunnelling is also favoured by high and narrow energy barriers; for low and wide barrier shapes, transfer is dominated by the classical route.

Thus, different strategies are required for optimising enzyme structure for reactions to proceed by quantum tunnelling rather than classical transfer. For classical transfers, the enzyme has evolved to reduce the height of the potential energy barrier and to stabilise the transition state (rather than ground state). In the quantum regime, it is reduction of barrier width and not height that optimises rate. Quantum tunnelling from the ground state requires little or no structural reorganisation of the substrate, and the need to stabilise a transition state is thus eliminated. Exclusion of water from the active sites of enzymes prevents coupling of solvent motion to the transfer reaction, and this leads to a reduction of mass for the transferred particle. In the following sections, we review the evidence for quantum tunnelling in biological catalysis and discuss the strategies employed by enzymes to optimise the transfer process. Surprisingly – and unlike for biological electron transfers – reports of hydrogen tunnelling in enzymatic reactions have been restricted to only a small number of enzyme molecules. The realisation that hydrogen tunnelling occurs in enzymes has been relatively recent. This may, in part, be due to (i) the misconception that the much larger mass of the hydrogen nucleus is inconsistent with tunnelling, and (ii) the erroneous assumption that measured kinetic isotope effects <7 are always indicative of classical hydrogen transfer. Our recent work has demonstrated that hydrogen tunnelling in proteins is inextricably coupled to protein dynamics. This provides a link to the established theories for electron tunnelling in proteins. To provide a framework for the discussion of hydrogen tunnelling in enzymes, protein-mediated electron transfer is discussed below.

2.5 Electron tunnelling in proteins

The transfer of electrons in proteins by a quantum mechanical tunnelling mechanism is now firmly established. Electron transfer within proteins

occurs between two 'centres' (known as redox centres since one reduces the other, and in so doing is itself oxidised) – the 'electron donor' (which is thereby oxidised) supplies an electron to the 'electron acceptor' (which is thereby reduced). This can be modelled using quantum mechanics.

It is well established that electron transfer in proteins is driven by distortion in the nuclear (protein) geometry of the reactant state. This is facilitated by the natural, thermally activated breathing of the protein molecule. Thermal activation of the reactant state leads to overlap with the potential energy curve for the product state – the point of overlap is the nuclear geometry that is compatible with electron tunnelling. At this intersection point, there is an energy barrier through which the electron must tunnel to arrive on the product side. The theory for protein-mediated electron transfer reactions illustrates an important role for protein dynamics in driving the tunnelling process. The importance of dynamic fluctuations in the protein can be appreciated by considering those reactions that have a nonzero change in overall energy for the electron transfer reaction. Since tunnelling is significant only between states of nearly equal energy, tunnelling is unlikely in such instances. However, dynamic fluctuations in the protein overcome this problem. These equalise the energy between the reactant and product at the intersection point of the R (reactant) and P (product) potential energy curves (i.e. their configurations are identical), thus enabling transfer by quantum tunnelling. The term 'vibrationally assisted tunnelling' is therefore appropriate for protein electron transfer reactions. As described below, our recent work has also demonstrated a similar role for dynamic fluctuations of the protein during enzyme-catalysed hydrogen tunnelling. Electron transfer theory therefore provides a useful framework for understanding enzymatic hydrogen tunnelling. Despite this, until very recently tunnelling derivatives of transition state theory – that do not take into account the fluctuating nature of the enzyme – have been used to account fully for enzymatic hydrogen tunnelling. As a backdrop to the very recent dynamic treatments of hydrogen tunnelling in enzymes, we describe below static barrier approaches – i.e. tunnelling correction theories of transition state theory – that have been applied to some enzyme systems.

2.6 Transition state theory and corrections for hydrogen tunnelling

Deviations from classical behaviour are usefully probed via the kinetic isotope effect (Section 2.2). For non-enzymatic reactions, several factors – in addition to inflated kinetic isotope effects (i.e. kinetic isotope effect >7) – have been used to indicate quantum tunnelling of hydrogen. A particularly striking indication of quantum tunnelling comes from studying the temperature dependence of the reaction rate – this manifests itself as curvature in the plot of $\ln(\text{rate})$ vs. $1/T$ (the so-called ‘Arrhenius plot’; where ‘ \ln ’ is the natural logarithm, \log_e , and T is the temperature in kelvin, $=^{\circ}\text{C} + 273$) over an extensive temperature range. Interestingly, this has been observed in non-enzymatic radical reactions. However, curvature in Arrhenius plots is not a useful indicator of quantum tunnelling because the limited experimental temperature range available in studies using enzymes make it impossible to detect any such curvature. An alternative approach is to estimate, from the Arrhenius plot, the activation energy for the reaction (from the slope) and the so-called ‘preexponential factors’ (from the intercept). Large differences in the activation energies for protium and deuterium transfer ($>5.4\text{kJmol}^{-1}$) and values deviating from unity for the ratio of Arrhenius preexponential factors, can indicate non-classical behaviour. In conjunction with inflated kinetic isotope effects, these parameters have been used to demonstrate quantum tunnelling in enzyme molecules.

Small deviations from classical behaviour have been reported for the enzymes yeast alcohol dehydrogenase, bovine serum amine oxidase, monoamine oxidase and glucose oxidase. More recently, the enzyme lipoxxygenase has been shown to catalyse hydrogen transfer by a more extreme quantum tunnelling process. In this case, the apparent activation energy was found to be much smaller than for reactions catalysed by yeast alcohol dehydrogenase, bovine serum amine oxidase, monoamine oxidase and glucose oxidase, suggesting a correlation between apparent activation energy and the extent of tunnelling. Use of a static (transition state theory-like) barrier in the treatment of hydrogen tunnelling in enzymes has allowed the construction of (hypothetical) relationships between the reaction rate and temperature. These relationships are readily visualised in the context of an Arrhenius plot and are observed in studies that employ isotope (i.e. H, D and T) substitution within the reactive bond. The plot can

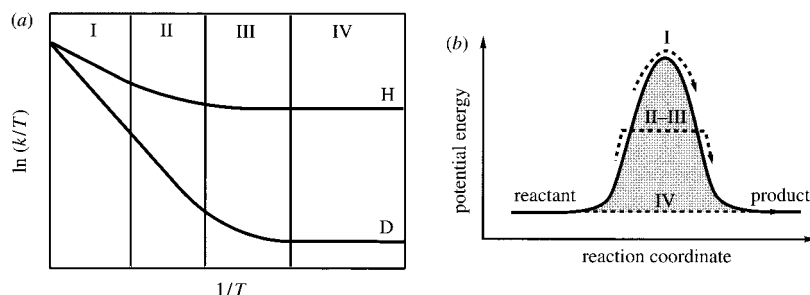


Figure 2.5. The static barrier (transition state theory-derived) model of H-tunneling and definition of tunneling regimes. Panel (a), the four different hydrogen tunnelling regimes. On the plot, 'ln' is the natural logarithm, \log_e , and T is the temperature in kelvin ($=^{\circ}\text{C} + 273$). Panel (b), a static barrier indicating transfer to the product side in each of the regimes shown in (a). In regimes II and III, additional thermal activation may be required to populate higher vibrational energy states of the reactive C–H bond.

be divided into four regimes (Figure 2.5): regime I describes classical (transition state theory) behaviour. Regimes II to IV reveal the effects of quantum tunnelling on the temperature dependence of the reaction rate – the extent of quantum tunnelling increases from regime II to regime IV. In regime II, protium tunnels more extensively than deuterium, thus giving rise to inflated values for the kinetic isotope effect, and a preexponential factor ratio for (H:D) < 1. Regime III is characterised by extensive tunnelling of both protium and deuterium, and the preexponential factor ratios are difficult to predict. Finally, regime IV is the predicted regime for transfer solely by ground state tunnelling. In this case the preexponential factor ratio equals the kinetic isotope effect and the reaction rate is not dependent on temperature (the reaction passes through, and not over, the barrier, thus there is no temperature-dependent term).

Relationships between reaction rate and temperature can thus be used to detect non-classical behaviour in enzymes. Non-classical values of the preexponential factor ratio (H:D \neq 1) and difference in apparent activation energy ($>5.4 \text{ kJ mol}^{-1}$) have been the criteria used to demonstrate hydrogen tunnelling in the enzymes mentioned above. A major prediction from this static barrier (transition state theory-like) plot is that tunnelling becomes more prominent as the apparent activation energy decreases. This holds for the enzymes listed above, but the correlation breaks down for enzymes

catalysing the breakage of C–H bonds – a direct result of the type of potential energy barrier that is used. Temperature-independent tunnelling is a direct result of invoking a static (Eyring-like) potential energy barrier. However, an alternative approach comes from invoking a fluctuating (Kramers-like) potential energy barrier. This is conceptually more realistic as it takes into account the dynamic motion of the protein. These dynamic effects will give rise to more complex temperature dependencies for rates of hydrogen transfer than those illustrated in Figure 2.5. The role of protein dynamics in driving enzymatic hydrogen tunnelling is discussed below.

2.7 Hydrogen tunnelling driven by protein dynamics

In recent years, attempts have been made to model theoretically enzymatic hydrogen tunnelling by incorporating thermal vibrations. However, and importantly, none of these approaches have been verified experimentally. Recently, the kinetic data for bovine serum amine oxidase have been re-evaluated for thermally activated substrate vibrations, but with the protein molecule treated as rigid. Computational molecular dynamics simulation studies have also suggested a dynamic role for the protein molecule in enzymatic hydrogen tunnelling. Indeed, some theoretical treatments have recognised the role of thermal motion in the protein in hydrogen tunnelling, but fail to predict the experimentally observed kinetic isotope effect – and again experimental verification of these theories is lacking.

The only (to the best of our knowledge) theoretical treatment of hydrogen transfer by tunnelling to explicitly recognise the role of protein dynamics, and relate this in turn to the observed kinetic isotope effect, was described by Bruno and Bialek. This approach has been termed ‘vibrationally enhanced ground state tunnelling theory’. A key feature of this theory – and one that sets it apart from many other theoretical approaches – is that tunnelling occurs from the ground state vibrational energy levels of the substrate, i.e. there is no thermal activation of the substrate. The temperature dependence of the reaction is therefore attributed to the natural thermally induced breathing of the enzyme molecule, thus shortening the distance the hydrogen must tunnel. Thus, the natural breathing of the enzyme molecule can be visualised in the context of the familiar R (reactant) and P (product) potential energy curve depiction encountered in discussions of electron transfer in proteins (Section 2.5). Hydrogen tunnelling does not occur until the geometry of the protein is distorted so that the R

and P curves intersect (Figure 2.6). At the intersection point (X) of the two curves, the hydrogen tunnels – the average tunnelling probability is decreased when heavier isotopes (e.g. deuterium) are transferred, thus giving rise to a kinetic isotope effect >1 . At the intersection point, tunnelling is from the vibrational ground state – since vibrational energy differences are comparable to barrier height, and therefore vibrational excitation would lead to a classical ‘over-the-barrier’ transfer.

Clearly protein dynamics is hypothesised to have a major role in driving hydrogen tunnelling in enzymes. However, like all hypotheses, this requires experimental verification. The activation energy of the reaction is associated with distortion of the protein molecule. Following the tunnelling event, rapid movement away from the intersection point along the P curve prevents coherent oscillations of the hydrogen between the R and P curves. As such, the reaction is modelled in much the same way as electron transfer in proteins (i.e. Fermi’s Golden Rule applies and the non-adiabatic regime operates). A key prediction of this theory is that hydrogen tunnelling can occur even when the value of the kinetic isotope effect <7 , thus suggesting that (contrary to current dogma) kinetic isotope effects may be poor indicators of quantum tunnelling in enzymes. This is an important point, since static barrier models of hydrogen tunnelling suggest that hydrogen tunnelling does not occur when kinetic isotope effect <7 . This indicates that detailed temperature dependence studies are required to demonstrate unequivocally that tunnelling is a feature of an enzyme catalysed reaction.

The fluctuating enzyme model of hydrogen tunnelling can be divided into two reaction components: (i) a thermally activated nuclear reorganisation step, and (ii) the hydrogen tunnelling event at the intersection point of the potential energy curves. This leads to three possible rate-limiting regimes in which either (i) nuclear reorganisation is rate-limiting, (ii) quantum tunnelling is rate-limiting, or (iii) both factors contribute to the observed rate. The value of the kinetic isotope effect is affected directly by these steps. When nuclear reorganisation is rate limiting, the kinetic isotope effect is unity (since this is independent of isotope) and reaction rates are dependent on solvent viscosity (i.e. the ease with which the protein structure can reorganise). In the quantum tunnelling limiting regime, the kinetic isotope effect is not dependent on solvent viscosity and is not unity (since tunnelling rate is a function of isotope). However, when both nuclear reorganisation and quantum tunnelling contribute to the

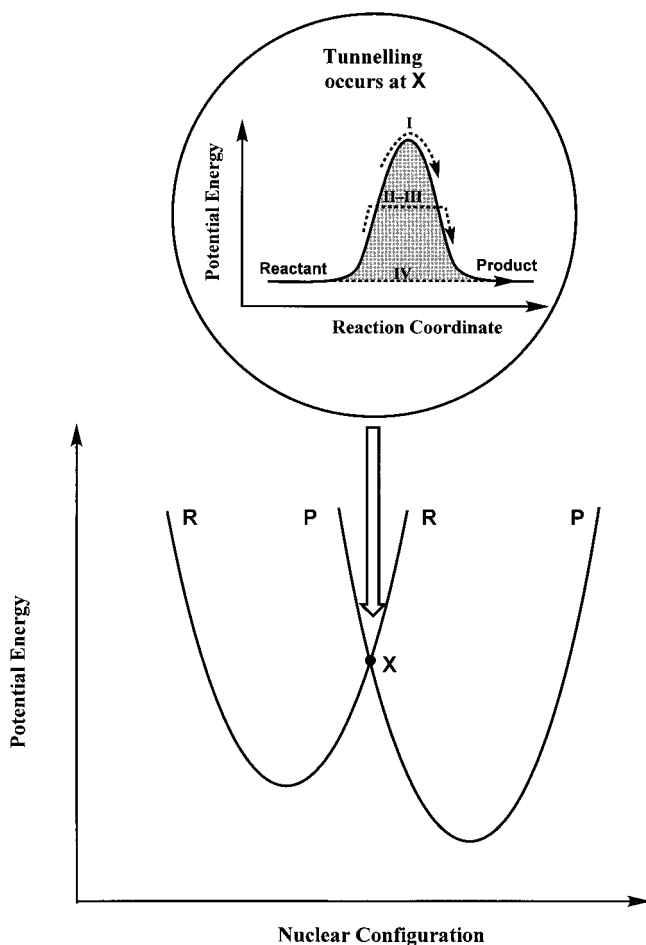


Figure 2.6. The dynamic barrier model of hydrogen tunnelling. Reactant (R) and product (P) energy curves for distortion of the protein scaffold. Hydrogen tunnelling does not occur until the geometry of the protein is distorted so that the R and P curves intersect. Thus, the intersection point (X) is the optimum geometry required for hydrogen transfer. At the intersection point, transfer can be by the classical (I), ground state tunnelling (IV) or intermediate regimes (II) and (III). In regimes II and III, additional thermal activation (other than that required to distort the protein scaffold to the optimum geometry for transfer, i.e. the R–P intersection) may reflect (i) population of higher vibrational energy levels of the reactive C–H bond and/or (ii) transfer via a combination of classical (over-the-barrier) and quantum mechanical routes.

observed rate the kinetic isotope effect is viscosity-dependent – as viscosity increases the nuclear reorganisation step becomes rate limiting, and thus the kinetic isotope effect tends to unity. In experimental studies, measurements of (i) increased viscosity or (ii) decreased temperature effects on the kinetic isotope effect may be used to discriminate between these possible regimes, since both would be expected to selectively perturb geometrical distortion of the protein.

The vibrationally enhanced ground state tunnelling theory assumes that hydrogen transfer occurs entirely by quantum mechanical tunnelling. The model is therefore appropriate for those enzymes catalysing ground state tunnelling (see below). The model is likely to be incomplete for those enzymes where tunnelling occurs just below the saddlepoint of the energy surface (i.e. the reactant passes up the energy barrier before tunnelling) – in these situations hydrogen transfer is likely to occur by a combination of classical and quantum mechanical behaviour. In the case where hydrogen transfer is by a combination of classical and quantum mechanical effects, the activation energy will reflect partitioning of energy into a wide range of modes within the protein, e.g. changes in protein geometry, bond angles of reacting substrate etc., as well as thermal excitation of the reactive C–H bond. However, experimental verification of the vibrationally enhanced ground state tunnelling theory would demonstrate the importance of protein dynamics in enzymatic hydrogen tunnelling. By analogy, therefore, protein dynamics would also be expected to play a major role in those enzymes where hydrogen tunnelling is not from the ground state, but from an excited state of the substrate molecule. Experimental verification of a role for protein dynamics is thus a key milestone in developing theories for enzymatic hydrogen tunnelling – this verification is described below.

2.8 Experimental demonstration of vibration-driven tunnelling

Kinetic data for bovine serum amine oxidase were originally analysed in terms of the tunnelling correction derivatives of transition state theory, but the data are also consistent with – although not verification of – the vibrationally enhanced ground state tunnelling theory. Alternatively, the bovine serum amine oxidase data can also be interpreted in terms of a hydrogen tunnelling reaction driven by substrate oscillations. Thus, ambiguity remains concerning the correct theoretical treatment of the bovine serum amine oxidase kinetic data. This ambiguity arises because the

complex temperature dependence of the reaction can be modelled in a variety of ways. Our recent studies on enzymatic C–H bond cleavage have, however, provided verification of vibrationally enhanced ground state tunnelling theory and also, for the first time, proved the existence of a ground state H- and D-tunnelling regime in an enzyme molecule.

Our kinetic isotope effect and temperature-dependent studies of the reaction catalysed by the bacterial enzyme methylamine dehydrogenase have revealed that the rate of reduction of the enzyme redox centre (tryptophan tryptophylquinone) by substrate has a large, temperature independent kinetic isotope effect. Reduction of this redox centre is a convenient way of following C–H bond breakage in this enzyme, since breakage of the bond and reduction of the cofactor occur simultaneously. An Arrhenius-like plot revealed that ground state quantum tunnelling is responsible for the transfer of the hydrogen nucleus. This is indicated by the linear and parallel nature of the plots for C–H and C–D bond breakage, which should be compared with regime IV of the corresponding hypothetical plot for a static potential energy barrier (Figure 2.7). However, contrary to the static

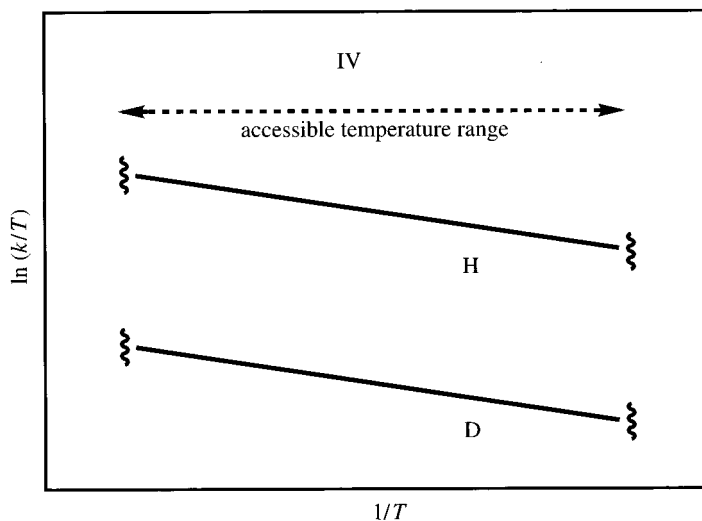


Figure 2.7. Expected temperature dependence (in the experimentally accessible temperature range) in regime IV in the context of Figure 2.5. Ground state tunnelling occurs in regime IV. The experimental data for methylamine dehydrogenase are apparently linear in regime IV but, as noted in the text, this linearity will likely not extend to cryogenic temperatures.

potential energy barrier model for hydrogen tunnelling, reaction rates are strongly dependent on temperature (apparent activation energy $\sim 45 \text{ kJ mol}^{-1}$) and, importantly, this activation energy was found to be independent of isotope. These observations indicate that thermal distortion of the protein scaffold – but not vibrational excitation of the substrate – are required to *drive* hydrogen transfer. Thus, a fluctuating energy surface is a feature of the tunnelling process. The vibrationally enhanced ground state tunnelling theory equivalent of regime IV of the static barrier plot (Figure 2.7) recognises that thermal motions of the protein molecule are required to distort the protein scaffold into conformations compatible with hydrogen tunnelling. Regime IV of the vibrationally enhanced ground state tunnelling theory plot therefore has a nonzero value for the slope, the value of which is the energy required to distort the protein into the geometry compatible with hydrogen tunnelling. With methylamine dehydrogenase, it has thus been possible to quantify the energy term associated with structural distortion of the protein during an enzyme catalysed reaction.

The temperature dependence in regime IV – i.e. ground state tunnelling – for vibrationally enhanced ground state tunnelling theory contrasts markedly with that for the static barrier model. Although there is a sizeable energy term in this regime for the vibrationally enhanced ground state tunnelling theory model (apparent activation energy $\sim 45 \text{ kJ mol}^{-1}$), the apparent linearity seen in the accessible temperature range for methylamine dehydrogenase probably does not extend to lower temperatures. At low temperatures, nuclear vibrations will be frozen, thus preventing distortion of the nuclear scaffold into geometries compatible with hydrogen tunnelling. Thus, over a large temperature range, complex temperature dependencies of the reaction rate are predicted.

Ground state tunnelling driven by protein dynamics (vibrationally enhanced ground state tunnelling theory) is the only theoretical treatment consistent with our work on methylamine dehydrogenase. As indicated above, a prediction of vibrationally enhanced ground state tunnelling theory is that ground state tunnelling may occur even when the kinetic isotope effect < 7 – a regime interpreted previously as indicating classical behaviour. The kinetic isotope effect with methylamine dehydrogenase is large (~ 18), and thus the presence of tunnelling is predicted by current dogma. In the case of sarcosine oxidase, our studies on hydrogen tunnelling have shown that the kinetic isotope effect approaches the classical limit. Furthermore, our recent analysis of hydrogen tunnelling in trimethylamine

dehydrogenase has indicated that, under certain conditions (and contrary to current dogma), ground state tunnelling occurs even when the kinetic isotope effect <7 . This observation lends support to the validity of vibrationally enhanced ground state tunnelling theory in describing enzymatic hydrogen tunnelling.

2.9 Significance of hydrogen tunnelling in enzymes

Both methylamine dehydrogenase and trimethylamine dehydrogenase catalyse the breakage of stable C–H bonds. These are difficult reactions if viewed in terms of the classical transition state theory approach to catalysis, but the structural plasticity of methylamine dehydrogenase and trimethylamine dehydrogenase (in common with other enzymes) provides a means of circumventing this problem by facilitating ground state tunnelling. Vibration driven ground state tunnelling may therefore be a common mechanism for the breakage of C–H bonds by enzymes and this may extend to other types of hydrogen transfer reactions.

The dynamic barrier approach to catalysis has major implications for how hydrogen transfer reactions – and indeed other reactions – are modelled theoretically. Given the dynamic nature of protein molecules, it is perhaps surprising that the indiscriminate use of transition state theory has persisted for so long. For classical transfers, Kramers' theory seems appropriate, and this is an excellent platform from which to develop theories of quantum tunnelling in enzymes. For those reactions that proceed by quantum tunnelling, it is the energy barrier width that is important in determining reaction rate. Tunnelling probability depends on the mass of the transferred particle, the net driving force and the height and width of the reaction barrier. Proteins can facilitate this by (i) reduction of mass (e.g. exclusion of water), (ii) an equalisation of energy states for reactants and products and, most importantly, (iii) a reduction in barrier width. Exclusion of water from enzyme active sites is achieved readily and documented amply in the literature. The exploitation of protein dynamics to equalise energy states and shorten tunnelling distance is, however, less well appreciated but nevertheless pivotal.

2.10 Enzymology in the future

An in-depth understanding of biological catalysis is central to the successful exploitation of enzymes by mankind. At the end of the last century the

'Lock and Key' mechanism propounded by Emil Fischer – in which the enzyme accommodates a specific substrate like a lock does a key – opened the door to our understanding of enzyme catalysis. This has evolved to take account of protein motion in the 'Induced Fit' model of catalysis in which the enzyme has one conformation in the absence, and another conformation in the presence, of substrate. The induced fit model of catalysis recognises preferred complementarity to the transition state and has provided a conceptual framework for transition state theory. Now, moving into the new Millennium, our understanding has progressed yet further by highlighting the role of (i) protein dynamics and (ii) quantum tunnelling in enzyme catalysis. Thus, the rules underpinning our design and understanding of enzymes have changed significantly. Important areas where these rules apply include enzyme redesign, the production of catalytic antibodies, design of enzyme inhibitors (drugs and pesticides), enzymatic fine chemical synthesis and use of enzymes in bulk processing (e.g. paper manufacture, food industry and detergents).

Enzyme redesign strategies currently attempt to reduce the activation energy (i.e. the barrier height) by seeking maximum complementarity with the transition state and destabilisation of the ground state. This is the approach adopted in producing catalytic antibodies. Here, an animal's immune system is exposed to a transition state analogue, thus inducing antibodies with surface complementarity to the transition state. Although in principle this is an elegant approach to producing novel catalysts, in practice it is usual for catalytic antibodies to have poor catalytic rates. These studies imply that knowledge of the transition state alone is not sufficient to develop a good catalyst. Insight into additional factors required for efficient catalysis has come from recent work. An important determinant of catalytic efficiency is the role of protein dynamics. The structural plasticity of protein molecules is important in driving both classical and quantum mechanical transfers. As we have seen, in quantum mechanical transfers distortion of the enzyme molecule transiently compresses barrier width and equalises reactant and product energy states. In contrast to classical models of catalysis, for vibrationally driven ground state tunnelling maximum complementarity with the ground state should be sought. Additionally, the exclusion of water will reduce the mass of the transferred particle (thus increasing tunnelling probability). The challenge will therefore be to incorporate these new aspects into programmes of rational enzyme redesign and to provide a unified theory for enzyme catalysed reactions. Over the past century, our understanding of catalysis has been based

primarily on static pictures of enzymes and enzyme-ligand complexes. As we start the new millennium, our quest for a better understanding will be driven by an appreciation of a role for protein dynamics – both experimental and computational – in driving enzyme-catalysed reactions. The future will thus witness a flurry of activity directed at understanding the role of quantum mechanics and protein motion in enzyme action.

2.11 Further reading

- Bendall, D. S. (ed.) 1996 *Protein electron transfer*. Oxford: Bios Scientific Publishers.
- Fersht, A. 1985 *Enzyme structure and mechanism*. New York: W. H. Freeman.
- Kohen, A. & Klinman, J. 1999 Hydrogen tunneling in biology. *Chem. Biol.* **6**, R191–R198.
- Scrutton, N. S., Basran, J. & Sutcliffe, M. J. 1999 New insights into enzyme catalysis: ground state tunnelling driven by protein dynamics. *Eur. J. Biochem.* **264**, 666–671.
- Sutcliffe, M. J. & Scrutton, N. S. 2000 Enzymology takes a quantum leap forward. *Phil. Trans. R. Soc. Lond. A* **358**, 367–386.



3

World champion chemists: people versus computers

Jonathan M. Goodman

*Department of Chemistry, Cambridge University, Lensfield Road, Cambridge
CB2 1EW, UK*

Making molecules has been important to human society from prehistoric times. The extraction of tin and lead from their ores has been possible for thousands of years. Fermentation has also been controlled to produce alcohol for millennia. In the past century, carbon-containing molecules have become increasingly important for the development of new substances, including plastics, other new materials and health products. Organic chemistry was originally the study of compounds connected with life, but, more than a century and a half ago, Wöhler showed it was possible to make an organic compound (urea, which may be extracted from urine) from inorganic (that is, not living) compounds. What had seemed a precise distinction between living and non-living compounds became hazy. The subject may now be defined as the study of molecules which contain carbon atoms, although the precise boundaries of the area are not clear, as the overlaps with biology, with materials science, with inorganic chemistry, and with physics can all be debated and boundaries drawn and re-drawn. However, it is clear that understanding of organic chemistry advanced tremendously in the closing century of the second millennium.

Increasing knowledge of the properties of molecules has made it possible to synthesise very complicated compounds. Organic synthesis is engineering on an atomic scale, and requires delicate operations to be performed on objects which are too small to see. It also requires techniques of mass production, because single molecules are usually not useful by themselves. A car factory may produce tens of thousands of cars each year, but

this is very small scale compared to the job of a synthetic chemist. A pint of beer contains approximately 10^{25} (ten million million million million) molecules. If you were to pour a pint of beer into the sea, wait for the waves to mix it well all around the world, and then take a pint of sea water from any part of any ocean, that pint would probably contain a thousand molecules from the original pint. A successful synthesis of a new molecule would not make hundreds or thousands of copies of the molecules, but millions of millions of millions. For this to be possible, every step of the synthesis must work well.

In order to make a complex molecule, it is necessary to have methods which join simpler molecules together, and also techniques to make small changes to different bits of the molecule, once the framework has been constructed. There is an enormous variety of reagents which can be used to transform one arrangement of atoms into another. A common transformation is to turn alcohols into ketones (Figure 3.1). Every reagent which is

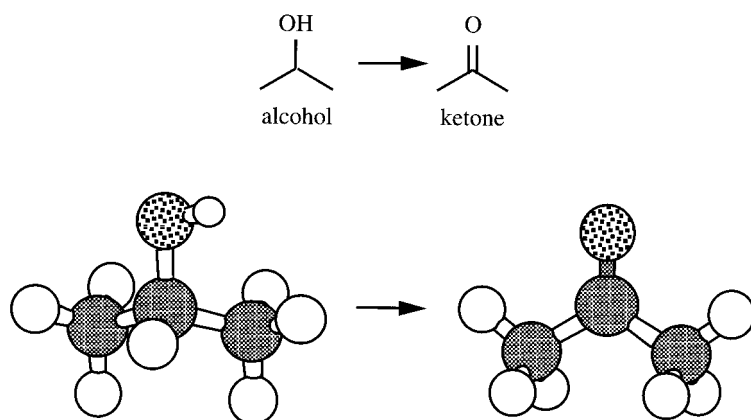


Figure 3.1. The transformation of an alcohol to a ketone. The line drawings at the top show the same molecules as the ball and stick representations below. In the lower version, hydrogen atoms are white, carbon atoms are dark grey, and oxygen atoms are speckled. In the more concise representation at the top, hydrogen atoms attached to carbon are omitted, and the carbon–oxygen double bond in the ketone is drawn with a double line. The lower diagram shows the two hydrogens which must be removed to turn the alcohol into the ketone. One of these is on the oxygen, and the other on the central carbon atom. Many reagents are available which will transform an alcohol into a ketone, removing these two hydrogens and turning the carbon–oxygen single bond into a double bond.

added will act on the whole molecule. This is not a problem for the structures illustrated in Figure 3.1, because there is only one alcohol group in the starting material and so all of the alcohols are transformed into ketones. It is a problem if the same transformation is used to make a more complicated molecule, such as PM-toxin (Figure 3.2). In this molecule,

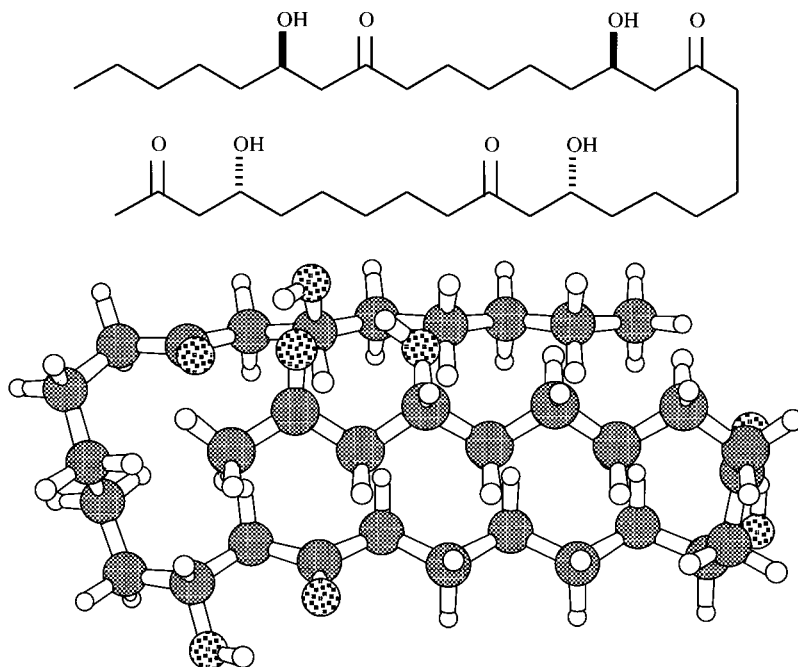


Figure 3.2. PM-toxin A. This molecule, which is produced by the fungal pathogen *Phylosticta maydis*, has been the cause of major epidemics of leaf blight disease in the United States. The molecule is toxic only to specific plants. As in Figure 3.1 both a line drawing and a ball and stick representation of the same molecule are shown. The top representation is much more concise, but does not give information about the shape of the molecule. The lower illustration shows how the atoms are arranged in space as well as how they are connected to each other. A low energy conformation of the molecule is illustrated. This molecule has four alcohol groups (an oxygen joined to a carbon atom and a hydrogen atom by single bonds). Changing one of these to a ketone (an oxygen joined to a carbon by a double bond) without affecting the others will be difficult, as all the reagents which can do this are likely to act on the whole molecule, not just on a specific part of it.

there are several alcohols, and several ketones. A synthesis could not finish by oxidising just some of the alcohols to ketones, because the reagent would not know which alcohols should be oxidised and which should not. This is a major problem for synthesis. How is it possible to selectively make ketones in the presence of alcohols? More generally, how can a transformation be made to act on only a part of a molecule?

Two general approaches can be used to find solutions to this problem. Selective reagents could be developed, so that it is possible to change some alcohols into ketones without affecting others in the same molecule. For example, the lower left hand ketone in PM-toxin is close to the end of the carbon chain. Might it be possible to develop a reagent that only oxidises alcohols which are close to the end of carbon chains? An approach of this sort would require a very good knowledge of the properties of reagents. Alternatively, a strategic approach could be tried. The molecule could be joined together in such a way that the question does not arise, because the alcohols and ketones are already in the right places as a result of the choice of joining processes. In practice, a combination of these methods may be required in order to make complex molecules.

As a result, organic synthesis is an extremely demanding discipline, requiring both a wide knowledge of chemistry and also the ability to develop complete strategies for the construction of molecules. If the last step of a synthesis does not work, then it may be necessary to begin again by altering the first step. The science fiction character, Dr Who, has machines which can synthesise molecules, just given the target structure. Might it be possible to build such a machine? The physical manipulations of mixing and purifying compounds can be automated to a large extent, and it is possible to imagine building a machine which could do the mechanical tasks of a highly trained synthetic chemist, although it would be far more expensive and probably less effective than a skilled individual. The main difficulty in the construction of such a machine would be to provide the machine with suitable instructions for the synthesis.

Organic synthesis is sometimes compared with a game of chess, where the effects of the opening moves are felt right through to the end game, and where the total number of possible situations is greater than can be comprehensively analysed by any computer. Chess games require an opponent, whose responses to the strategy chosen by the opening moves determine the course of the game. Organic synthesis may be regarded as a similar challenge. A good chess player may reasonably be expected not to make a

foolish response to any situation that is presented, but the details of the response are not predictable. The same is true of organic synthesis, contending with the properties of molecules. Organic reactions are well understood, but if a reaction is performed in a completely new context, then the molecule's response may not be exactly as expected from the experience gained through earlier studies of related systems. The variety of possible responses makes chess a demanding game, and organic synthesis a challenging subject.

Chess is, however, succumbing to computers. Only the very best human chess players can compete on a level with the best chess-playing computers, and every year the computers become more powerful. It is unlikely that the chess champion of the world will be human for any of the third millennium. At the end of the second millennium, the best designers of organic syntheses were unquestionably human. For how much longer will this pre-eminence continue?

A molecule-building computer would need to understand chemistry. This is possible. Quantum mechanics provides a method for calculating how molecules behave with a high level of precision, using Schrödinger's equation. In 1929, Dirac wrote 'The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble' (Dirac 1929). Since that time, advances in computers have made some of these complicated equations not only soluble, but routinely used. However, the equations become more complicated very rapidly as larger systems are considered, and so the exact application of these laws remains out of reach, except for the smallest molecules. Many useful approximations have been developed in order to extend the range of possible calculations, and the effects of these simplifications are now well known. The 1998 Nobel prize in chemistry was awarded to Pople and Kohn for the development of methods for calculating chemistry.

Solving quantum mechanical problems is a conceptually straightforward way of solving organic chemistry. The problem is simply one of computer power. In order to calculate the energy of a molecule the size of PM-toxin (Figure 3.2) or bryostatin (Figure 3.3), an extremely complex calculation must be done. It is now possible to do this, using advanced quantum chemistry programs. If lower accuracy is acceptable, then the calculation may even be made easy using the much greater approximations of

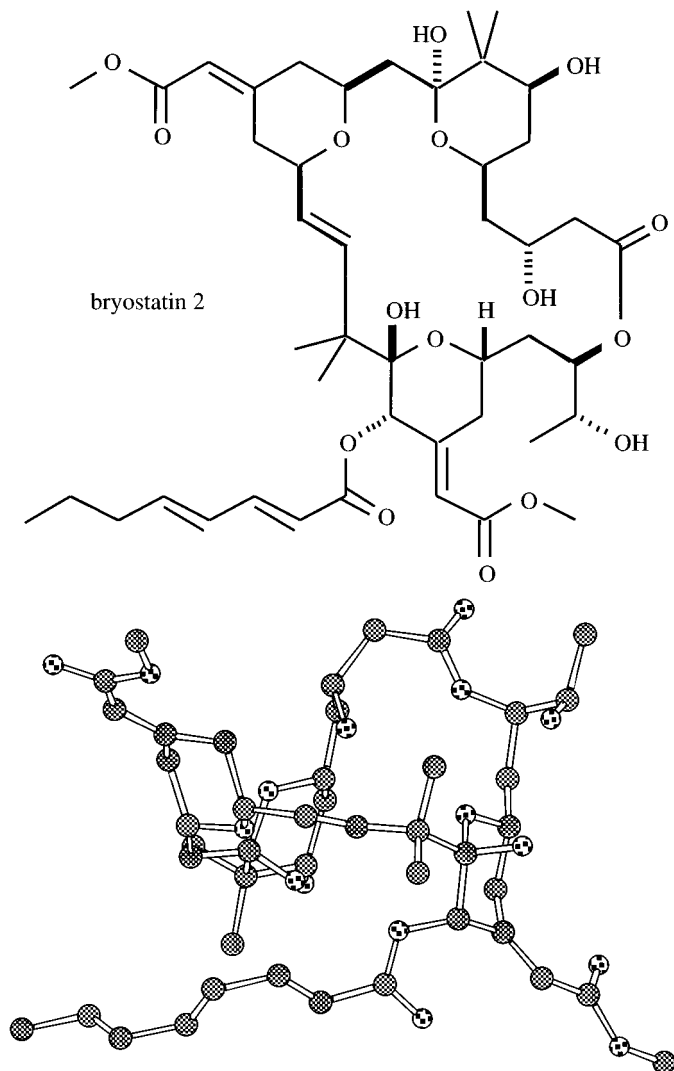


Figure 3.3. Bryostatin 2. Bryostatin 2 ($C_{45}H_{66}O_{16}$) is a biologically active marine natural product which may have useful anti-cancer properties. It was recently synthesised at Harvard by Professor David Evans and his research group. In this illustration, all of the hydrogen atoms are omitted in order to simplify the structure. The lower diagram shows a low energy conformation of bryostatin 2, but it may only be a local minimum and not a global minimum. Many other conformations are accessible at room temperature.

force fields. Once the energy has been found, it is possible to calculate the preferred shape of the molecule, by finding alterations to the shape of the molecule which lower the total energy. This process of altering the structure and recalculating the energy is continued until all small changes to the structure lead to an increase in energy. The shape that the molecule has now reached is called a minimum energy conformation. This requires many calculations of the energy of the structure.

This does not solve the problem of synthesis. A minimum energy conformation is the lowest energy point in the immediate vicinity, but it may not be the lowest energy geometry available to the molecule. The lowest energy point of all is called the global minimum. There can only be one global minimum for any molecule, but there may be very many local minima. These are geometries for which any small change will increase the energy of the structure, but for which larger changes may lead to a decrease in energy, so they must be higher in energy than the global minimum. This can be compared with a mountainous landscape. Only one point can be the lowest point of all, the global minimum, but there may be many points from which every direction you choose to walk will be up hill.

For a molecule containing several alcohol groups, some conformations may have particular alcohols tucked into the centre of the molecule. This may be helpful, if it means that these alcohols will not react, and others in the molecule may do so. But will each conformation be accessible? One way to assess this is to make a list of all the minima on the surface, and to examine the properties of each. The higher energy minima will be less likely to be occupied than the lower energy minima, and this difference can be quantified. This process, called conformation searching, requires many minimisations, each of which requires many energy calculations, and so multiplies the total time required for the analysis. This leaves out all of the parts of the landscape between the minima, and this can be a problem. There are ways of taking these into account, but they are even more time consuming.

A simple molecule is illustrated in Figure 3.4. This is pentane, a chain of five carbon atoms, with hydrogen atoms ensuring that each carbon makes four connections to its surroundings. Pentane has four minimum energy conformations, as illustrated. The conformation analysis is straightforward, but pentane is a simple molecule. It is not easy to assess accurately the number of conformations accessible to PM-toxin, but the answer is certainly well into four figures, for structures only slightly higher

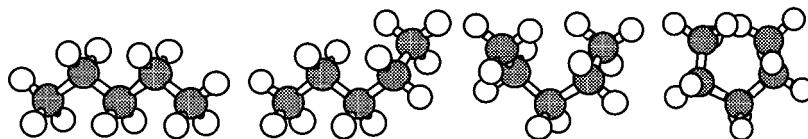


Figure 3.4. Pentane. The diagram shows the four minimum-energy conformations of pentane. The global minimum is on the far left. Reflection and rotation of some of these geometries would generate more structures, but nothing with a different energy. Pentane is a simple molecule. More complicated molecules have many more conformations. Bryostatatin 2 and PM-toxin A have so many minimum-energy conformations that to list them all would be a major undertaking and would require a large library to store the result.

in energy (a few tens of kilojoules) than the global minimum. For bryostatatin, there are probably many more accessible conformations. Simply being able to calculate the energy of one of these molecules is a long way from understanding its structural properties, which will require many energy calculations.

In addition to finding the minima for each intermediate in a synthesis, it is also necessary to be able to analyse reactivity. This is a more difficult problem than conformation searching, because it is now possible for bonds to break. The range of movements available to the molecule is far larger, and it is also necessary to consider which bond will break most easily, and what factors are present which will drive the reaction forwards. If there are many competing reactions, then these calculations may have to be very precise in order to distinguish between them.

This problem is made easier because the different reactions have similar starting points, so the question of the most favourable reaction only requires the comparison of similar systems, and this is a great advantage. It is easy to compare two pieces of string to find which is longer, but only if the strings have similar conformations. If one is straight, and the other tied in knots, it may be very hard. Even if both strings are untangled, then it may still be hard to decide which is longer, if they have very similar lengths. Comparing possible reaction pathways is usually like comparing two pieces of string which are both untangled, or, at least, tangled in much the same way. However, the energy differences between processes may be very small compared with the total energy of the system, and so it may be hard to decide which will be preferred.

Analysing structure, conformation and reactivity means that the mol-

ecules' reactions, or the opponent's move, may reasonably be predicted for each possible reaction, but such a calculation will be very difficult. Even if we assume that this problem is solved, to a sufficient extent for useful answers to be obtained, then the problem of designing a total synthesis is still not complete.

Molecules such as bryostatin are synthesised by joining together small fragments. How many ways can the fragments be joined together? If we assume that we can buy any molecule with four carbon atoms or fewer, which is a crude approximation, bryostatin (Figure 3.3) will require about ten joins, which suggests that there are ten factorial (ten times nine, times eight, times seven, times six, times five, times four, times three, times two, which is about three and a half million) strategies to consider. In practice, the problem is not so straightforward, because many different starting molecules could be considered, and the adjustments between alcohols and ketones, and similar transformations, mean that it is necessary to consider many, many times this number of steps. Two steps for each join might be a more realistic estimate of the number of steps expected, so the number of possible approaches is closer to twenty factorial, which is more than a million million million. Each of these strategies will require the calculation of the outcome of many reactions, as outlined above, and each of these calculations is demanding, by the standards of the fastest computers available today. A complete solution would not be made possible by an increase in computer power of an order of magnitude, nor even by many orders of magnitude.

Several orders of magnitude increase in computer power would be useful to make the calculation of an individual structure rapid, rather than a major project (for molecules of this size). The conformation searching problem then requires that many such calculations are performed. To analyse reactivity many competing reaction processes must be considered in order to determine the best conditions for a particular transformation. Many reagents should be considered for each transformation. There are millions of potential transformations that need to be considered in order to fully analyse competing strategies for synthesis. To complete these calculations in a reasonable amount of time, which is to say, faster than a synthesis could be accomplished by expert organic chemists without all of this computational help, will require much faster computers than are currently available. These calculations will generate an extraordinary quantity of information which will all need to be analysed. Computers are becoming

more powerful very rapidly, but will they become more powerful by a sufficient amount for this problem?

We can obtain a crude estimate the time required for a precise quantum mechanical calculation to analyse possible syntheses of bryostatin. First, the calculation of the energy of a molecule of this size will take hours. Many such calculations will be required to minimise the energy of a structure. A reasonable estimate may be that a thousand energy calculations would be required. Conformation searching will require many such minimisations, perhaps ten thousand. The reactivity of each intermediate will require a harder calculation, perhaps a hundred times harder. Each step will have many possible combinations of reagents, temperatures, times, and so on. This may introduce another factor of a thousand. The number of possible strategies was estimated before as about a million, million, million. In order to reduce the analysis of the synthesis to something which could be done in a coffee break then computers would be required which are 10^{30} times as powerful as those available now. This is before the effects of solvents are introduced into the calculation.

Dr Gordon E. Moore, the co-founder of Intel, the computer chip company, predicted that computers would double in power about every two years, without increasing in price. 'Moore's Law' has held good for almost 30 years. If Moore's law continues to hold true, it will be 200 years before it is possible to analyse a synthesis in a coffee break, and then begin to think about solvents. Moore's law is based on the idea that it will be possible to double the density of components on computer chips every two years. If this is to continue for the next two centuries, it will be necessary to have circuits very much smaller than atoms! It is unlikely that Moore's law will continue to hold for so long. The estimate of the time required is a crude one, and algorithmic advances will undoubtedly play a part in making the problem easier, but it will certainly be a long time before computers can conquer synthesis by brute force.

Can computers, therefore, have any hope of being competing with humans at synthesis, or will people maintain supremacy over machines for the foreseeable future? Fortunately for computers, there is another approach to solving the problem of chemistry. In the introduction to his book, *The Nature of the Chemical Bond* (Pauling 1945), Pauling gives his opinion that it should be possible to describe structural chemistry in a satisfactory manner without the use of advanced mathematics. Books such as this have probably been more influential in the development of modern

chemistry than the direct application of quantum mechanics. A computer may do better to read Pauling than to solve Schrödinger if it wishes to contribute to the development of chemistry.

A huge amount of information has been built up by chemists over the last century which is directly useful for solving new problems in organic synthesis. The difficulty lies in retrieving the right information to help with a specific problem. This may simply be finding one piece of information, for example, a particular reaction which has already been done in a similar way, or it may be finding two or more disparate pieces of information which add together to give a better knowledge of what may happen in a new reaction. The advantage of having a large amount of data at chemists' disposal is also a problem. How can this data be handled effectively?

The textual content of chemistry papers can easily be held in a database, and searched for key words. More sophisticated procedures may also be used, to search for groups of words which tend to appear close to each other, so enabling relevant papers to be discovered. However, chemistry papers are written in many languages and even chemical names are not used consistently. The international language of organic chemistry is structures, such as those drawn in the figures in this article, and these contain more information than can easily be manipulated in words. In the past few years, computers have become sufficiently powerful that an ordinary desktop machine can draw chemical structures, and be used to search a database of structures. This has revolutionised the way that the chemical literature is used. Instead of having to translate a structure to a name, and then search a printed index of chemical names, in order to find references to abstracts of papers, it is possible to sketch the structure, or transformation, of interest and be presented with an abstract, or a diagram, or even the full paper which uses the structure.

Such techniques mean that the chemical literature may be used more effectively, and that its use can be partially automated. Might this lead to a way of automating organic synthesis? To make most molecules there are many strategies which may be successful. If each reaction of each strategy can be evaluated for similarity to a reaction recorded in the literature, it should be possible to develop a route to most molecules by mechanically searching the chemical literature, so that suitable precedent is found for every transformation.

There are two difficulties with this approach. First, there is the problem of performing all of the necessary searches. As discussed above,

there may be billions of possible strategies for making a new molecule. Each reaction in each strategy must be compared with the literature in order to discover if similar reactions have been done before, and if the possible side reactions which could occur are unlikely to do so. This is an enormous task. So much information is available that each search would take a significant length of time. The task is made easier, but less reliable, because much information is not available in a computer readable form. Despite this, the time for each search is significant, and there are a great many to do. It is also possible that a key reaction has been performed in the past, but was not included in the available on-line databases, or else an erroneous result is recorded as if it were true.

Second, it is hard to define similarity in this context. A synthesis may require the transformation of an alcohol to a ketone, and there is ample literature precedent for this. But if there are other alcohols in the molecule, or other groups of atoms which may be affected by the same conditions, it may not be possible to establish this from the literature. If an alcohol is in an unusually crowded position, it may be rather hard to change it into a ketone. Literature precedent may include some crowded alcohols, but nothing quite as crowded, or nothing quite as crowded in the same way. This may be because nobody has tried a similar reaction, or it may be that similar reactions have been tried but found not to work. In the latter case, the unsuccessful result may not have been recorded in the literature.

For both these reasons, a strategy based simply on literature searching is unlikely to be competitive with the best synthetic chemists, who would, of course, use the literature to aid their synthetic designs. It may seem, then, that organic synthesis will remain a skill in which computers cannot compete with humans for some considerable time to come. However, this is not necessarily so.

Information technology enables computers to know the chemistry literature better than any person, but this, in itself, is not sufficient to design syntheses of new compounds. The use of information technology, coupled with methods for the computational analysis of novel reactions, may enable computers to design better syntheses.

The development of the WWW has shown that it is possible for computers to communicate on a global scale, and this, coupled with developments in theoretical chemistry, may lead to computers making useful contributions to synthetic strategies in the near future. The internet has

been growing very rapidly, but it is unlikely to grow without limit. The Cambridge Chemistry WWW server now handles about seventy thousand requests for information each week and has been running for five years. After two years, the growth in use appeared to be approximately exponential, and so it was possible to estimate how the load on the server would increase. Based on just two years of data, the general shape of the following three years of growth was predicted with surprising precision, despite the constant addition of new material and new techniques. When the growth of the internet levels off, the access to this server is also likely to level off. A recent report suggests that this may happen as early as 2003, with around fifty million computers connected together. This suggests that accessible computer power is growing at many times the Moore's law prediction, but it is unlikely to continue to do so for very much longer. It will not give the thirty or so orders of magnitude that are required in order to solve organic synthesis by a brute force approach.

The internet allows the linking of computers which are tuned for database searching (and which may access a world wide database of information, which is not limited by the published literature but also includes research results which are available only on the internet) with computers which are capable of calculating chemical reactivity. It is now easy for me, for example, to do different sorts of literature searches on computers in Bath, Daresbury, and in Manchester, and to analyse the data using computers in Cambridge, all without leaving my office.

The next step would be to allow computers which can calculate chemical properties to interact automatically with computers which can search the chemical literature. This would enable the literature results to be extended to the precise systems of interest for a particular synthesis. If a new alcohol is being oxidised, then the effect of the surroundings could be calculated, while the experimental protocol could be taken from the paper. Thus, the literature results would guide the calculations. The calculations would also guide the literature searching, because the calculation may suggest a side reaction which could be checked in the literature. Literature precedent may be a more reliable guide than calculation as to which of several possible reactions is likely to work best.

It is only just becoming possible to use information technology to routinely search the chemical literature and to do chemical calculations which are directly useful to synthetic chemists. Each of these fields is likely to develop in a powerful way in its own right over the next decades.

However, it is the interaction between these fields which gives the best chance of computers becoming the world's best synthetic chemists.

Chess is not solved, in the way the simple game noughts and crosses is solved, because the outcome of every game is not completely predictable. However, computers will usually win. In the same way, it may not be necessary for computers to analyse all possible routes to a molecule to be best at organic synthesis. It may be enough simply to be successful at finding good routes. This makes the problem much easier, if it is assumed that there are many good routes. The computer would begin by guessing a route, and if it did not work, partially retracing steps, and trying again, thus reusing the information which had already been gathered or calculated so far as possible. Thoroughly exploiting the information that was developed with each potential synthesis would be a crucial step. The time required for conformation searching is dramatically reduced, if similar molecules have already been investigated. For example, PM-toxin has a very complicated potential energy surface, which may be searched directly by traditional methods, or which may be mutated from the conformation search of an alkane, which is easier as it is particularly susceptible to a genetic algorithm based approach. A web page illustrating these results is available: <http://www.ch.cam.ac.uk/MMRG/alkanes/> Some more technical details and a more complete list of references can be found in the Millennium issue of *Philosophical Transactions of the Royal Society* (Goodman 2000).

Will this allow syntheses to be automated? It depends how difficult syntheses are (and will provide a way of quantifying this). It may be that the best possible synthesis is not required, provided that a good route is available, as assessed by total cost (including waste disposal and safety precautions), by time required, by certainty of success, by ease of using robots to follow the procedure, and so on.

Organic synthesis is, and will remain, a very demanding discipline. Brute force methods of calculating new synthetic routes will not be feasible for a very long time, and pure literature based methods will also be very time consuming, and will be restricted by the data available. A hybrid approach provides the best hope for designing a synthetic machine, and it is likely that such programs will become increasingly useful in the new millennium. Most of the elements of these programs are available now, but they are not sufficiently useful that they are an essential part of every chemist's work. An exhaustive solution may not be possible, so it is

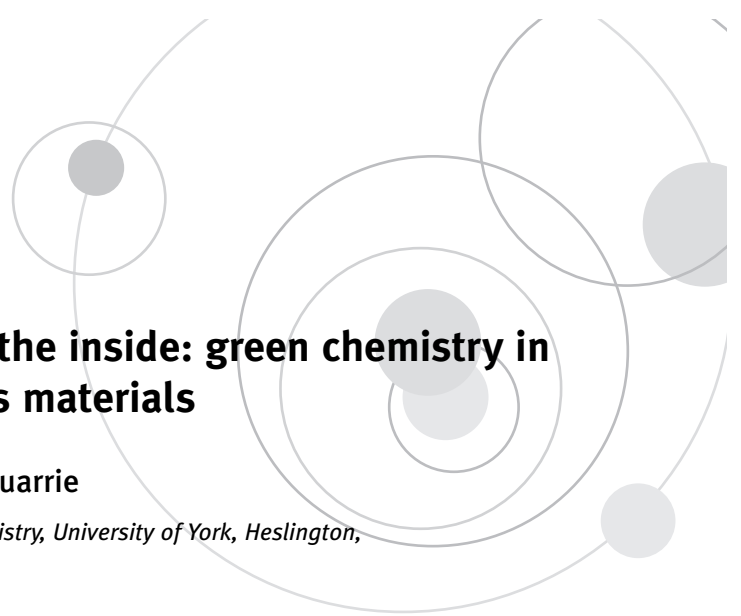
not certain that computers will beat people. However, the odds are stacked in favour of the computer, which will be able to develop and optimise many more routes than it is possible for synthetic chemists to consider directly. How difficult is organic synthesis? We will be taught the answer by the computers which extend the art beyond the heights reached by human scientists.

3.1 Further reading

Dirac, P. A. M. 1929 Quantum mechanics of many-electron systems. *Proc. R. Soc. Lond. A* **123**, 714–733.

Pauling, L. 1945 *The nature of the chemical bond*. Ithaca, New York: Cornell University Press.

Goodman, J. M. 2000 Solutions for chemistry: synthesis of experiment and calculation. *Phil. Trans. R. Soc. Lond. A* **358**, 387–398.



4 Chemistry on the inside: green chemistry in mesoporous materials

Duncan J. Macquarrie

*Department of Chemistry, University of York, Heslington,
York YO10 5DD, UK*

4.1 Green chemistry

The chemical industry today is one of the most important manufacturing industries in the world. The ability of chemists to produce a wide range of different molecules, both simple and staggeringly complex, is very well developed, and nowadays almost anything can be prepared, albeit maybe only on a small scale. On an industrial scale, a great variety of products are synthesised, using chemistry which varies from simple to complex. These products go into almost all the consumer goods we take for granted – colours and fibres for clothes, sports equipment, polymers which go into plastics for e.g. computer and television casings, furnishings, and photographic materials, cleaner fuels, soaps, shampoos, perfumes, and, very importantly, pharmaceuticals. Unfortunately, many of these processes generate a great deal of waste – often more waste is produced than product.

One of the major challenges for chemistry in the opening years of the new millennium is therefore the development of new methods for the clean production of these chemicals. Traditional, so-called end-of-pipe solutions – i.e. treating the waste generated from reactions to render it less harmful – are of limited value in the long term. In the last few years a new, intrinsically more powerful approach has been pioneered. Green chemistry, as it has been called, involves the redesign of chemistry, such that the desired products from a reaction are obtained without generating waste. This massive undertaking involves a wide range of approaches, from the

invention of new reactions to developing new catalysts (chemicals which are themselves not used up in the reaction, but which allow the reaction partners to be transformed more rapidly, using less energy, and often more selectively, generating fewer byproducts) which allow more selective reaction to take place, to biotransformations and novel engineering concepts, all of which can also be used to minimise waste. Catalysts can sometimes be developed which allow inherently clean reactions to be invented.

A very important part of such an undertaking is to be clear about what stages of a chemical process generate the most waste. Often this is found to be the separation stage, after the transformation of reactants to products, where all the various components of the final mixture are separated and purified. Approaches to chemical reactions which help to simplify this step are particularly powerful. Such an approach is exemplified by heterogeneous catalysis. This is an area of chemistry where the catalysts used are typically solids, and the reactants are all in the liquid or gas phase. The catalyst can speed up the reaction, increase the selectivity of the reaction, and then be easily recovered by filtration from the liquid, and reused.

One of the newest areas in the realm of catalysis is that of tailored mesoporous materials, which are finding many uses as highly selective catalysts in a range of applications. A mesoporous material is one which has cavities and channels (pores) in the range of 2–5 nm (a nanometre is 10^{-9} m) – for comparison, a typical chemical bond is of the order of 0.1 nm, and a small organic molecule is around 0.50 nm across. Such mesoporous materials can be thought of as being analogous to the zeolites, which came to prominence in the 1960s. Zeolites are highly structured *microporous* inorganic solids (pores <2 nm), which contain pores of very well defined sizes, in which catalytic groups are situated. A wide range of zeolites is known, each having different pore sizes and channel dimensions. Many are used in large-scale industrial applications. For example, many of the components of petrol are prepared using zeolites, as are precursors for terephthalic acid, used for the manufacture of PET bottles, processes in which millions of tonnes of material is produced annually.

Zeolites are prepared by the linking of basic structural units around a template molecule. The structural units are typically based on oxides of silicon and aluminium, and the templates are usually individual small molecules. Under the right conditions, the silicon and aluminium oxide precursors will link up around the template to form a crystalline three-dimensional matrix containing the template molecules. The template

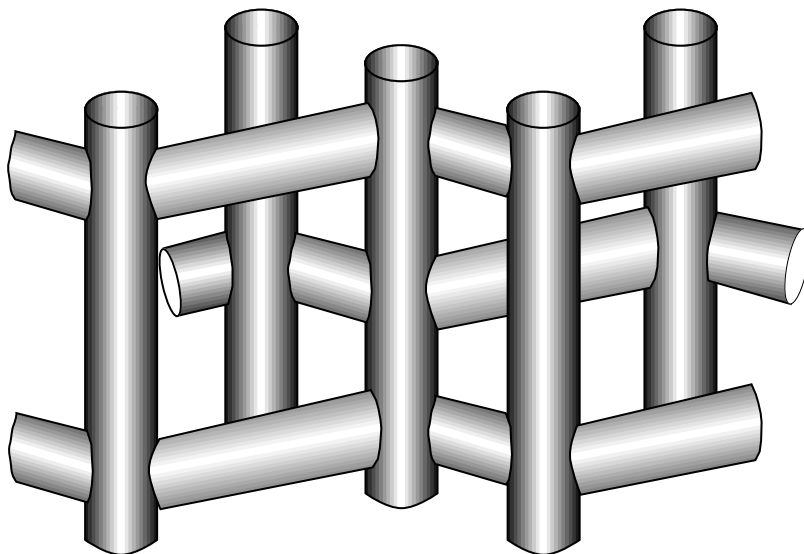


Figure 4.1. Representation of the pore structure of HZSM5, one of the most important zeolites industrially. The vertical cylinders represent one pore network, and the other cylinders an interconnecting network. The narrow pores, and their almost complete uniformity, means that only some molecules can enter. Others are excluded, and cannot react at the active sites, which are found within the structure. Thus, the reactivity of a molecule is determined by its shape and size, rather than by its electronic properties. Such a situation is almost unique, with the only exception being enzymes, where molecules must fit into the enzyme active site in order to react.

molecules can be removed by calcination – i.e. by treatment at high temperatures in air, where the template is effectively burnt out of the structure. This leaves a highly regular structure which has holes where the template molecules used to be. These holes are connected to form pores and cages. It is in these pores and cages, also of very regular size and shape, that the catalytically active groups can be found (Figures 4.1 and 4.2). As we will see, it is this exceptional degree of regularity which is the key to the success of these materials.

Zeolites based on silicon and aluminium are acidic catalysts and are extremely thermally stable. This makes them ideal for use in the petrochemical industry, where some of the largest scale and most high energy transformations are carried out. These transformations are carried out in

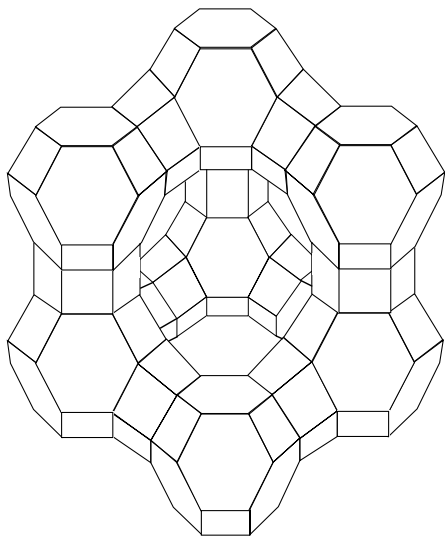


Figure 4.2. The structure of Faujasite, a more open, larger pore zeolite. Larger molecules can enter this structure, which is more open, and slightly less regular than HZSM5 (Figure 4.1). Nevertheless, there are still many important molecules which cannot enter the pores of this zeolite, one of the most accessible of the class.

the gas phase at high temperatures and involve small molecules such as dimethyl benzenes and small alkanes – these are the materials which are used in petrol and PET, as mentioned above. Since the catalytic groups of the zeolite are found within the structure, the molecules must be able to diffuse into the structure before they can react. The size of the pores and channels of the zeolites are designed to be very close to the dimensions of the molecules to be reacted. This means that small changes in size and shape can dramatically alter the ability of the molecule to reach the active site. Under ‘normal’ chemical conditions, molecules react according to their electronic properties – i.e. since the electrons in the molecule must be rearranged during a reaction, their exact positioning and energy within the molecule usually determines both the rate and the nature of the reaction in a given situation. Harsh conditions usually allow many different reactions to take place, and are thus to be avoided if, as is almost always the case, a selective reaction is required. However, in the case of zeolites, the only molecules which can react are those which can fit into the pore structure and get to the active site. Similarly, the only products which can be formed are those which are of the right shape and size to escape from the catalytic sites, migrate through the pores, and out of the catalyst. This phenomenon is known as shape selectivity, although size selectivity might be a more accurate description.

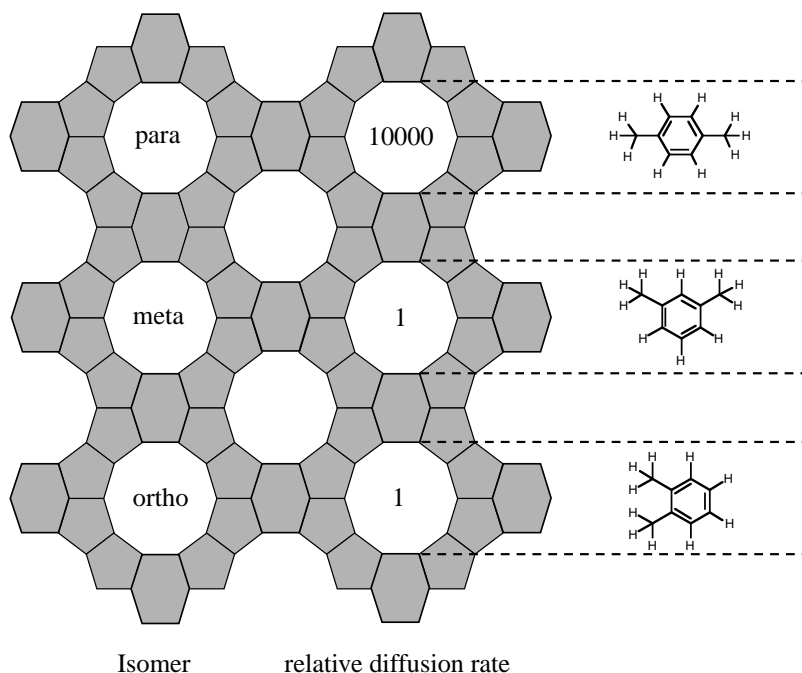


Figure 4.3. Relative diffusion rates in HZSM5. The shaded areas are the pore walls, the unshaded parts the vertical pore system from Figure 4.1. As can be seen, the rate of diffusion varies enormously with only very small changes in molecular size and shape. This allows the zeolite to discriminate almost completely between the three molecules shown, a situation which is unprecedented in traditional, homogeneous chemistry.

An example of this is the commercial process for preparing *para*-xylene, the precursor to terephthalic acid, which is polymerised to give poly(ethylene terephthalate) (PET). In this case, the mixture of xylenes obtained from crude oil is reacted in a zeolite (known as HZSM5). The relative rates of diffusion in and out of the pores are sufficiently different (by a factor of about ten thousand) to allow the extremely efficient and selective conversion of all the isomers to the desired *para* isomer, which is the narrowest and can thus move through the structure most rapidly (Figure 4.3).

This type of selectivity is extremely valuable, as it gives chemists the opportunity to direct reactions in different ways to those available using conventional, electronically controlled, systems. With this in mind, chemists

have searched for many years for materials with the same degree of uniformity displayed by the zeolites, but with larger pores. This would allow the concept of shape selectivity to be extended to larger molecules such as pharmaceutical intermediates, and other highly functional compounds. Other forms of selectivity will also benefit from a very regular structure.

The pore size of most zeolites is <1.5 nm. This microporosity limits their utility in most areas of chemistry, where the molecules used are much larger, and for which mesoporous materials would be necessary. Unfortunately, attempts to use larger template molecules in the zeolite synthesis, an approach which should in theory lead to larger pore size zeolites, have met with very little success. Indeed, some zeolitic materials have been prepared which have mesopores – none of these has ever displayed any real stability and most collapse on attempts to use them. A new methodology was thus required.

4.2 New mesoporous materials

In the past 20 years or so, the field of supramolecular chemistry has become enormously important, with Jean-Marie Lehn, Donald Cram and Charles Pedersen winning the Nobel Prize in 1987. The concept of supramolecular chemistry is that molecules can self-organise into definite structures, without forming covalent bonds, but rather through weaker interactions such as hydrogen bonding. A hydrogen bond is a special type of weak chemical bond, which holds water molecules together, giving water many unique properties – the same bond is critical to the formation of the double helix of DNA, and is often of extreme importance in biological systems. Hydrophobic interactions, also important in self-assembly, are interactions between oily molecules which minimise contact with water by causing the oily parts to huddle together. One example of the latter, although not at all new, is the ability of molecules containing a polar head group and a long non-polar hydrocarbon tail (surfactants) to form micelles in polar, aqueous environments. These micelles form because the water-repelling hydrocarbon tails gather together in the centre of a sphere, or sometimes a cylinder, to avoid contact with water. The polar head groups then form a layer on the surface of the sphere or cylinder, forming a barrier between the hydrocarbon tails and the water. The best-known example of these micelle-forming materials are detergents.

The diameter of the micelles depends on the exact nature of the sur-

factant, but is typically of the order of 2–4 nm. Interestingly, these dimensions are exactly those required for the pores in a mesoporous catalyst. The high profile of supramolecular chemistry helped to highlight such systems, and chemists from Mobil were the first to realise that this chemistry could be applied to catalyst design. Whereas initial approaches to mesoporous zeolites relied on larger and larger *individual* template molecules, Mobil researchers found that they could use supramolecular *assemblies* of molecules as templates. They chose long chain quaternary ammonium salts as the micelle forming agent, and reacted Si and Al precursors around these using conditions similar to those for zeolite manufacture: removal of the template micelle, again by calcination, leaves a solid with pores, where the micelles were.

These materials, known as MTSs (Micelle Templated Silicas) can be prepared with a range of pore sizes (see Figure 4.4). As the pore size is essentially the diameter of the micelle template, it is easy to estimate the pore size obtained with a given template. For example, a MTS made with a dodecyl trialkylammonium (C_{12}) template would have a pore diameter approximately twice the length of the dodecyl trialkylammonium species – roughly 2.2 nm. As the chain length of the template molecules decreases, there comes a point where they do not form micelles. This happens around C_8 , meaning that the smallest pores achievable using this method are around 1.8 nm. Luckily, this is almost ideal in many ways, since the largest zeolites have pore sizes of *c.* 1.3 nm, almost seamlessly extending the range of pore sizes available to the chemist. At the other extreme, as the chain length increases, the ability of the quaternary salt to form micelles decreases, due to lack of solubility, and the largest template molecule which can easily be used is the C_{18} trialkylammonium salt. This gives a pore size of *c.* 3.7 nm. This range of sizes is sufficiently broad to allow ingress and reaction of many large molecules, but the Mobil researchers managed to increase the pore dimensions even further by expanding the micelle. They did this by inserting hydrophobic mesitylene (trimethylbenzene) molecules into the interior of the micelle. The rationale is that the mesitylene molecules will preferentially exist in the hydrocarbon interior of the micelle, rather than in the aqueous environment outside the micelle, causing the micelle to expand (see Figure 4.5).

MTS materials grown using these expanded micelles have pore sizes from 4.0 to 10 nm, depending on the quantity of mesitylene added during synthesis.

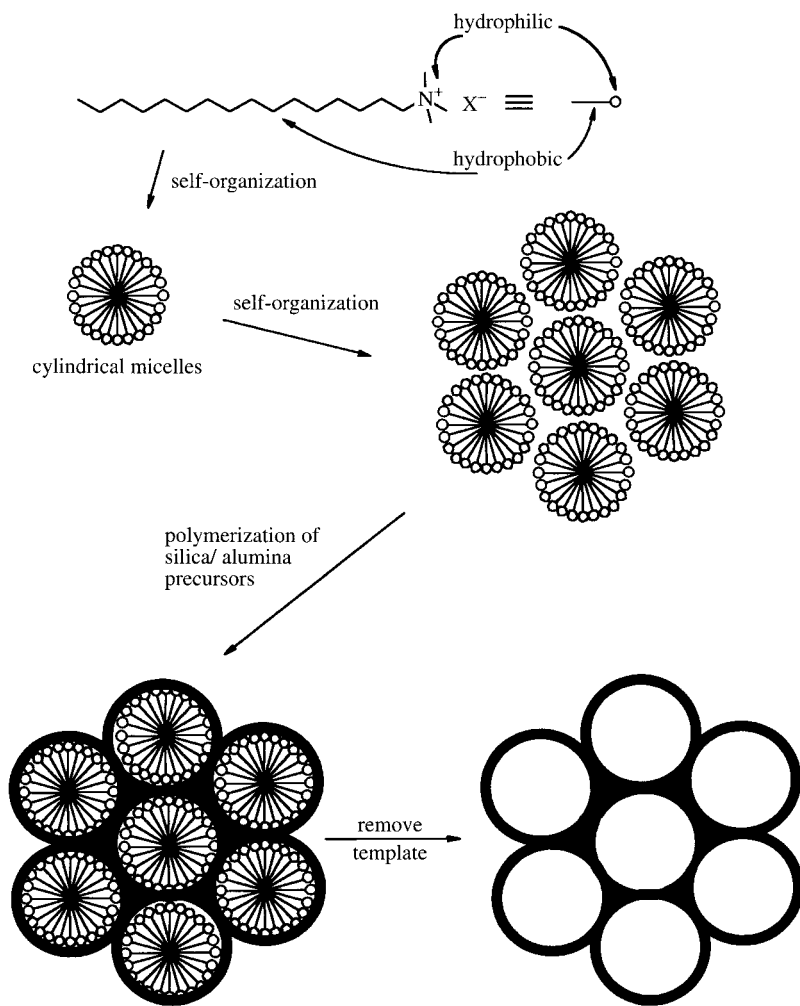


Figure 4.4. Preparation of MTS materials. The diagram shows self assembly of the surfactant into micelles followed by condensation of silica around the micelles. The micelles arrange themselves into an approximately hexagonal array. After the formation of the silica around the micelles, the micelles are burnt out, leaving pores where the micelles were. The pores are an accurate reflection of the size and shape of the micelles. This makes the pores uniformly sized and shaped.

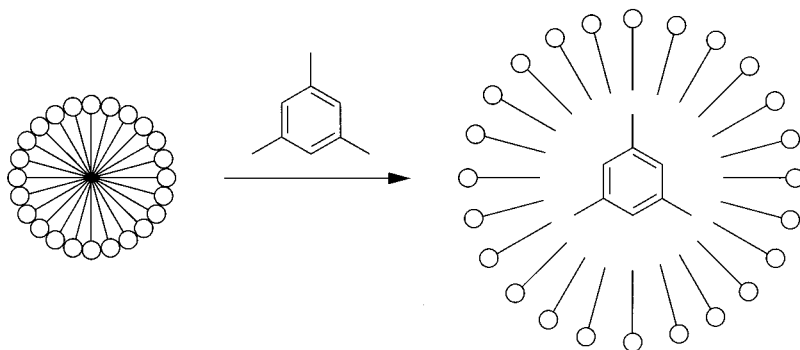


Figure 4.5. Expansion of a micelle by inclusion of a hydrophobic guest into the hydrophobic interior of the micelles. The guest is hydrophobic, and thus does not like being in water. The interior of the micelle is similarly water-repellent, and thus is a much more comfortable environment for the guest. The incorporation of the guest into the centre of the micelle causes an expansion, which in turn leads to larger pores in the resultant material.

A further consideration in porous materials is the shape of the pores. Molecules have to diffuse through the pores to feel the effect of the catalytic groups which exist in the interior and, after reaction, the reaction products must diffuse out. These diffusion processes can often be the slowest step in the reaction sequence, and thus pores which allow rapid diffusion will provide the most active catalysts. It is another feature of the MTSs that they have quite straight, cylindrical pores – ideal for the rapid diffusion of molecules.

One final extension of the original methodology is that different templates can be used to structure the materials. Two of the most useful systems developed were discovered by Tom Pinnavaia of Michigan State University. These methods allow for the complete recovery of template, so that it can be reused, minimising waste in the *preparation* of the materials, and giving a much greater degree of flexibility to the preparation, allowing the incorporation of a great variety of other catalytic groups.

More recently, many workers have concentrated on controlling the size and shape of particles, with an eye on industrial applications, where such features must be well defined and controllable. Many shapes have been made, including fibres, spheres, plates, as well as membranes cast on

surfaces. All these shapes could one day find application, not only in catalysis, but in adsorption of e.g. pollutants from water, molecular wires, and a host of other devices.

4.3 Applications

The initial applications of MTSs were, perhaps not surprisingly, simply attempts to reproduce zeolite chemistry on larger molecules. This chemistry is based on the fact that the aluminium centres in zeolites cause a negative charge to exist on the framework of the solid; this charge must be balanced by a cation. When the cation is a hydrogen ion (proton), the material is an acid, and indeed some zeolites are very strong acids indeed. However, the acidity of the corresponding MTSs is much lower, and initially this limited their applicability somewhat. Nevertheless, the MTSs are often found to be very effective as mild acid catalysts. Much work has therefore been aimed at the production of other materials using the same concept, but with either different templating systems, or with combinations of elements other than Si and Al in the framework.

However, many industrial processes are based on the use of very strong acids, and there is great pressure to find replacements for the liquid acids currently used in industrial processes. One method which has been successfully applied to increase the acidity of these systems is the immobilisation of aluminium chloride onto the pore walls. Aluminium chloride is itself a very strong acid, and is one of the commonest in industrial chemistry. It is used in a wide range of transformations, but cannot be recovered intact from reactions. Its destruction leads to large quantities of waste being generated. Aluminium chloride has been successfully attached to the walls of HMS materials, without any reduction in activity – i.e. the resultant material has the same activity as unsupported aluminium chloride. A major advantage over free aluminium chloride is the ease of removal of the solid catalyst from reaction mixtures, simplifying the process and reducing waste dramatically. The catalyst can then be easily recovered from the reaction mixture, and reused. A second important advantage is the ability to control product distributions by tailoring the pore size of the material. This is best illustrated by the preparation of linear alkyl benzenes (LABs) which are precursors to detergents, and are produced on a massive scale using either aluminium chloride or hydrogen fluoride, both of which have many problems associated with their use. The general scheme is shown in Figure 4.6.

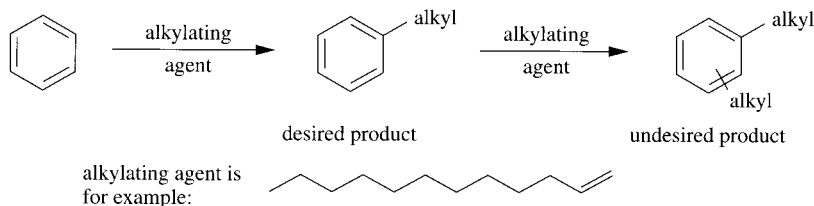


Figure 4.6. General scheme for the synthesis of linear alkyl benzenes, precursors to surfactants. Control over pore size of the catalyst can suppress the second alkylation almost completely. Given the ease with which the pore size can be chosen, one can design an effective catalyst for any particular reaction, and allow the selective and clean production of the desired mono-alkyl product, thus eliminating much of the waste associated with the process.

As can be seen, the reaction will proceed to the monoalkylated product, but does not stop there. The alkylated product is more reactive than the starting material, and will alkylate again, giving products which are useless. Control over this aspect of the reaction can only be achieved with difficulty in traditional systems, and very high dilutions are used to control the product distribution. The use of the new mesoporous materials allows a more concentrated (and thus more efficient) process to be developed. This is because the dialkylated product is bigger than the monoalkylated product. Careful choice of the pore size of the material will mean that the space inside the pore is too small for the dialkylated product to form, but is big enough for the desired monoalkylated product to form readily. Thus, the reaction can run selectively at high concentrations, solving the selectivity problem and using a catalyst which can be easily recovered. Waste is thus reduced dramatically.

While most work has been concentrated on aluminium-containing zeolites, the discovery of titanium-containing zeolites by an Italian company, Enichem, in the 1980s represented another major breakthrough in zeolites. They showed that these titanium-containing zeolites are excellent catalyst for the selective oxidation of a variety of simple, small molecules. Such oxidations are amongst the most important reactions in organic chemistry, as they allow the introduction of a huge range of important functions into the basic hydrocarbon feedstocks derived from oil. Larger pore size versions of the material would allow a much wider range of organic molecules to be functionalised. This type of reaction is of enormous importance in large molecule chemistry too, with some

existing processes being far from 'green'. Thus researchers have been active in preparing analogous MTS structures containing titanium. Results with these MTS materials have shown that these materials are indeed capable of carrying out many of the desired reactions, but without the limitations of size which hamper the zeolites. For example, one of the important applications of titanium-containing zeolites is the hydroxylation of benzene to phenol (which is used as a precursor to antioxidants for food and cosmetic use), and then further to hydroquinone, used in the photography industry as a developer. Ti containing MTSs are known and can carry out the same type of transformations as the corresponding zeolite. Larger molecules such as naphthalenes, which cannot enter the pores of the zeolites, *can* access the pores of the MTSs, and react in the expected manner. One important target is Vitamin K3, a derivative of naphthalene, formed by hydroxylation. Current practice still involves the use of acidic chromium reagents which are used in large quantities, and are highly toxic. Significant success has been reported with the use of Ti containing mesoporous materials of this reaction type, and further progress is expected (see Figure 4.7).

The organically modified versions of these materials have also been investigated as catalysts. These materials have great potential, as the incorporation of organic groups will allow a much wider variety of materials to be prepared, and thus a much wider range of applications can be investigated. Simple amine (an example of a basic group, which will remove a proton from a molecule, thus making it able to react in many ways) con-

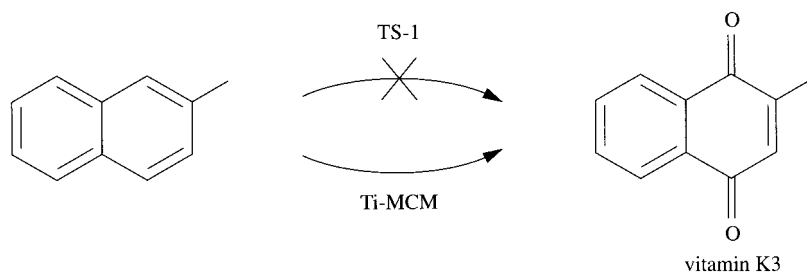


Figure 4.7. Possibilities for the synthesis of Vitamin K3. The small pore titanium zeolite TS-1 cannot fit the large naphthalene molecule into its pore system, and thus is ineffective in this transformation. The larger titanium MTS material is capable of interacting with the molecule, and the desired transformation can take place.

taining materials have been prepared by three different routes. These materials are capable of the fast and efficient catalysis of several reactions with excellent selectivity. Activity is greater than that found with the traditional amine-containing silicas, as is the stability of the catalyst, allowing more product to be prepared with a given amount of catalyst. The increased amount of amine groups which can be attached to the MTS materials gives them even more of an advantage over the traditional catalysts. Initial results on a catalyst with both amine and phenyl (non-polar) groups indicate a substantial rate increase over the simple amine-only material. The reasons for this are not yet understood, but may be due to improved transport of reagents and products onto and off the surface. Many important reactions can be carried out with such solid bases, and their uses in chemistry will increase. In particular, many reactions which either do not generate any side products or only generate water (condensation reactions) are amenable to catalysis using these materials. Early work on such systems indicates that the future for these materials is very rosy.

Sulphur-containing materials have been found to be excellent adsorbents for heavy metals. The sulphur atom is known to complex strongly to heavy metal ions, with gold and mercury being two particularly interesting examples. The higher amounts of sulphur which can be attached to the MTS materials means that their capacity for binding these metals, and removing them from e.g. drinking water, is much greater than that achieved with traditional materials.

Solid acids can also be prepared from these materials by transformation of the sulphur group to the sulphonic acid, very closely related to sulphuric acid, one of the most commonly used acids industrially. The material can be easily recovered and easily handled; since the acidity resides within pores, it cannot come into contact with living tissue. Important transformations, such as the formation of synthetic lubricants and intermediates for fragrances, have already been reported using these materials. The scope for such materials in future is enormous.

More sophisticated materials have been made by attachment of transition metal complexes to the surface. These materials are designed to enhance the fundamental activity of the metal ion, by providing it with an environment tailored to make it as active as possible, and to aid in its recovery afterwards. The heterogenisation of such (normally homogeneous) complexes has attracted a lot of attention, since the heterogeneous equivalents can be much more easily separated and recycled than the

homogeneous ones, leading to much less waste being produced. These materials have been shown to be very active in a range of reactions, leading to many important product types. One particularly important area of chemistry is the selective preparation of one of a pair of mirror images of a compound. This so-called chiral (from Greek; *chiros* – hand) catalysis requires great control over the exact chemical environment of the catalytic site, and is one of the major challenges in synthetic chemistry. Many drugs and agrochemicals can exist as two forms which are mirror images of one another, only one of which is useful, the other being useless or even dangerous. It is therefore important to be able to prepare only the desired form. As an example of mesoporous materials containing chiral metal-centred catalysts, the group of Daniel Brunel in Montpellier has published work on transformations using zinc species. Selectivity to the desired form was good, approaching that achievable with conventional systems. Further refinement of these systems will lead to improvements in the design of the catalytic site, and its surrounds, and the prospects for this area of catalysis are exciting.

4.4 Future prospects

The chemical industry produces an enormous range of products, from petrol and other fuels, to additives which improve the performance of the fuels, to plastics and fabrics (including the colours and fire retardants which make them attractive and safe), colours, flavours and fragrances, and further to the most complex molecules, which find use as agrochemicals and pharmaceuticals. One of the strongest current trends in the industry is towards green chemistry, which will involve redesign of many of these processes for the preparation of this bewildering array of products. Much success has already been achieved, and many major products are now produced using green technologies. Much remains to be done, however, and several approaches are currently being investigated. One of the most exciting is the development of new materials which can function as catalysts, and whose structures can be fine tuned for the application in mind. The rate of the advances made in the last eight years of the twentieth century has been remarkable, and further advances will allow these fascinating materials to contribute greatly to the quality of life for everyone in the twenty-first century.

The future of these designer materials is very exciting indeed. Further

work will reveal advanced catalytic systems, possibly containing more than one type of active site, and the control over pore dimensions will allow an ever-increasing level of control over selectivity towards the desired product. The ability to incorporate polarity-modifying groups will also play a major role in transport processes, of great importance in both catalysis and membrane processes.

Many other opportunities exist due to the enormous flexibility of the preparative method, and the ability to incorporate many different species. Very recently, a great deal of work has been published concerning methods of producing these materials with specific physical forms, such as spheres, discs and fibres. Such possibilities will pave the way to new application areas such as molecular wires, where the silica fibre acts as an insulator, and the inside of the pore is filled with a metal or indeed a conducting polymer, such that nanoscale wires and electronic devices can be fabricated. Initial work on the production of highly porous electrodes has already been successfully carried out, and the extension to uni-directional bundles of wires will no doubt soon follow.

The ability to produce threads, discs and spheres of defined size and structure will be of great importance when the very promising initial results from catalytic studies are applied on a larger scale. Processes using heterogeneous catalysts require the ability to control particle size and shape in order to ensure good mixing of all the reaction components, and separations after reaction.

A further application of this technology will certainly be the fabrication of membranes of these materials. Membrane reactors have shown great utility in many systems, where one component of a reaction mixture can be separated by permeation through a membrane, thus driving a reaction forwards, by continuous separation. Such continuous processes can themselves save a great deal of waste.

Looking further ahead, the pores in these materials could be considered as analogous to ion channels in cell walls. If a hollow sphere of MTS could be fabricated with e.g. an enzyme (or other cell component) inside, one could imagine this as being an 'inorganic cell'. The encapsulation of the enzyme inside the cell could then possibly be used to protect the enzyme from harsh conditions outside the cell, while allowing reaction components to diffuse in, react, and diffuse out again. Already, some effort is being expended on silica/biological composites, with significant advances being made. Given the enormous strides made since the

discovery of the MTSs in 1992, such major advances will no doubt become reality in the early years of the twenty-first century.

4.5 Further reading

- Anastas, P. T. & Williamson, T. C. (eds.) 1998 *Green chemistry – frontiers in benign chemical syntheses and processes*. Oxford University Press.
- Clark, J. H. (ed.) 1995 *Chemistry of waste minimisation*. Glasgow: Blackie.
- Corma, A. 1997 From microporous to mesoporous molecular sieve materials and their use in catalysis. *Chem. Rev.* **97**, 2373.
- Gates, B. C. 1992 *Catalytic chemistry*. New York: Wiley.
- Macquarrie, D. J. 2000 Chemistry of the inside: green chemistry in mesoporous materials. *Phil. Trans. R. Soc. Lond. A* **358**, 419.



5

Diamond thin films: a twenty-first century material

Paul W. May

School of Chemistry, University of Bristol, Bristol BS8 1TS, UK

Diamond has some of the most extreme physical properties of any material, yet its practical use in science or engineering has been limited due its scarcity and expense. With the recent development of techniques for depositing thin films of diamond on a variety of substrate materials, we now have the ability to exploit these superlative properties in many new and exciting applications. In this paper, we shall explain the basic science and technology underlying the chemical vapour deposition of diamond thin films, and show how this is leading to the development of diamond as a twenty-first century engineering material.

5.1 The diamond in history

Probably more so than any other gemstone, diamonds feature more predominantly in the history and cultural heritage of the human race. They were prized for their scarcity for centuries, and still remain a symbol of wealth and prestige to this day. The word diamond comes from the Greek *adamas*, meaning indestructible. Diamonds were first mined in India over 4000 years ago, but the modern diamond era only began in 1866, when huge diamond deposits were discovered in Kimberley, South Africa, creating a huge rush of European prospectors. The wealth this created helped to underwrite the British Empire, and changed the fates of many African countries.

Apart from their appeal as gemstones, diamonds possess a remarkable

Table 5.1. *Some of the outstanding properties of diamond*

-
-
- Hardest known material giving extreme wear resistance
 - Highest bulk modulus, i.e. stiffest material
 - Least compressible
 - Highest room temperature thermal conductivity
 - Extremely low thermal expansion at room temperature
 - Broad optical transparency from the deep ultraviolet to the far infrared
 - Highest speed of sound
 - Very good electrical insulator
 - Diamond can become a wide band gap semiconductor
 - Very resistant to chemical corrosion
 - Biologically compatible
 - Some surfaces exhibit very low or 'negative' electron affinity
-
-

range of physical properties. Indeed, a glance at any compendium of material data properties will prove that diamond is almost always 'the biggest and best'. A selection of some of these is given in Table 5.1. Amongst other properties, diamond is the hardest known material, has the highest thermal conductivity at room temperature, is transparent over a very wide wavelength range, is the stiffest material, the least compressible, and is inert to most chemical reagents. With such a wide range of exceptional properties, it is not surprising that diamond has sometimes been referred to as 'the ultimate engineering material'.

Unfortunately, it has proved very difficult to exploit these properties, due both to the cost and scarcity of large natural diamonds, and the fact that diamond was only available in the form of stones or grit. It had been known for 200 years that diamond is composed solely of carbon, and many attempts were made to synthesise diamond artificially using as a starting material another commonly occurring form of carbon, graphite. This proved extremely difficult, mainly because at room temperature and pressure, graphite is more stable than diamond. Although the difference in stability between the two forms of carbon is actually quite small, their structures are so different that it would require a large amount of energy to convert between them. Ironically, this large energy barrier which makes diamond so rare is also responsible for its existence, since diamond, once

formed, cannot spontaneously convert to the more stable graphite phase. Diamonds are, indeed, forever!

To overcome these problems, researchers realised that in order to form diamond, they had to choose conditions where diamond, and not graphite, is the more stable phase. The knowledge of the conditions under which natural diamond is formed deep underground, suggested that diamond could be formed by heating carbon under extreme pressure. This process forms the basis of the so-called high-pressure high-temperature growth technique, first marketed by General Electric, and which has been used to produce 'industrial diamond' for several decades. In this process, graphite is compressed in a hydraulic press to tens of thousands of atmospheres, heated to over 2000 °C in the presence of a suitable metal catalyst, and left until diamond crystallises. The diamond crystals this produces are used for a wide range of industrial processes which utilise the hardness and wear resistance properties of diamond, such as cutting and machining mechanical components, and for polishing and grinding of optics. However, the drawback of this method is that it still produces diamond in the form of single crystals ranging in size from nanometres to millimetres, and this limits the range of applications for which it can be used. What is required is a method to produce diamond in a form which can allow many more of its superlative properties to be exploited – in other words, diamond in the form of a thin film.

5.2 Chemical vapour deposition

Rather than try to duplicate Nature's method, diamond could conceivably be produced if carbon atoms could be added one-at-a-time to an initial template, in such a way that a tetrahedrally bonded carbon network results (see Figure 5.1). These ideas led to experiments in which carbon-containing gases were heated under reduced pressure until the molecules broke apart, and then these fragments were condensed onto a nearby surface. Analysis showed that the thin film that resulted from this did, indeed, contain diamond. However, the rate of growth in these early experiments was low, and the films were impure, containing a large proportion of unwanted graphite. The breakthrough came in the late 1960s, when researchers in the USA discovered that the presence of atomic hydrogen during the deposition process would remove graphite from a surface much faster than diamond. This meant that the impure components were removed from the

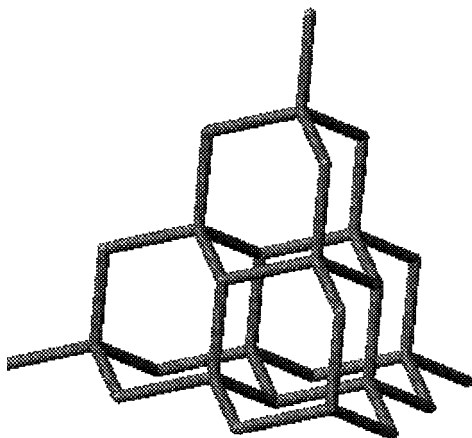


Figure 5.1. In diamond, every carbon atom is bonded to four others in a strong, rigid tetrahedral structure.

growing film, leaving only pure diamond behind. This process became known as '*chemical vapour deposition*' (CVD), since it involves a chemical reaction occurring within a vapour over a surface, leading to deposition of a thin coating onto that surface. Over the next few years more breakthroughs were made which allowed diamond films to be grown at significant rates on many useful materials. This series of discoveries stimulated world-wide interest in diamond CVD, in both academia and industry, which continues to the present day.

5.3 Methods for production of CVD diamond

All CVD techniques for producing diamond films require a means of 'activating' gas phase carbon-containing precursor molecules. This activation can involve heating (e.g. a hot filament), or an electric discharge, such as a plasma. Figure 5.2 illustrates two of the more popular experimental methods. While each method differs in detail, they all share a number of features in common. For example, growth of diamond (rather than graphite) normally requires that the precursor gas (usually methane, CH_4) is diluted in excess of hydrogen – typically the mixing ratio is 1 per cent methane to 99 per cent hydrogen. Also, the temperature of the substrate is usually greater than 700°C in order to ensure the formation of diamond rather than amorphous carbon.

Hot Filament CVD (see Figure 5.2(a)) is relatively cheap and easy to operate and produces reasonable quality polycrystalline diamond films at

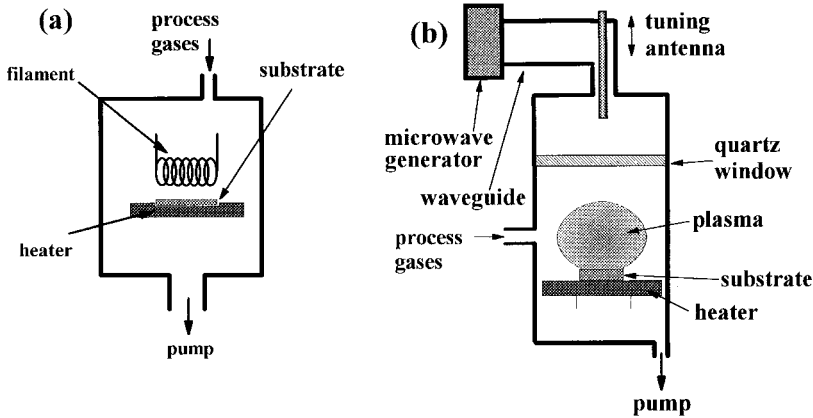


Figure 5.2. Two of the more common types of low pressure CVD reactor. (a) Hot Filament Reactor – these utilise a continually pumped vacuum chamber, while process gases are metered in at carefully controlled rates (typically a total flow rate of a few hundred cubic centimetres per minute). Throttle valves maintain the pressure in the chamber at typically 20–30 torr, while a heater is used to bring the substrate up to a temperature of 700–900°C. The substrate to be coated – e.g. a piece of silicon or molybdenum – sits on the heater, a few millimetres beneath a tungsten filament, which is electrically heated to temperatures in excess of 2200°C. (b) Microwave Plasma Reactor – in these systems, microwave power is coupled into the process gases via an antenna pointing into the chamber. The size of the chamber is altered by a sliding barrier to achieve maximum microwave power transfer, which results in a ball of hot, ionised gas (a plasma ball) sitting on top of the heated substrate, onto which the diamond film is deposited.

a rate of around 1–10 μm per hour, depending upon exact deposition conditions (1 μm is one thousandth of a millimetre). However, it also suffers from a number of major disadvantages. The hot filament is particularly sensitive to oxidising or corrosive gases, and this limits the variety of gas mixtures which can be employed. It is also very difficult to avoid contamination of the diamond film with filament material. For diamond to be used in mechanical applications, metallic impurities at the tens of parts per million level are not a significant problem, but it becomes unacceptable for electronic applications.

Microwave Plasma CVD reactors use very similar conditions to hot filament reactors, and despite being significantly more expensive, are now among the most widely used techniques for diamond growth. In these

systems, microwave power is coupled into the chamber in order to create a discharge or plasma. This leads to heating and fragmentation of the gas molecules, resulting in diamond deposition onto a substrate which is immersed in the plasma. The most common type of microwave reactor in use is shown in Figure 5.2(b). Nowadays, microwave powers of up to 60 kW can be utilised in such systems giving growth rates well in excess of 0.1 mm per hour. As well as high powers and hence higher growth rates, other advantages of microwave systems over other types of reactors are that they can use a wide variety of gas mixtures, including mixtures with high oxygen content, or ones composed of chlorine- or fluorine-containing gases. The fact that no filament is involved makes microwave systems inherently cleaner than hot filament systems, and so they have become the system of choice for making diamond for electronic applications.

A number of other deposition methods have been used for growing diamond, with varying degrees of success. These include oxyacetylene welding torches, arc jets and plasma torches, laser ablation and liquid phase crystallisation, but none of these yet realistically compete with the hot filament or microwave systems for reliability and reproducibility.

5.4 The chemistry of CVD diamond growth

The complex chemical and physical processes which occur during diamond CVD are comprised of a number of different but inter-related features, and are illustrated in Figure 5.3. At first sight, this may seem like a daunting array of physical and chemical reactions which need to be grasped if diamond CVD is to be understood. But over the past 10 years there have been a large number of studies of the gas phase chemistry, and we are now beginning to obtain a clearer picture of the important principles involved. The first clue was that diamond growth appeared to be independent of the chemical nature of the gas phase precursors – it was only the total number of carbons, hydrogens and oxygens in the reactant molecules that mattered. This meant that the gas phase chemistry is so rapid that it simply and effectively breaks down the constituent gases to smaller, reactive components.

It is now believed that the most critical component in the gas phase mixture is atomic hydrogen, and indeed, this reactive atom drives the whole chemical system. Two hydrogen atoms are made when a hydrogen molecule (H_2) splits apart. In a hot filament system, the thermal energy

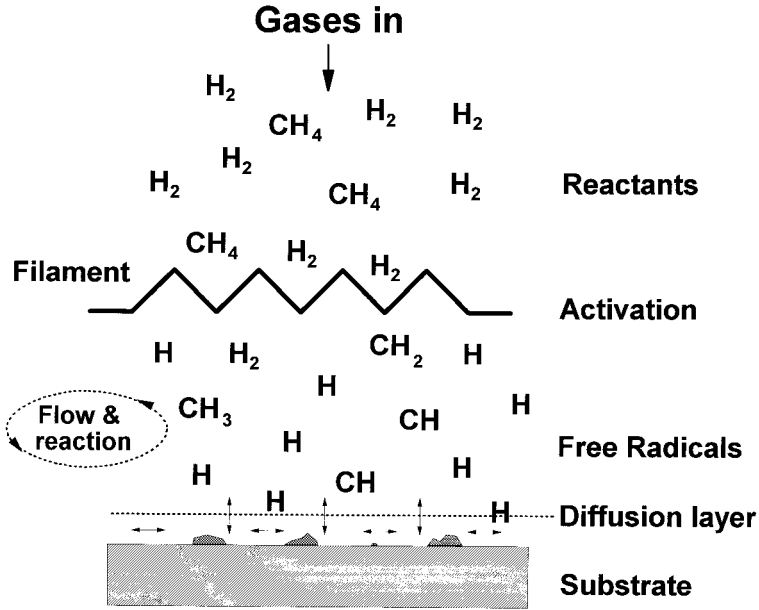


Figure 5.3. Schematic representation of the physical and chemical processes occurring during diamond growth. The process gases first mix in the chamber before diffusing toward the substrate surface. En route, they pass through an activation region (e.g. a hot filament or electric discharge), which provides energy to the gaseous species. This activation causes molecules to fragment into reactive radicals and atoms, creates ions and electrons, and heats the gas up to temperatures approaching a few thousand degrees Celsius. Beyond the activation region, these reactive fragments continue to mix and undergo a complex set of chemical reactions until they strike and stick to the substrate surface. At this point, the species can either react with the surface, escape again back into the gas phase, or diffuse around close to the surface until an appropriate reaction site is found. If a surface reaction occurs, one possible outcome, if all the conditions are suitable, is diamond.

from the hot filament surface is enough to do this, whereas in a plasma system, the H_2 molecules are broken apart as a result of impacts by high energy electrons. The resulting high concentration of atomic hydrogen is crucial for a number of main processes.

- (i) Although in bulk diamond the carbon atoms are all fully tetrahedrally bonded (see Figure 5.1), at the surface there is effectively a 'dangling

bond', which needs to be terminated in some way. If too many dangling bonds are left unterminated, they will tend to join together (cross-link), and the surface structure will begin to resemble that of graphite. The vital surface termination is normally performed by hydrogen which attaches to the dangling bond and thereby keeps the tetrahedral diamond structure stable. During diamond growth, some of these surface hydrogen atoms need to be removed and replaced by carbon-containing species. A large number of reactive hydrogen atoms close to the surface can quickly bond to any excess dangling bonds, so preventing surface graphitisation.

- (ii) Atomic hydrogen is known to etch graphite-like carbon many times faster than diamond-like carbon. Thus, the hydrogen atoms serve to remove back to the gas phase any graphite-like clusters that may form on the surface, while leaving the diamond clusters behind. Diamond growth could thus be considered as 'five steps forward, but four steps back', with the net result being a (slow) build up of diamond.
- (iii) Hydrogen atoms are efficient scavengers of long chained hydrocarbons, breaking them up into smaller pieces. This prevents the build up of polymers or large ring structures in the gas phase, which might ultimately deposit onto the growing surface and inhibit diamond growth.
- (iv) Hydrogen atoms react with hydrocarbons such as methane (CH_4) to create reactive radicals such as methyl (CH_3) which can then attach to suitable surface sites.

There have been many suggestions for the identity of the diamond growth species, however, the general consensus is now that the bulk of the evidence supports CH_3 as being the important radical. The basic picture which emerges for CVD diamond growth is believed to be as follows. During growth, the diamond surface is nearly fully saturated with hydrogen. This coverage limits the number of sites where hydrocarbon species (probably CH_3) may stick. A schematic illustration of the resulting processes is shown in Figure 5.4, suggesting that diamond growth can be considered to be a one-by-one addition of carbon atoms to the existing diamond structure, driven by the presence of reactive atomic hydrogen. In oxygen-containing gas mixtures, it is believed that the hydroxyl (OH) radical plays a similar role to atomic hydrogen, except that it is more effective at removing graphitic carbon, leading to higher growth rates and better quality films at lower temperatures.

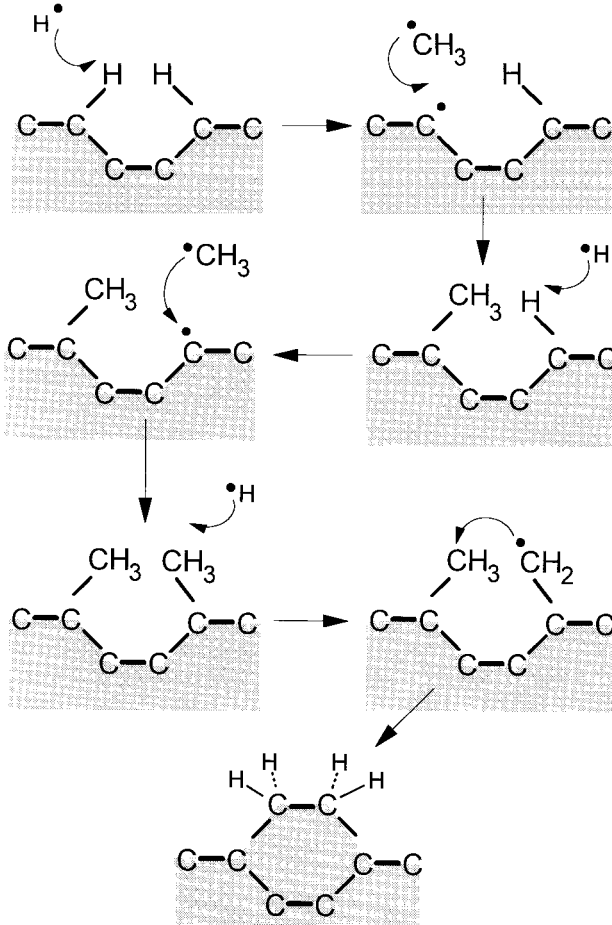


Figure 5.4. A schematic illustration of the reaction process occurring at the diamond surface. Atomic hydrogen removes a surface hydrogen to form an H_2 molecule, leaving behind a reactive surface site (illustrated by the dot). The most likely fate for this surface site is for it to react with another nearby hydrogen atom, returning the surface to its previous stable situation. However, occasionally a gas phase CH_3 radical can collide and react with the surface site, effectively adding one carbon to the structure. This process of hydrogen removal and methyl addition may then occur on a site adjacent to the attached methyl. Further hydrogen removal reactions will lead to completion of the ring structure, locking the two carbons into the diamond structure. Thus, diamond growth can be considered to be a one-by-one addition of carbon atoms to the existing diamond structure, catalysed by the presence of excess atomic hydrogen.

5.5 The substrate material

Most of the CVD diamond films reported to date have been grown on single crystal silicon wafers, mainly due to its availability, low cost, and favourable properties. However, this is by no means the only possible substrate material – although any candidates for diamond growth must satisfy a number of important criteria. One requirement is obvious – the substrate must have a melting point higher than the temperature required for diamond growth (normally $>700^{\circ}\text{C}$). This precludes the use of existing CVD techniques to coat low-melting point materials, like plastics, aluminium, some glasses, and electronic materials such as gallium arsenide.

Another criterion is that the substrate material should expand by the same amount as diamond when heated. This is because at the high growth temperatures currently used, a substrate will tend to expand, and thus the diamond coating will be grown upon, and bonded directly to, an expanded substrate. Upon cooling, the substrate will contract back to its room temperature size, whereas the diamond coating will be relatively unaffected by the temperature change. Thus, the diamond film will experience significant compressive stresses from the shrinking substrate, leading to bowing of the sample, and/or cracking and flaking of the entire film.

Another issue is that at the high deposition temperatures many substrate materials react with carbon directly to form a carbide. The presence of a thin carbide layer is not a problem – in fact it is desirable, since the carbide layer can be pictured as the ‘glue’ which aids the adhesion of the diamond layer by (partial) relief of stresses at the interface. Without this carbide glue, any diamond layer will not adhere well to the surface, and the films will often readily delaminate after deposition. This can be utilised as one method to make free-standing diamond films, using non-carbide-forming substrate materials such as copper, tin, silver and gold. Conversely, if the substrate material is too reactive toward carbon, then the deposited carbon (even when it’s in the form of diamond) simply dissolves into the surface forming a solid solution. This can result in large quantities of carbon being transported into the bulk, rather than remaining at the surface where it can promote diamond growth. Metals where this is significant include titanium, nickel and iron. The latter metal is of particular concern, because this means that at present all industrially important ferrous materials (such as iron and stainless steel) cannot be diamond coated using simple CVD methods.

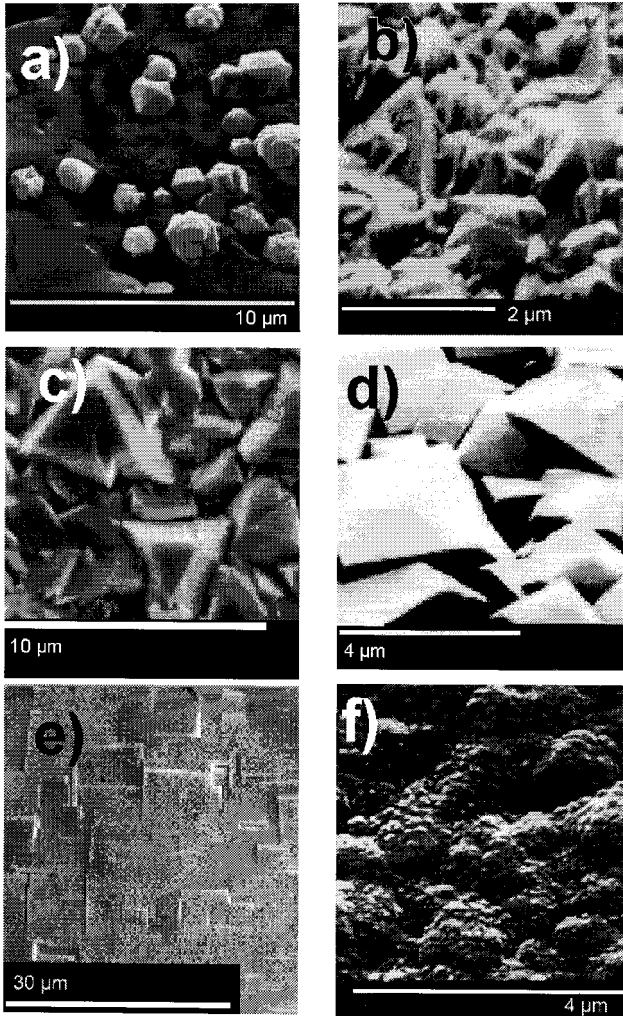


Figure 5.5. Electron micrographs of different types of diamond film grown on silicon. The white bar shows the scale in micrometres (μm) (thousandths of a millimetre). (a) The initial stages of diamond growth on a nickel substrate, showing individual diamond crystallites nucleating in scratches and crevices created on the surface by mechanical abrasion, (b) a randomly oriented film, (c) a triangular-faceted film, (d) a square-faceted film, (e) a diamond film showing highly aligned crystals, and (f) a nanocrystalline diamond film grown by using high methane concentration. (Figures 5.5(b)–(e) are reproduced with permission of Professor John Wilson, Heriot Watt University, UK.)

It is apparent that many of the problems with expansion mismatch and carbon solubility could be eliminated if the deposition were to occur at much lower temperatures. Many groups world-wide are focusing their research efforts in this direction, and the answer may lie in different gas chemistries, such as use of carbon dioxide or halogen containing gas mixtures. Until then, the difficulties associated with diamond growth on problematic materials have ensured the continuing popularity of silicon as a substrate material. It has a sufficiently high melting point (1410°C), it forms only a localised carbide layer (a few atoms thick), and it expands relatively little upon heating. Molybdenum and tungsten display similar qualities, and so are also widely used as substrate materials. They can also be used as barrier layers – thin coatings deposited upon certain of the more problematic substrate materials to allow subsequent diamond CVD.

5.6 Nucleation

Nucleation is the process whereby gas phase carbon atoms join together on a surface to make the beginnings of a new crystal structure. Growth of diamond begins when individual carbon atoms nucleate onto the surface in the specific diamond-like tetrahedral structure. When using natural diamond substrates (a process called *homoepitaxial* growth), the template for the required tetrahedral structure is already present, and the diamond structure is just extended atom-by-atom as deposition proceeds. But for non-diamond substrates (*heteroepitaxial* growth), there is no such template for the carbon atoms to follow, and those carbon atoms that deposit in non-diamond forms are immediately etched back into the gas phase by reaction with atomic hydrogen. As a result, the initial induction period before which diamond starts to grow can be prohibitively long (hours or even days). To combat this problem, the substrate surface often undergoes a pre-treatment prior to deposition in order to reduce the induction time for nucleation and to increase the density of nucleation sites. This pre-treatment can involve a number of different processes. The simplest is abrasion of the substrate surface by mechanical polishing using diamond grit ranging in size from 10nm to 10µm. It is believed that such polishing aids nucleation by either (a) creating appropriately-shaped scratches in the surface which act as growth templates, or (b) embedding nanometre-sized fragments of diamond into the surface which then act as seed crystals, or (c) a combination of both. An example is given in Figure 5.5(a), which

shows the initial stages of nucleation, with individual diamond crystallites growing in scratches on the surface. Another, better-controlled version of this is to use ultrasonic agitation to abrade the substrate immersed in a slurry of diamond grit in water. Whatever the abrasion method, however, the need to damage the surface in such a poorly-defined manner prior to deposition may severely inhibit the use of diamond for applications in, say, the electronics industry, where circuit geometries are frequently on a sub-micron scale.

5.7 The CVD diamond film

Once individual diamond crystallites have nucleated on the surface, growth proceeds in three dimensions until the isolated crystals meet their neighbours and coalesce. At this point a continuous film is formed, and the only way growth can then proceed is upwards. The resulting film is 'polycrystalline' with many grain boundaries and defects, and exhibits a columnar structure extending upward from the substrate. Furthermore, as the film becomes thicker, the crystal size increases whilst the number of defects and grain boundaries decreases. This means that the outer layers of thicker films are often of much better quality than the initial nucleating layers. For this reason, if the diamond film is to be used as a heat spreader or optical window (applications where good quality and small number of grain boundaries are paramount), the film is often separated from its substrate and the lower 50–100 μm are removed by mechanical polishing.

The surface morphology of the diamond film obtained during CVD depends critically upon the various process conditions, especially the gas mixing ratio. Depending upon the ratio of methane to hydrogen, the film can be randomly oriented (Figure 5.5(b)) or have some degree of preferred orientation, such as triangular- (Figure 5.5(c)) or square-faceted films (Figure 5.5(d)). By employing growth conditions which favour one particular orientation, highly textured films can be produced which are very closely aligned to the structure of the underlying substrate (Figure 5.5(e)). The ultimate aim, for electronic applications, is to produce diamond films which are essentially single crystal, but although a number of groups have recently made significant progress, this goal still hasn't been achieved. With increasing methane concentrations, the crystal sizes decrease, until above about 3 per cent CH_4 in H_2 the crystalline morphology disappears altogether (see Figure 5.5(f)). Such a film is referred to as 'nanocrystalline'

or 'ballas' diamond, and may be considered to be an aggregate of diamond nanocrystals and disordered graphite. Although this type of film might be considered inferior to the more crystalline and therefore better quality diamond films, it still possesses many of the desirable properties of diamond while being much smoother and considerably faster to deposit. Thus, by the simple expedient of changing the growth conditions, films can be deposited with properties ranging from almost graphitic to essentially those of natural diamond. This allows the quality, appearance and properties of a diamond film, as well as its growth rate and cost, to be easily tailored to suit particular applications. With the advent of high power microwave deposition systems, it is now possible to produce CVD diamond films over areas up to 8 inches in diameter and of thicknesses exceeding 1 mm (see Figure 5.6).

5.8 Applications

The applications for which CVD diamond films can be used are closely related to the various extreme physical properties it exhibits. Some of these applications are already beginning to find their way into the marketplace; however, some, including some of the more sophisticated electronic applications, are still a number of years away. Until recently, the main issue preventing the wide-scale use of CVD diamond has been economic – the coatings were simply too expensive compared to existing alternatives. However, as higher power deposition reactors become standard, the cost for 1 carat (0.2 g) of CVD diamond fell below US\$1 in the year 2000, making the use of CVD diamond much more economically viable, and finally allowing engineers the opportunity to exploit its vast array of outstanding properties in a wide variety of different applications.

5.8.1 Cutting tools

The extreme hardness of diamond, coupled to its wear resistance, makes it an ideal candidate for use in cutting tools for machining non-ferrous metals, plastics, chip-board and composite materials. Indeed, industrial diamond has been used for this purpose since the 1960s, and remains a lucrative commercial process today. This involves either gluing the diamond grit to a suitable tool (saw blades, lathe tools, drill bits), or consolidating the diamond grit with a suitable binder phase (e.g. cobalt or silicon carbide) to make a hard, tough and durable composite. CVD

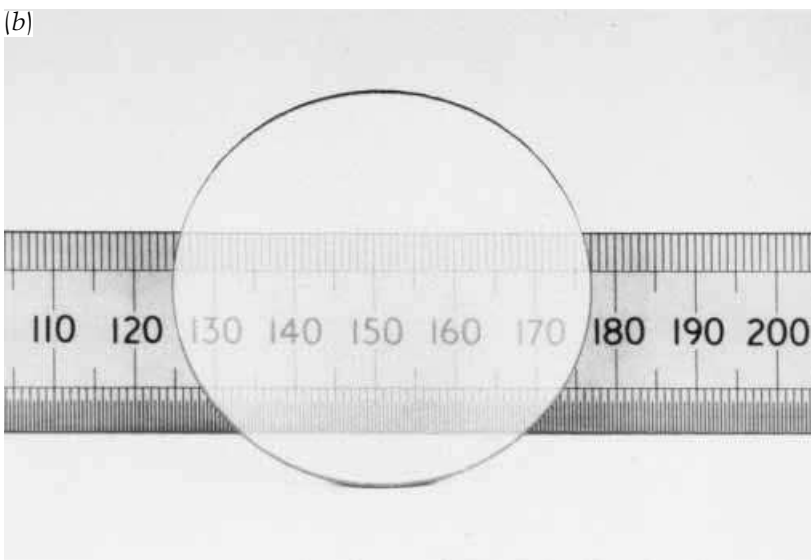
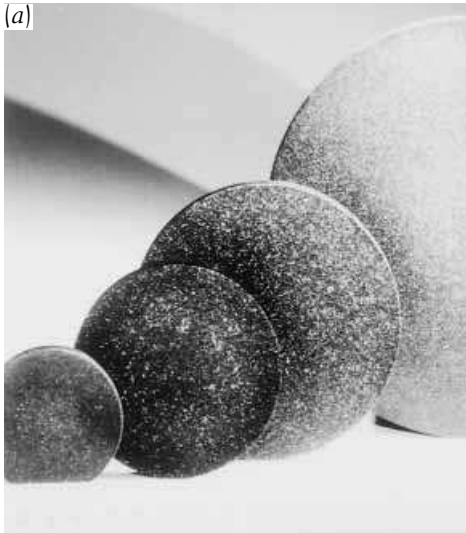


Figure 5.6. (a) 2–6 inch diameter diamond wafers on silicon, and (b) optically transparent free-standing diamond window (reproduced with permission of Dr Christoph Wild, Fraunhofer Institut für Angewandte Festkörperphysik, Freiburg, Germany).

diamond is beginning to be used in a similar way, by coating the diamond directly onto the surface of the tungsten carbide tool pieces. Initial tests indicate that such CVD diamond-coated tools have a longer life, cut faster, and provide a better finish than conventional tungsten carbide tools. However, the term non-ferrous should be emphasised here, since this highlights one disadvantage that diamond has over other tool materials – it reacts with iron at high temperatures, and so cannot be used to cut ferrous materials such as steel. However, some of the newer composite metals that are beginning to be used in the aerospace and automobile industries, such as aluminium–silicon alloys, are excellent candidates for diamond-coated cutting tools, as they are very difficult to machine with conventional materials.

5.8.2 Thermal management

Modern high power electronic and opto-electronic devices suffer severe cooling problems due to the production of large amounts of heat in a small area. In order to cool these devices, it is essential to spread this heat by placing a layer of high thermal conductivity between the device and the cooling system (such as a radiator, fan, or heat sink). CVD diamond has a thermal conductivity that is far superior to copper over a wide temperature range, plus it has the advantage of being an electrical insulator. Now that large area CVD diamond plates with very high thermal conductivities are available, this material is beginning to be used for a variety of thermal management applications. For example, using CVD diamond heat spreaders to cool microchips should result in higher reliability and higher speed operation, since devices can be packed more tightly without over-heating.

5.8.3 Optics

Because of its optical properties, diamond is beginning to find uses in optical components, particularly as a free-standing plate for use as an infrared window in harsh environments. Conventional infrared materials (such as zinc sulphide, zinc selenide, and germanium), suffer the disadvantage of being brittle and easily damaged. Diamond, with its high transparency, durability and resistance to thermal shock, is an ideal material for such applications. An example of an optical quality diamond window can be seen in Figure 5.6(b).

5.8.4 Electronic devices

The possibility of adding impurities to diamond (a process called *doping*) and so changing it from being an electrical insulator to a semiconductor, opens up a whole range of potential electronic applications. However, there are a number of major problems which still need to be overcome if diamond-based devices are to be achieved. First, CVD diamond films are polycrystalline containing many grain boundaries and other defects, which all reduce electrical conductivity. For effective device operation, single crystal diamond films are required, and this still hasn't been achieved. Another problem, which to some extent has recently been solved, is the requirement that the diamond films must be patterned to produce features of similar size to those used in microcircuitry, typically a few microns. Fortunately, diamond can be etched in oxygen-based plasmas, provided a suitable non-erodible mask is used. Diamond films can now be patterned to geometries suitable for all but the most demanding devices. The final, and probably the most difficult problem to solve in order to be able to create diamond devices, is that of doping – changing the conductivity of the diamond reliably and reproducibly by incorporation of suitable impurity atoms. Unfortunately, most electronic devices require two different types of impurities to be present, ones that lead to an excess of positive charge, and ones that lead to an excess of negative charge. Creating excess positive charge (*p*-type doping) is relatively straightforward, since addition of a small amount of a boron-containing gas such as diborane to the CVD process gas mixture is all that is required to incorporate boron into the structure. However, the close packing and rigidity of the diamond structure make incorporation of atoms larger than carbon very difficult. This means that impurities (such as phosphorus or arsenic) which are routinely used to create excess negative charge (*n*-type doping) in other semiconductor materials like silicon, cannot easily be used for diamond. The development of a successful *n*-type doping process has taken a considerable time, and only very recently have a few reports appeared from Japan claiming success in this area using sulphur as the necessary impurity. Despite these difficulties, diamond-based devices are gradually beginning to appear, and it may become the material of choice for electronic applications involving high power and/or high temperature.

5.8.5 Field emission displays

Another device which can utilise polycrystalline CVD diamond, and which is causing a great deal of interest at the moment, is the idea of using diamond as an electron emitter in flat panel displays. The electronic properties of diamond are such that when a negative voltage is applied across it in vacuum, electrons are ejected from its surface. This process is also common in most metals, except that in metals the electrons have to overcome an energy barrier, or work function, to escape from the surface. In diamond this barrier has been measured and found to be very small, maybe even negative, and this has given rise to the term 'negative electron affinity'. In practice, this means that devices based on the electron emission properties of diamond could consume very low power levels and hence be extremely efficient. The electrons emitted from the surface are accelerated using a positive grid, and strike a phosphor screen, causing light to be emitted. Each emitting diamond crystal, or group of crystals, would form a 'pixel' on a flat panel display screen. Unlike their major competitors (liquid crystal displays), diamond cold cathode field emission displays would have high brightness, have a large viewing angle, and be insensitive to temperature variations. Also, because of their relative simplicity, it is possible that diamond emitting displays could be scaled up to large areas that would be impossible with liquid crystals – maybe even metres square!

5.8.6 Electrochemical sensors

CVD diamond films can be used for electrochemical applications, especially in harsh or corrosive environments. Conducting diamond electrodes, made by adding boron to CVD diamond films, are very inert compared to other electrode materials (such as platinum). Such diamond electrodes may find applications in analysis of contaminants, such as nitrates, in water supplies, and even in the removal of those contaminants.

5.8.7 Composite reinforcement

Diamond fibres and wires have been fabricated (see Figure 5.7), which show exceptional stiffness for their weight. If growth rates can be increased to economically viable levels, such diamond fibres may find uses as reinforcing agents in advanced composites, allowing stronger, stiffer and lighter load-bearing structures to be manufactured for use in, say, aerospace applications. Hollow diamond fibres and two-dimensional diamond fibre mat-

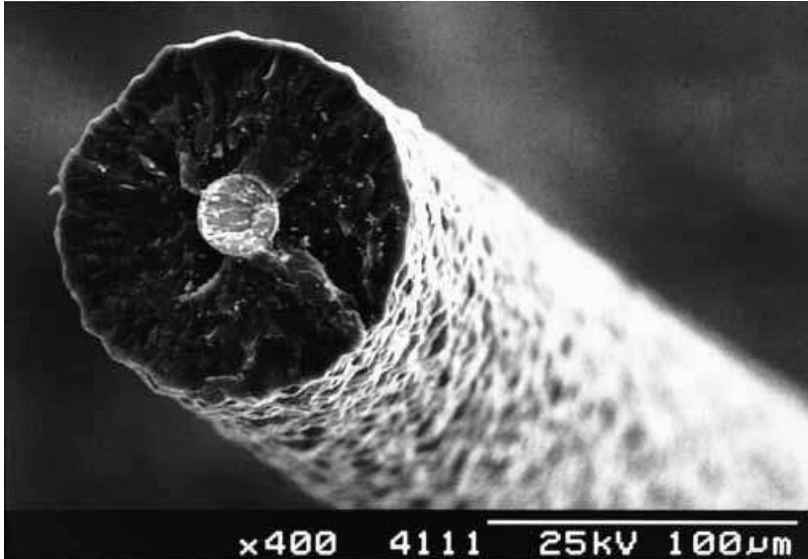


Figure 5.7. A diamond-coated tungsten wire that is about the same diameter as a human hair.

tings have also been demonstrated, and could form the basis of smart composite structures.

5.8.8 Particle detectors

One area where CVD diamond is beginning to find a market is as a detector for ultraviolet light and high energy particles. High performance ultraviolet detectors have already been demonstrated and are now in commercial production. Diamond can be used to detect other high energy particles (such as alpha- and beta-particles and neutrons), and be used as a replacement for silicon in the next generation of particle accelerators. Since diamond has a similar response to damage by X-rays and gamma rays as human tissue, a possible application is in medical applications, as a dosimeter for radiation exposure.

5.9 Summary

Despite the rapid progress made in the past 10 years in the science and technology behind diamond film CVD, the commercialisation of this

amazing material is still in its infancy. Researchers and industry are currently concentrating upon developing methods to scale up the CVD processes and reduce production costs to the point at which it becomes economically viable to use diamond as the material of choice. With the twenty-first century now upon us, we are still some way from diamond becoming the engineer's dream of being 'the ultimate engineering material'. However, some applications are already in the marketplace, such as diamond heat spreaders, windows and cutting tools. In the next few years we can expect to see diamond films appearing in many more applications, especially in electronics. Perhaps the most likely 'killer applications' which will firmly establish diamond as a twenty-first century material will be in the area of specialised flat panel displays and high temperature electronics, for which the total available market in the year 2000 has been estimated at US\$435 million and US\$16 billion, respectively. In some ways this may be a shame, since familiarity with diamond as just another commonplace material may remove some of the glamour and mystique surrounding the world's most sought-after gemstone.

The author thanks the Royal Society for funding. He also thanks Professor John Wilson (Heriot Watt University) and Dr Christoph Wild (Fraunhofer Institut für Angewandte Festkörperphysik, Freiburg, Germany) for giving permission to reproduce their figures and photographs.

5.10 Further reading

- Dischler, B. & Wild, C. (eds.) 1998 *Low-pressure synthetic diamond*. Berlin: Springer. This book goes into more detail about the technical aspects of making CVD diamond.
- May, P. W. 2000 Diamond thin films: a 21st-century material. *Phil. Trans. R. Soc. Lond. A*, **358**, 473–495. This gives a much more thorough and detailed scientific account of the subject.
- Spear, K. E. & Dismukes, J. P. 1994 *Synthetic diamond: emerging CVD science and technology*. New York: Wiley. This book gives a useful description of the chemistry and physics behind diamond CVD, as well as various novel applications for CVD diamond.
- Ward, F. 1998 *Diamonds*. Bethesda, MD, USA: Gem Book Publishers. A compact book with many photographs telling the history of diamond gemstones.



6

The secret of Nature's microscopic patterns

Alan R. Hemsley¹ and Peter C. Griffiths²

¹ *Department of Earth Sciences, Cardiff University, PO Box 914, Cardiff CF10 3YE, UK*

² *Department of Chemistry, Cardiff University, PO Box 912, Cardiff CF10 3TB, UK*

There is little doubt that the information encoded in the genes of living things has a great impact on their ultimate form. Dogs, daisies and diatoms (Figure 6.1(a)) are what they are, largely because they have a set of genes that, working in combination, code for the production of such things as fur, flowers or frustules. However, working in tandem with the genetic code is a diversity of mechanisms which cannot be mapped to any gene, but which contribute much to the production of structure, architecture and pattern. The existence of such mechanisms is in part obvious. As explained by Cohen, the imaginary changeling-like introduction of fly DNA into the egg of a chicken would produce neither fly nor chicken since fly information and chicken constructional mechanisms would be largely incompatible. A fly needs fly construction mechanisms while the constructional apparatus in a chicken's egg cannot use fly information to make a chicken.

6.1 The biology of microarchitecture and self-assembly

6.1.1 Message and machinery

Man-made structures and architecture operate under similar constraints. Three factors come together to produce the final object. There is a design in the form of a blueprint, the workforce to manipulate the components, and the components themselves whose physical properties also play a role in determining the ultimate form. One cannot build a car engine from rubber or Wellington boots from steel. Classical Greek architecture

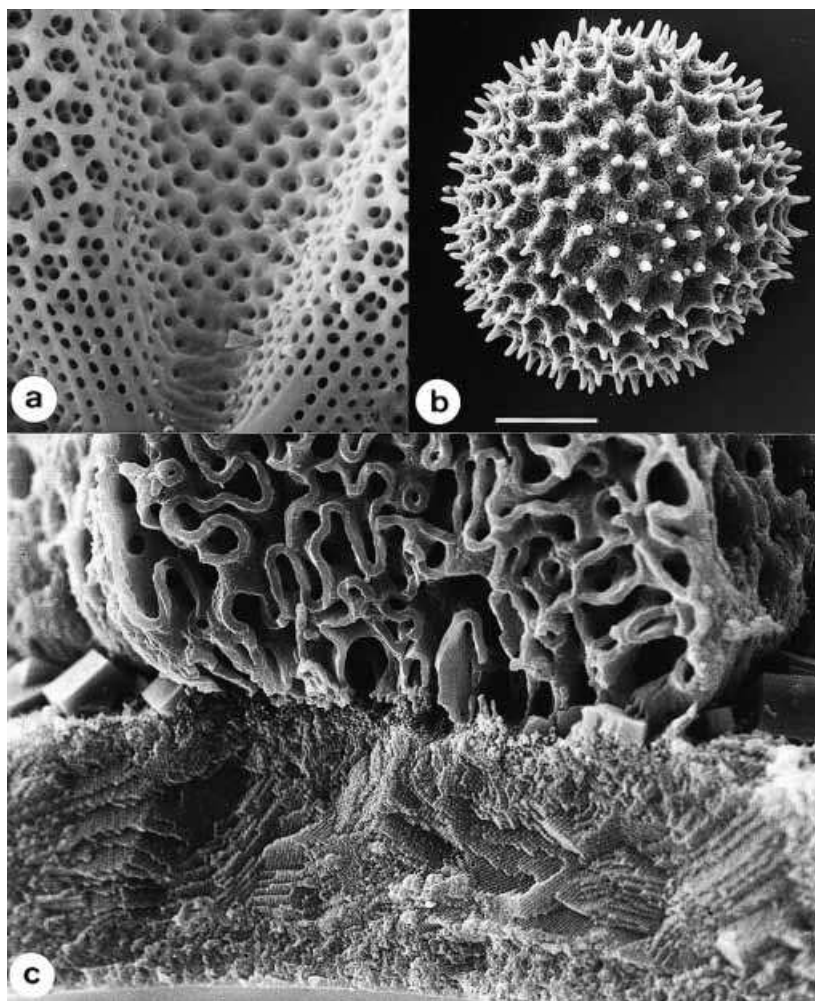


Figure 6.1. All scales refer to bar in (b). (a) The silica frustule (shell) of the colonial diatom *Actinoptycus*. Microscopic algae may use silica, calcium carbonate, or organic polymers to produce such shells. Scale = 10 μm . (b) Pollen grains of the Morning Glory, *Ipomoea indica*, have a complex surface pattern. Such patterns are common among spores and pollen, but how do they arise? Scale = 40 μm . (c) A broken section of a spore wall from *Selaginella myosurus* (a type of club moss) showing complex internal structure including a colloidal crystal region composed of numerous tiny spherical particles. It is spore walls such as these that have led botanists to consider colloids as fundamental in the production of complex wall patterns.

embodies these principles and has provided structures which were functional and have stood the test of time. Furthermore, the Greeks were aware of some fundamental patterns in nature. Their architects recognised the intrinsic aesthetic value of the 'golden ratio' (1:1.618) which is derived from adjacent numbers in the Fibonacci series. The same mathematical series governs many space-filling operations in nature, seen most obviously in the arrangement of scales in a pine cone or of seeds on a sunflower head.

The DNA (our blueprint) gives rise to proteins (commonly our components) by converting the genetic code into a sequence of linked amino acid units. The proteins roll up in a specific (self-assembling) way governed by the interactions of the side chains. Some, by a long history of chance and evolutionary selection, behave as efficient catalysts (enzymes) to bring about the formation of other types of molecule from the same simple components. Others break apart molecules releasing energy to power these processes. The self-assembly of biological molecules and synthetic analogues has received some attention from biochemists, but exactly how does an organism progress from such a molecular cocktail to something with a spinal column, a stem or a complex silica shell? What is the workforce that operates to achieve construction from our genetic blueprint?

6.1.2 The inertia of natural patterns

In his inspiring work *On growth and form*, D'Arcy Thompson saw that the production of many relatively small scale biological structures such as radiolarian skeletons and the spiral shells of many marine organisms resulted from packing phenomena (as in pine cones or sunflowers) upon surfaces or in three dimensions. Today his work is perhaps seen as being overly directed to the description of nature by 'natural' mathematical rules, very much in the Greek tradition. However, the nub of his argument still has great merit; rules do apply in development and, as expounded by Kauffman, they are those of biophysics and chemistry acting at the interfaces of components derived from the molecular soup within cells. Further, it is the interaction between cells so constructed and constrained that gives rise to the varied shapes of multicellular organisms, including ourselves. Nonetheless, it is at the scale of single-celled organisms that the mechanisms of self-assembly are most apparent and close observation of the often spectacular architecture displayed at this level, should give clues to the nature of these mechanisms. These interactions, as noted by Thompson,

occur at the colloidal dimension. Given this connection, it is surprising that there are few studies attempting to correlate architecture and colloid chemistry.

Proteins are not the only structures within cells to adopt a particular form dependent upon the intrinsic characteristics of their components. Self-assembly has been demonstrated in microtubules; cell components built from proteins that act like tug boats and guide large components to the interaction sites. Their various conformations are a result of concentration specific self-assembly processes. Similarly, the form taken by membranes is governed by the concentration of the components, the nature of the surrounding fluids, and physical parameters such as temperature. The formation of periodic minimal surfaces and other bicontinuous structures may be an inherent consequence, as seen in the prolamellar bodies of chloroplasts in plants. In both cases, the genetic code need not define all possible conformations, merely the required concentration of the components in order to initiate the 'desired' structure. It is perhaps noteworthy that the formation of complex membrane systems, and indeed the positioning of the structural units, is often aided by microtubules presenting clear evidence of a hierarchy of developmental self-organisation and assembly.

Microorganisms may produce complex microscopic architecture involving inorganic components. Common amongst these additions are calcium and silica. Small, golden-brown algae produce surface discs of calcium carbonate (coccoliths) which can resemble miniature car hub caps. These structures, although small, are the principal component of the White Cliffs of Dover, having accumulated for millennia upon a Cretaceous sea bed. The siliceous frustules (shells) of diatoms (Figure 6.1(a)) enclose the single-celled alga in much the same way as a petri dish; one larger half, overlapping the edges of the smaller. Like the coccoliths, and many other microstructures, these shells are composed of networks of bars, ridges, pores and spines. Siliceous architecture also occurs on the surface of some higher plant spores (Figure 6.2(a)) and has been shown to have a colloidal origin.

6.1.3 Mimicking and modelling nature

The production of artificial microscopic structures with similar architecture to that produced by microorganisms has been pioneered by Stephen Mann. As in our experiments (below), the production of microstructure

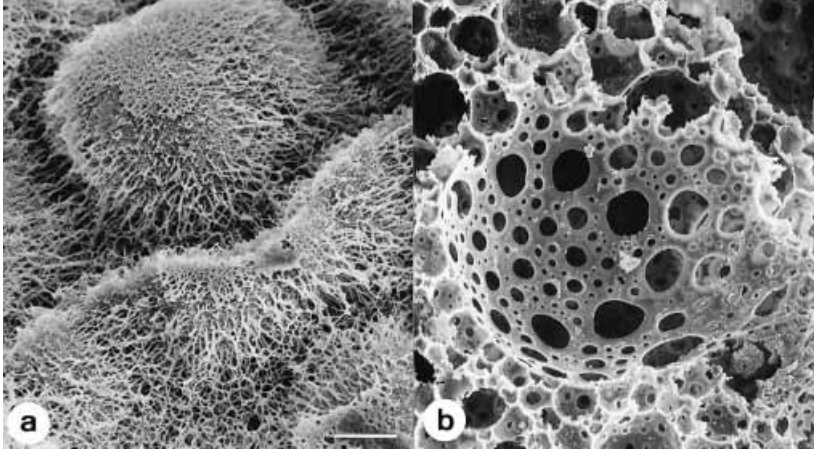


Figure 6.2. (a). Colloidal silica network on the surface of spores from *Isoetes pantii* (quill wort). Scale = 20 μm . (b). Polystyrene networks and foams produced as a biproduct of colloidal latex formation. Both types of colloidal system are typical of the diversity of patterns that can be derived from the interactions of minute particles. Scale (in (a)) = 50 μm .

relies on the behaviour of the component which will form the structure (in Mann's case calcium bicarbonate) in a bicontinuous mixture of oil-water-surfactant. We concur with the views of Mann and Ozin that complex three-dimensional surfaces (such as that of the prolamellar body) provide a potential template for the accumulation of more robust structural units, be they inorganic or organic.

In the case of diatom frustules, foam-like aggregations adjacent to the surface membrane of the organism restrict the deposition of the mineral phase. This is the self-assembling aspect of pattern formation. What is less clear (and probably more directly under genetic influence) is how consistency of form is maintained within a species and how different forms are produced by different species. This is not a problem restricted to mineral microarchitecture. The organic (sporopollenin) surfaces of spores and pollen (Figure 6.1(b)) all seem to have species-unique patterning, of great use to taxonomists working with both living and extinct plants. These very different microarchitectures can only arise through slight modifications in the building process – the question that needs addressing is how?

Flexibility of pattern formation may well be the consequence of self-assembly mechanisms acting upon 'digital' information such as that

contained within DNA. The nature of proteins is such that a single base change in the genetic sequence can code for a different amino acid which, in turn can give rise to a different molecular configuration. A different protein within a construction sequence will have no effect (possibly by coding for the same amino acid), cause it to fail, or occasionally cause it to produce something different that does something useful within the organism. The nature of such mechanisms is essentially chaotic in that they exhibit both robustness and fragility. The substitution of many amino acids within a protein need not significantly change its folding pattern if these are chosen with care with respect to the influence they have on folding (robustness). However, the substitution of any one critical amino acid will cause the adoption of a different configuration (fragility). Alongside such potential generators of robust difference are so-called 'antichaotic' factors as proposed by Kauffman. In antichaotic systems, the great complexity of components within the cellular soup are seen to be fully interactive with each other. These systems can be perturbed, but are in a sense self-generating and in a state of balance. Such systems act to maintain this equilibrium but if distorted to excess, will 'snap' to an alternative stable state.

It is against this background that we have been investigating the structure and development of spores from the club moss *Selaginella*. These show complex microscopic architecture within their relatively thick walls (Figure 6.1(c)). The presence of an apparently colloidal crystal region within the wall, which consists of more or less spherical particles of sporopollenin, has been determined. This has focused attention on constructional mechanisms involving processes of colloidal interaction in order to account for the crystalline region and the other structures encountered within the complex walls. It has become apparent that a full understanding of this mode of microarchitectural construction lies as much with an appreciation of colloid and surfactant interactions as it does with 'biological' control mechanisms.

6.2 Consideration of colloidal interactions and self-assembly

6.2.1 The unexpected behaviour of tiny objects

Whilst our understanding of the relevant factors important to colloid science in terms of synthetic applications and materials (e.g. paints) is quite advanced, as we have seen the same cannot be said for the 'colloid

science' operating in natural environments. This is surprising since many of the same types of materials are present in the natural environment, e.g. spherical particles comprising silica or a monomer, free polymer, salt and other additives such as fatty acids. Furthermore, the synthetic colloid scientist can manipulate the components within a system in ways that are not accessible to nature, i.e. there is unlikely to be a genetic mechanism that can suddenly add 0.2g of polymer or increase the ionic strength to 0.1 M! Genetic input is simply not responsive enough in relation to the speed of reactions. However, Nature is a far better chemist than man – although she has had many more millennia to get it right – and discovering the finesse and natural controlling factors would certainly enhance the ability of the relatively crude synthetic chemist. By analogy to the chaotic systems proposed previously, Nature may prepare systems at the boundary of stability and through subtle changes in one parameter, tip the system over the edge resulting in significant architectural changes. The approach taken in our work has been to try to manipulate the behaviour of synthetic organic colloids with a view to reproducing patterns and architecture present in the natural materials; this will *inter alia* uncover the controlling factors used by nature. Utilisation of organic components in synthetic biological self-assembly is new and presents complexity of interpretation. However, it is essential if we are to progress beyond qualitative description to quantitative and defined understanding.

First though, we must outline albeit very briefly, the basic factors important to colloidal stability and self-assembly. It is these areas that clearly hold the insights we require. Throughout the section, we highlight possible control mechanisms available to the natural system.

The Greeks also believed that only two forces – love and hate – could account for all fundamental phenomena. There are in reality four distinct forces; the strong nuclear interactions that bind nuclei together, weak interactions associated with electron clouds and the two forces the Greeks 'missed', electrostatic and gravitational forces. In actual fact, the Greeks *did* observe these latter two interactions but could not explain them. In the seventeenth century, Newton showed that the interaction between molecules within an ensemble affected their bulk physical properties. Phenomena such as capillary rise – the way water creeps up the sides of a very thin glass tube – led to the suggestion that different glass/liquid and liquid/liquid interactions must exist. It was the Dutch scientist van der Waals who made the breakthrough; in order to explain why gases do not

obey the ideal gas law, van der Waals introduced a force (which now bears his name) to account for an attractive interaction between molecules. However, it was not until the advent of quantum theory in the 1920s and the ability to elucidate the electronic structure of molecules, that it became clear that all intermolecular interactions are in fact, electrostatic in origin. Today, intermolecular forces can be calculated from a knowledge of the distribution of electron clouds associated with the molecules.

The characteristics of colloidal particles, as described by Shaw, are somewhat different to those of a molecule, yet the same basic forces operate. The generalised interaction between identical spherical colloid particles dispersed in a solvent depends on the nature of the particles and the solvent and varies with the distance between the particles. Interestingly, and independent of the nature of the particles, it turns out that there is always an attractive interaction between such identical particles dispersed in a solution. This attractive interaction tends to induce aggregation and thus, colloidal dispersions are inherently thermodynamically unstable. If an organism can synthesise a colloidal dispersion, either through aggregation of dissolved minerals or polymerisation of self-assembled molecules, the formation of the colloidal crystals such as those present in some spore walls (Figure 6.1(c)) should come as no surprise! It is this very potential, i.e. to form aggregates rather than dispersions, that organisms have used to great effect.

This simple thermodynamic picture is substantially altered if we introduce dissimilar particles into our dispersion. The various interactions now depend on the nature of the two particles, relative to the solvent, and can either favour dispersal or aggregation. Again, this could be the basis for a natural control mechanism; as the number and composition of the colloidal building blocks evolve, subtle changes in the interactions could switch a dispersion from stable to unstable.

The overall interaction between colloidal particles in solution sometimes includes two further terms, an electrostatic term arising through the presence of charged groups on the surface of the particle or a steric term resulting from the presence of polymers adsorbed onto the surface of the particles. Several mechanisms lead to surface charge – dissociation of ionic groups, adsorption/desorption of potential determining ions and other ionic materials such as surfactants. The presence of surface charges induces a re-distribution of nearby ions; like-charges are repelled and unlike-charges attracted. Combined with their thermal motion, this leads

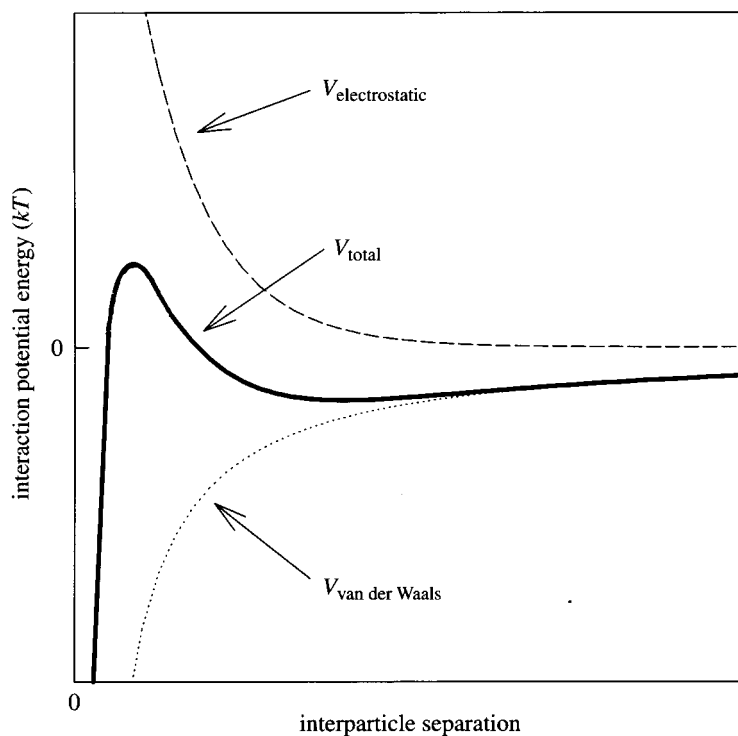


Figure 6.3. Schematic potential energy curve describing the interactions between colloidal particles. The overall potential is a sum of an electrostatic repulsive term which arises due to any charged groups on the surface of the particle and the attractive van der Waals term.

to an 'electric double layer' consisting essentially of two distinct regions; an inner region of adsorbed ions called the Stern layer and a more diffuse region. When two such diffuse layers overlap, a repulsive interaction is introduced. For typical ionic strengths, e.g. $10^{-3} \text{ mol dm}^{-3}$, the thickness of the double layer is about 10 nm. If the ionic strength is substantially higher, the double-layer interaction is sufficiently reduced and it can no longer provide stabilisation against the van der Waals driven aggregation. In contrast to the van der Waals interaction which falls off reciprocally with distance, the electrostatic repulsion falls off exponentially with distance. Consequently, the van der Waals interaction dominates at small and large distances, whilst the double-layer interaction dominates at intermediate distances. The interparticle interaction has the form shown in Figure 6.3.

The maximum in the potential corresponds to the barrier to aggregation – the inherent stability of the dispersion. If this barrier is larger than the thermal energy kT , the dispersion will be stable.

6.2.2 Creating pattern from instability

The stability of colloids can also be dramatically altered by inclusion of polymeric materials. If the polymer interacts favourably with the particle surfaces, i.e. it 'adsorbs', then both an increase and a reduction in stability is possible, via modification of the electrostatic interaction of the polymer is charged or a reduction in the van der Waals attraction.

The polymer layers, however, also introduces new contributions to the overall interaction between the particles. As two particles approach one another, compression of the polymer layer may occur which is unfavourable. Associated with this compression, is an increase in the local polymer concentration – this can be favourable or unfavourable depending on the solubility of the polymer.

If the polymer layers increases the stability of the dispersion, it is denoted 'steric stabilisation'. The polymer must fulfil two key criteria; (i) the polymer needs to be of sufficient coverage to coat all the particle surfaces with a dense polymer layer, and (ii) the polymer layer is firmly attached to the surface. How this is engineered is beyond the scope of this article, but the consequences of not satisfying these criteria are informative in understanding the effect that polymers have on the overall interparticle interaction. Since complete or incomplete coverage of the particles results in very different properties (i.e stability or instability), this is clearly one way in which minimal change in initial conditions can lead to major differences in product.

The presence of insufficient but very large polymers can also reduce the stability. When the particles attain a separation such that the polymer layers on an adjacent particles may bridge between the particles, a favourable interaction occurs and a loss of stability ensues. This is termed bridging flocculation.

A non-adsorbing polymer in solution can also destabilise a dispersion through a mechanism called depletion flocculation. When polymer molecules do not interact favourably with the particle surfaces from an enthalpic perspective, they are repelled from the surface regions due to entropic reasons. A 'depletion zone' around the particles is created which has a lower average polymer concentration than the bulk solution. The osmotic

pressure difference results in solvent being pulled from the depletion zone in essence, pulling the particles closer together. This is equivalent to an attractive interparticle interaction. Interactions involving surface polymers are of great interest in explaining biological microarchitectures as in many cases, the likely components will be separated from the supporting fluids by mixed polymeric membranes involving lipids, proteins and polysaccharides.

Another important interaction that needs to be considered is the 'hydrophobic interaction'. This can be most easily thought of in terms of two immiscible liquids such as oil and water being induced to mix by adding surfactants, to form (micro) emulsions. The exact structure of the phase formed depends heavily on the relative compositions of the various phases and the structure of the surfactant (see Figure 6.4).

Below some critical surfactant concentration, the system is two-phase with excess oil or water depending on the oil/water concentration. On adding more surfactant, the system moves into a one-phase region with normal micelles forming in water-rich systems. The water constitutes the continuous phase, solvating the headgroups of the surfactant whose hydrophobic tails solubilise oil in the core of the micelle. In oil rich systems, reverse-micelles form. With further increases in surfactant composition,

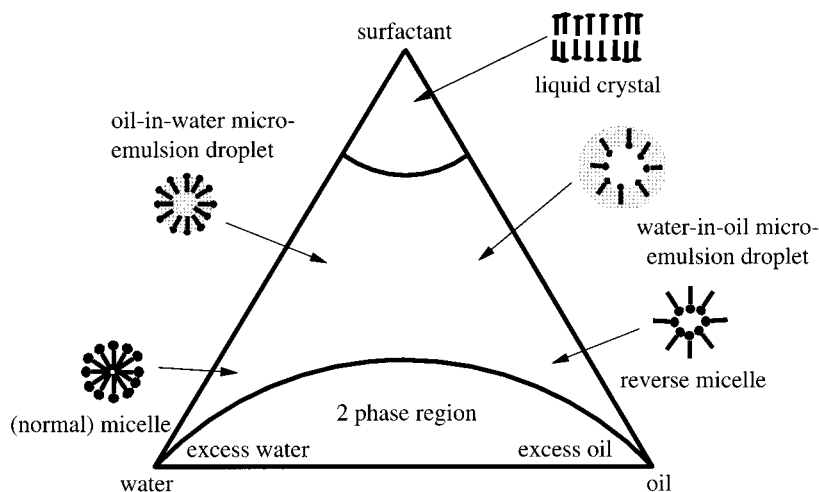


Figure 6.4. Schematic phase diagram for a three-component (oil, water, surfactant) system showing some of the self-assembled structures which form in the various regions.

oil-in-water or water-in-oil (micro) emulsion droplets form. Ultimately, at high surfactant compositions, liquid crystalline (lamellar) structures form.

In the natural system the sites of spore wall formation, i.e. the sporangial loculus, act as mini-reactor vessels in which the above interactions can occur. If a polymerisation occurs within one such structure, the resulting (polymer) architectures will probably closely resemble the self-assembled ones formed in our artificial sporangia.

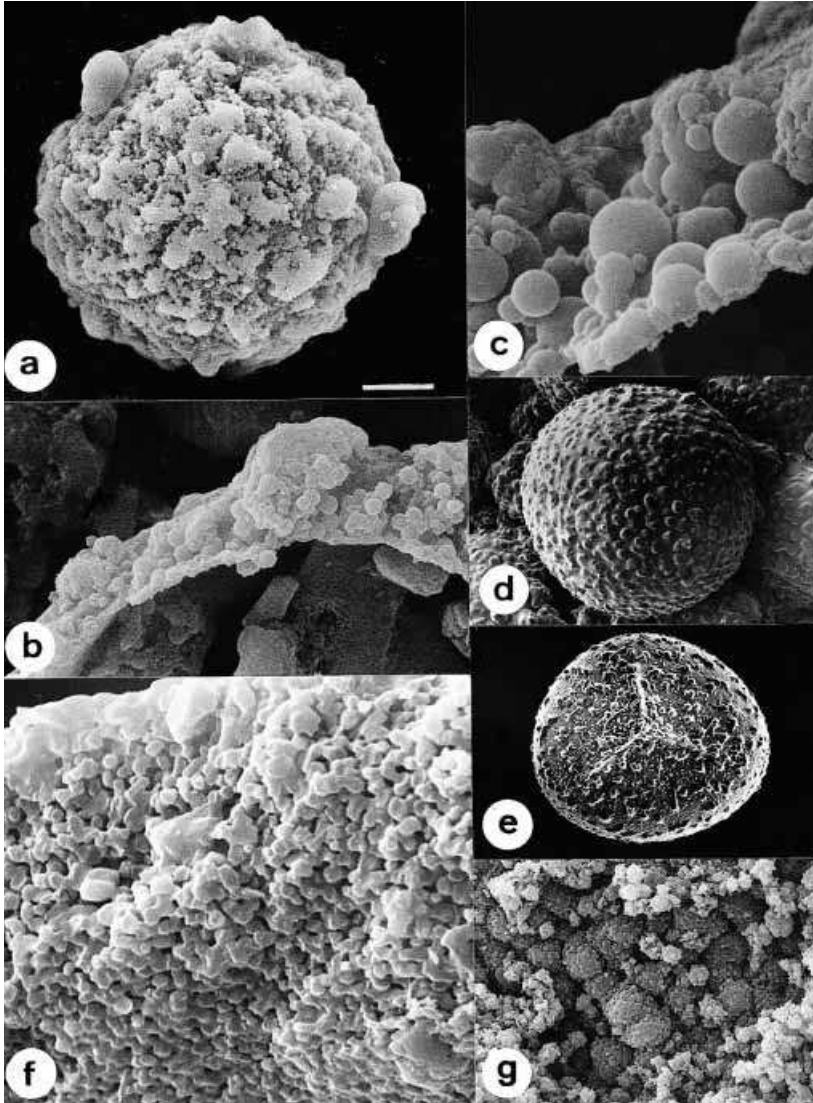
6.3 Synthetic self-assembled architecture and evolutionary implications

6.3.1 An experimental example

Following identification of the colloidal crystal layer within our spore walls, an attempt was made to utilise a simple colloid consisting of polystyrene particles in water (a latex) to mimic the natural structure. To cause flocculation of the particles, carboxymethylcellulose (CMC) was introduced with the intention of initiating a depletion interaction as described above. Although different from sporopollenin, polystyrene shares some properties and is at least reasonably well understood with regard to its colloidal behaviour. CMC was chosen as a relatively 'natural' polysaccharide. These initial experiments proved successful and resulted in the formation of colloidal crystals like those within the spore walls, but more significantly, they were built by processes and components which we believe behave in a similar manner to those in the natural system. Similar particle flocculations, but of an amorphous nature and formed from particles of inconsistent size could be produced by either depletion or bridging flocculation. Subsequent experiments have utilised hydrocarbons and lipids (known from the natural system of wall production) to synthesise mimics resembling other types of spore wall with some success.

It is disconcerting how 'life-like' some structures built from synthetic colloidal particles can be (Figures 6.2(b) and 6.5(a–d)). Hollow spheres of

Figure 6.5. Experiments involving mimics of sporopollenin (the principal component of spore walls) demonstrate that patterns very similar, if not identical to those of natural spores and pollen, can be produced from mixtures containing colloidal particles. All scales refer to bar in (a). (a) Spore-like structures of polystyrene particles and particle aggregates formed around a droplet of hydrocarbon. Scale = 10 μm . (b) A broken structure like that shown in (a). Scale = 5 μm . (c) Detail of the composition of the wall of the mimic spore shown in



(b). Scale = 2 μm . (d) Large scale particle aggregates formed in the presence of lipids, again around a hydrocarbon droplet. Scale = 500 μm . (e) A genuine spore of *Selaginella selaginoides* (club moss). Scale = 400 μm . (f) The wall structure of a broken spore of *Selaginella selaginoides*. Scale = 3 μm . (g) Natural sporopollenin particle aggregates and colloidal sporopollenin occurring during wall development in *Selaginella laevigata*. Scale = 10 μm .

aggregated particles and particle aggregates ('raspberries') are self-assembling from polystyrene latex in a water/cyclohexane emulsion. These are comparable to 'raspberries' and aggregated particles of sporopollenin formed during the development of *Selaginella* spores (Figure 6.5(g)). Similar structures occurring in water/rape seed oil emulsions (Figure 6.5(d)) closely resemble some *Selaginella* spores in surface architecture and internal organisation (Figure 6.5(e-f)).

The following hypothetical situation might arise, reflecting that found in synthetic systems. An oil-in-water emulsion forms, comprising a monomer such as a hydroxycinnamic acid (Figure 6.6) stabilised by fatty acids. The polymerisation resulting in sporopollenin can occur through a free radical mechanism involving the vinyl group, although the concentration of free radicals is likely to be low in natural systems, or through an alcohol + acid condensation to form an ester. The latter polymerisation, certainly in a synthetic application, is very slow in the absence of any added (acid) catalyst although a second molecule of acid could self-catalyse the reaction. Nevertheless, the kinetics of this reaction are very sensitive to concentration.

Furthermore, should free radicals be present, the vinyl groups would much more rapidly polymerise depleting the emulsion droplets of monomer, providing the control required for a particular particle size. The composition of the solution thus determines not only the phase behaviour, but the rate of polymerisation and the particle size. If, the organism has in its genetic code, the ability to synthesise the monomer, it presumably has

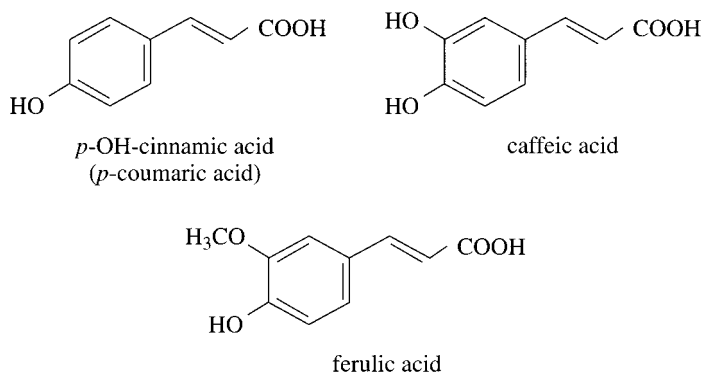


Figure 6.6. Three hydroxycinnamic acids common in plants and of interest as potential sporopollenin components.

the information to degrade any excess. This natural equilibrium could also create the initiator species as a by-product of the reaction which breaks down the excess monomer.

6.3.2 Of patterns and species

Differences in microarchitecture in relation to component concentration would appear to occur in our simulations of *Selaginella* megaspore wall construction. Imagine an example in which our synthetic wall structure is determined by concentration of styrene and cyclohexane (in the plant, these would be sporopollenin monomer and a fatty acid) all in water. Different arrangements (and sizes) of polystyrene particles occur depending upon the conditions at the initiation of polymerisation. In the hypothetical example shown in Figure 6.7, compositions and conditions represented by a and a' are different. They result from slightly different genetic codings but despite this, they both give rise to the same ultimate structure (they are within the same domain of the diagram). Examples b and b' may have much more similar genetic codings (they may differ only

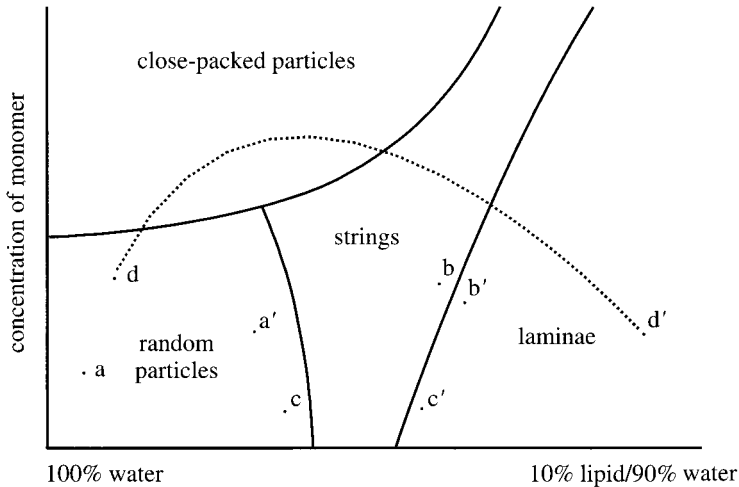


Figure 6.7. Hypothetical representation of a set of architectural domains defined by monomer concentration and proportion of lipid. Each defines structure regardless of the exact composition, providing this lies within its boundary. Letters a to d and a' to d' represent specific concentrations of components. The dotted line d to d' shows a pathway of changing concentration by which a spore wall such as that shown in Figure 6.1(c) might be constructed.

in a single base pair and thus produce similar concentrations of components) but because this gives rise to initiation points either side of a domain boundary, the resulting structure is different, possibly very different (strings or laminae). Points c and c' , although probably closer to each other (genetically) than a and a' , may be considered to exhibit the greatest difference in microarchitectural expression since these are separated by two domain boundaries. Significantly, it may not matter for any subsequent stage of development from where within each domain, the original composition was positioned since what matters is how the new components interact to initiate the next stage of development. It is abundantly clear from this illustration, that assessment of relationships of organisms based on comparison of the genetic code would differ somewhat from any assessment of based on patterning and structure. Consider the likely outcome of such an analysis on a , a' , c and c' . There are further complications in that composition will usually change as wall development occurs (consider arrow from d to d' and compare with Figure 6.1(c), development is from bottom to top). In addition, any *in vivo* self-assembly system such as this is reliant upon second hand manipulation by proteins/enzymes which have already been through a similar selection process.

The incorporation of self-assembly mechanisms in development is clearly advantageous to an organism if the processes involved are sufficiently robust and the results consistent. Such systems represent a saving in terms of both the required genetic code and its decryption (via ribosomal RNA) into enzymic regulatory proteins. The genetic code need only describe the initial conditions and not the complexity of the ultimate structure. Over the great expanse of time involved in the evolution of life (particularly simple, single-celled organisms) many self-assembly mechanisms have been included by chance, much as proteins with a specific function have been retained and elaborated. Amongst organisms, many self-assembly mechanisms are shared (although they may result in different patterns and architecture due to different initial conditions), whilst others may be unique. However, the identification of such mechanisms and an assessment of their distribution amongst organisms will surely assist in both an understanding of organismal relationships and the meaning of structural, architectural and pattern diversity between 'species'. The observation that self-assembly systems can switch from the production of one pattern to another with only minor modification of the initial conditions (supported by our colloidal work) adds weight to the view

that evolutionary change (in the form of speciation) could be relatively rapid.

The evidence we offer above for the microarchitectural development mechanisms occurring within spore walls serves to underline the significance of colloids in biological construction and pattern formation. As we have demonstrated, an understanding of colloidal mechanisms has the potential to explain certain aspects of biological complexity. As a first approximation to reality, our organic mimics have already revealed much about the way in which spore walls form. Furthermore, they have begun to indicate just how much of our ultimate structure is governed by the ways in which our microscopic components interact.

6.4 Future applications of biocolloid self-assembly

Clearly the improved understanding of colloidal behaviour within living systems that we are developing offers the eventual prospect of our being able to manipulate such systems. The control of microarchitecture in both living and synthetic systems has many potential applications. The most important aspect is the ability to define the particular conditions under which a certain pattern or structure will be formed such that the products will be uniform. This clearly happens in Nature, but natural systems have been subject to trial and error for considerably longer than any experiment involving synthetic systems.

Natural materials, particularly compounds such as sporopollenin with almost total resistance to digestion, could be used in the manufacture of cosmetic and drug delivery capsules, and would be both safe and efficient. Our studies of the colloidal construction of spore walls reveals how we might design such capsules with a high degree of control over size, wall thickness, solubility, and porosity leading to complete regulation of dosage. Such capsules could be self-assembled in vast numbers, possibly even around micelles containing the intended active content. As our understanding of the interaction of plant genetics and colloidal construction mechanisms improves, we may eventually be able to manipulate plants into producing both the required capsule and content.

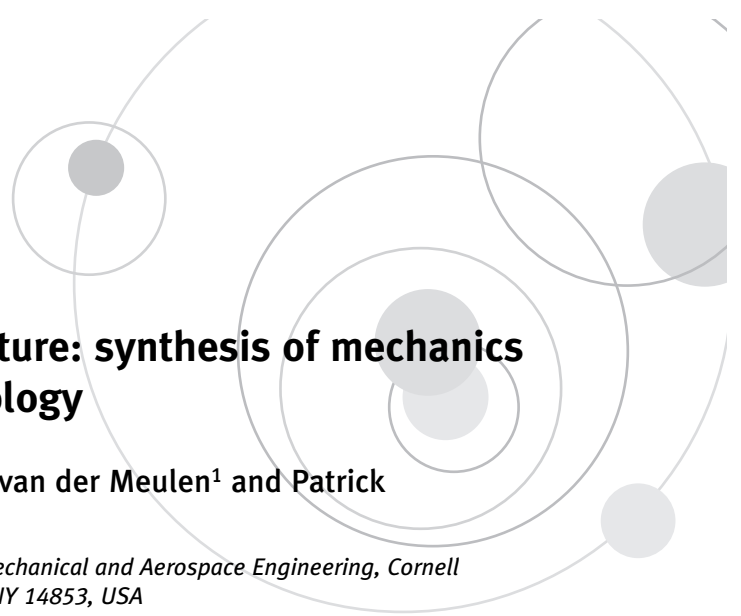
Regulation of microarchitecture has applications in the production of surface coatings. Again, control of the consistency of pattern offers the prospect of the self-assembly of periodic surface features on a scale that would interact with incident light. Paints could be designed to produce

iridescent effects or to produce specific finishes upon drying. The use of natural water-based colloidal systems could eliminate the need for potentially harmful or irritating volatile components. Were consistent surface patterns to be of a highly repetitive nature over relatively large scales, they may potentially be of use in the production of computer chip technology, providing a template for microcircuitry. Again, it might be feasible to extract the required component chemicals from genetically engineered plants, much as we can extract clove oil, ephedrine or opium now.

The use of colloidal chemistry in the production of synthetic organic microarchitecture based on that produced by living systems is in its infancy. Its development will naturally run parallel to the greater utilisation of genetic manipulation of organisms both as whole organisms and as organismal components in test tubes. We perceive a time, within the new millennium, in which we are able to control Nature, not just through genes, but by making use of the inherent properties of biological construction materials and processes. These substances and mechanisms will be, by their very nature, 'friendly' to both humans and the environment as a whole.

6.5 Further reading

- Cohen, J. 1995 Who do we blame for what we are. In *How things are* (eds. J. Brockman & K. Matson), pp. 51–60. London: Weidenfield and Nicholson.
- Hemsley A. R. & Griffiths, P. C. 2000. Architecture in the microcosm: biocolloids, self-assembly and pattern formation. *Phil. Trans. R. Soc. Lond. A.* **358**, 547–564.
- Kauffman, S. A. 1993 *The origins of order*. Oxford: Oxford University Press.
- Mann, S. & Ozin, G. A. 1996 Synthesis of inorganic materials with complex form. *Nature* **382**, 313–318.
- Shaw, D. J. 1980 *Introduction to colloid and surface chemistry*, 3rd edn. London: Butterworths.
- Thompson, D. W. 1961 *On growth and form* (abridged edition, ed. J. T. Bonner). Cambridge: Cambridge University Press.



7

Skeletal structure: synthesis of mechanics and cell biology

Marjolein C. H. van der Meulen¹ and Patrick J. Prendergast²

¹ *Sibley School of Mechanical and Aerospace Engineering, Cornell University, Ithaca, NY 14853, USA*

² *Department of Mechanical Engineering, Trinity College, Dublin 2, Ireland*

7.1 Introduction and historical background

Vertebrate skeletons serve several important functions, including structural support; storage of ions, particularly calcium; production of red blood cells; and protection of vital organs such as the heart and lungs. The structural role of the skeleton is particularly interesting to physicists and engineers because bone is an adaptive, living tissue containing cells that respond to their physical environment. Form follows function. Most of our knowledge comes from treating bone as a conventional engineering material, studying bone tissue and whole bones when placed under loads in the laboratory. This approach does not consider that bone is a living tissue or account for the influence of mechanical loading in regulating the biological processes that occur to form the skeletal structure. Separating the mechanics from the biology is impossible: mechano-biological coupling begins during early development when the primordial cellular structure first experiences deformations and pressures and continues throughout the growth, development and aging of the organism. The influence of biophysical stimuli on skeletal growth and adaptation is of great interest in treating diseases such as osteoporosis and osteoarthritis.

The structural adaptation of the skeleton is one of the most fascinating problems in the history of science. Galileo observed in 1638 that longer bones had to be thicker than shorter ones to have the same structural

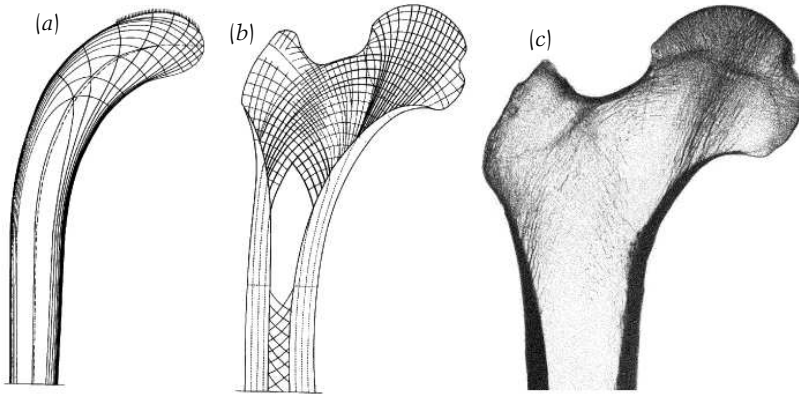


Figure 7.1. Early diagrams showing the relationship between stresses created by forces on bones and the internal architecture of the skeleton: (a) Culmann's calculation of the stress trajectories in a crane, (b) Wolff's drawing of the trabecular orientation in the upper part of the femur, and (c) a photograph of the cross-section of the upper part of the femur.

strength. In discussing heritable and acquired traits in *On the origin of species*, Charles Darwin noted that flying wild ducks have proportionally larger wing bones and smaller leg bones than their nonflying domestic relatives. Many natural philosophers of the nineteenth century used mechanical principles to explain bone geometry. In 1892 Julius Wolff studied many pathologically healed bones and concluded that bone tissue is distributed within the organ in ways to best resist mechanical forces. A famous exchange between the Swiss engineer Karl Culmann and his colleague Hermann von Meyer is considered the defining 'eureka' episode of modern biomechanics. The internal architecture of a femur was being demonstrated by von Meyer, and Culmann, who developed the methods of graphic statics, exclaimed, 'That's my crane' (Figure 7.1). These concepts were further developed and generalised by D'Arcy Thompson in his influential work *On growth and form* in 1917. The mechanism of bone adaptation was first addressed by the German embryologist Wilhelm Roux in 1895, who proposed the controversial hypothesis that bone cells compete for a functional stimulus, *à la* Darwin, and engage in a struggle for survival that leads to *Selbstgestaltung* (self-organisation).

Roux and his contemporaries were not able to advance much beyond

this philosophical and descriptive understanding of the role of mechanics in skeletal growth. As the twentieth century progressed, biology increasingly reduced the organism to the molecular level, and the interest in mechanics and other biophysical factors waned. In recent years, the emergence of several new technologies has fostered a reexamination of the old questions relating to the mechanical regulation of tissue growth and adaptation. The first of these is computer-based structural modeling, which allows a more valid analysis of effects of physical forces within complex skeletal geometries; the second is molecular biology, which localises individual gene expression and protein synthesis under different mechanical forces; and the third is the tremendous advances in imaging technologies that enable scientists to identify microstructural characteristics of tissues and the role of cells in constructing and maintaining skeletal strength. In this essay, we call on our current understanding of the role of mechanical forces in skeletal biology to highlight the interaction between the physical and biological sciences.

7.2 Form and function in bone

The musculoskeletal system consists of bones, blood vessels, nerves, ligaments, tendons, muscles, and cartilage, which work together to perform the structural and kinematic functions of the organism. These musculoskeletal tissues all have a composite structure of cells embedded in a matrix produced by the cells themselves.

7.2.1 Bone structure

The geometry and structure of a bone consist of a mineralised tissue populated with cells. This bone tissue has two distinct structural forms: dense cortical and lattice-like cancellous bone, see Figure 7.2(a). Cortical bone is a nearly transversely isotropic material, made up of osteons, longitudinal cylinders of bone centred around blood vessels. Cancellous bone is an orthotropic material, with a porous architecture formed by individual struts or trabeculae. This high surface area structure represents only 20 per cent of the skeletal mass but has 50 per cent of the metabolic activity. The density of cancellous bone varies significantly, and its mechanical behaviour is influenced by density and architecture. The elastic modulus and strength of both tissue structures are functions of the apparent density.

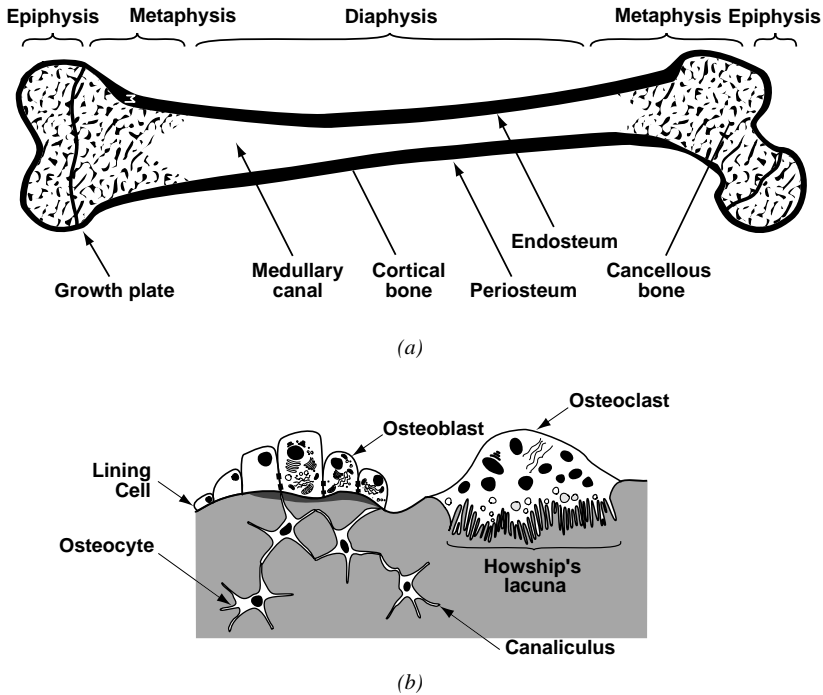


Figure 7.2. Schematics of bone anatomy: (a) the structure of a long bone demonstrating the distribution of the two different tissue structures, cortical and cancellous bone, and (b) the cells present in bone: osteoblasts, bone-forming cells found on surfaces; osteocytes, bone cells embedded in the mineralised matrix; and osteoclasts, bone-removing cells.

7.2.2 Cells and matrix

Cortical and cancellous bone tissue consists of cells in a mineralised matrix. All skeletal cells differentiate from a common precursor cell pool: the mesenchymal stem cells of the embryo (Figure 7.3). Mechanical stimuli influence the mode of stem cell differentiation and the resulting tissue type. Manipulation and control of stem cell differentiation holds considerable promise in the field of tissue engineering and is receiving much commercial and ethical attention. In addition to precursor cells, three principal cell types are present in bone: osteoblasts, osteocytes, and osteoclasts (Figure 7.2(b)). Osteoblasts are active bone-forming cells. All bone surfaces are covered by a single layer of precursor cells and resting

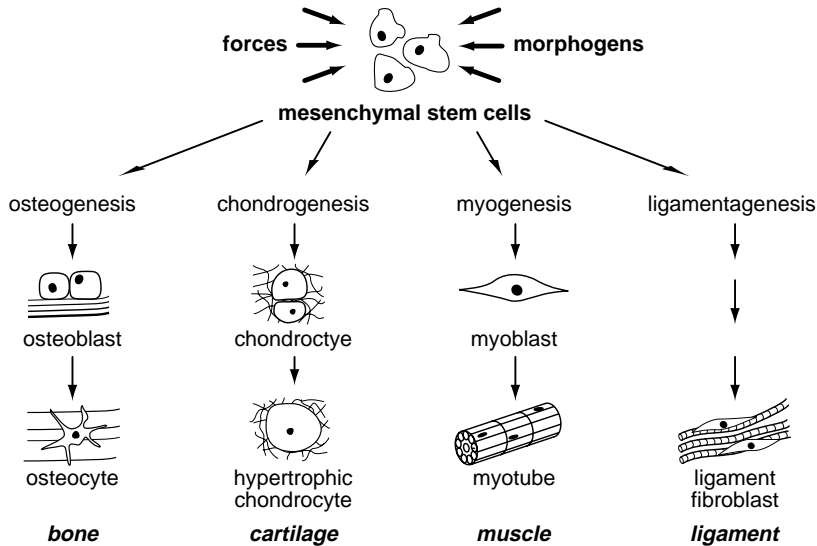


Figure 7.3. All skeletal tissues arise from a single cell type, the mesenchymal stem cell. Differentiation into bone, cartilage, muscle, or ligament occurs in response to the mechanical and biochemical stimuli of the stem cell's environment.

osteoblasts. Upon activation, osteoblasts secrete osteoid, the organic extracellular matrix into which mineral crystals are deposited. The organic matrix contains 90 per cent collagen and a ground substance consisting of large protein polysaccharides and a variety of matrix proteins. Gaps in the collagen fibrils serve as mineral nucleation sites for calcium phosphate, which forms the inorganic phase. The result is a composite of a ductile polymeric (collagen) phase and a strong ceramic phase. This combination gives bone its high mechanical strength and toughness.

Approximately 15 per cent of osteoblasts become entrapped in their own matrix to become osteocytes. Osteocytes have a vast three-dimensional network of cell processes (canaliculi), providing nourishment and cell-cell interactions. Because they are located throughout bone tissue and have an extensive canalicular network, osteocytes are assumed to be a vital component of sensing mechanical signals. Nutrients are essential for the vitality of bone tissue and are obtained from the blood supply, limiting most osteocytes to lie within $150\ \mu\text{m}$ of a blood vessel, resulting in a high cellular density: 25 000 osteocytes within a square millimetre of bone

tissue. The third cell type, the osteoclast, unlike the other two, is presumed to arise from the fusion of blood cells. Osteoclasts are large distinctive multinucleated cells that resorb bone. By sealing to a bone surface, the osteoclast forms an acidic cavity that dissolves the underlying bone (Figure 7.2(b)).

7.2.3 Bone growth and maintenance

Bone forms through two different developmental processes: endochondral ossification and intramembranous ossification. Endochondral ossification involves an intermediate tissue stage, cartilage, not present in intramembranous formation. The long bones all form endochondrally. In these bones, development begins with the condensation of mesenchymal cells, which differentiate into chondrocytes (Figure 7.3), creating a cartilage prepattern of the skeleton. The first bony tissue, known as the bone collar, appears spontaneously around the midshaft. Thereafter, ossification proceeds axially towards each bone end. The same identical sequence of ossification occurs at each location: the cartilage calcifies, blood vessels invade the site, and the cartilage is resorbed and replaced by bone. This sequence is regulated by genetic factors, systemic hormones, growth factors, and mechanobiologic effects. The timing of these signals is critical to the outcome. In the developing embryo, the first ossification of cartilage is coincident with the first muscle contractions – if a muscle is immobilised in the embryo, a distorted and disorganised bone forms, demonstrating the link between mechanics and bone tissue formation.

After embryonic bone formation, the skeleton continues to grow in length by dividing and enlarging cartilage cells, which then ossify to form cancellous bone. Bone diameter grows by direct deposition of bone on existing bone surfaces, accompanied by resorption of outer surfaces. As the skeleton continues to develop, mechanical forces generate an ever-increasing influence on the forming bone architectures and geometries. Cellular proliferation increases skeletal size and needs to be exquisitely controlled to maintain form and proportion throughout growth.

Once the skeleton is formed, continual 'remodelling' of bone tissue maintains structural integrity and creates more orderly tissue structures. Remodelling involves coupled resorption and formation on all bone surfaces in a well-defined sequence of events. The remodelling sequence has been described as activation of the surface, resorption by osteoclasts, reversal, formation by osteoblasts, and return to quiescence of the surface. In

the adult, remodelling serves to repair, renew, and adapt bone tissue. A primary function of remodelling is to replace damaged tissue such as microcracks resulting from repetitive functional loading. Without this continuous repair process, a much larger skeleton would be needed to prevent the accumulation of damage.

Repair of bone fractures is another important mechanically mediated process. A fracture initiates a multistage sequence of tissue regeneration which recapitulates tissue differentiation and development. This process also occurs in individual, fractured trabeculae. Initially a large granuloma forms containing undifferentiated mesenchymal stem cells whose differentiation is regulated by genetic and epigenetic factors. Following this immediate trauma response, the cells differentiate into cartilage to stabilise the fracture. The initial bridging and immobilisation are performed by tissues that can tolerate the high strains that preclude bone formation. Thereafter, endochondral ossification of the cartilage occurs and bone forms. Finally, the new bone is remodelled and integrated into the original structure. The mechanical environment is critical to the ability of the tissue to regenerate. Immobilisation of the fracture may enhance early healing at a time when stability is critical. Fractured bones that are dynamically loaded during healing regain strength more quickly, but if the applied strains are too large, cartilage or fibrous tissue forms and a pseudo joint may develop at the fracture site.

7.3 Mechanical regulation of bone structure

7.3.1 Adaptation experiments

The growth and development of organisms living within Earth's gravitational field are intricately linked to mechanical demands. Manipulation of forces in animal experiments has provided insights into the overall nature of the adaptation process. The characteristics of adaptation to increased or decreased *in vivo* loading include changes in bone quantity, not material quality; greater response in immature than mature tissue; and response to cyclic, not static, loading. These results were first demonstrated in a series of well-designed studies with loads applied to rabbit limbs and have been confirmed by a variety of studies since then. In the adult, in general, when the loads are increased over normal levels, bone mass is increased, and when the loads are decreased, bone mass is lost. Changes occur in the cross-sectional size and shape of cortical bone and in the apparent density

of trabeculae; bone length is seldom significantly affected, but curvature may be altered.

Most experiments examine cortical bone responses, in contrast to the historical interest in trabecular adaptation. Exercise often shows little or no effect, presumably because the overall activity level is not substantially elevated beyond the normal range. Demonstrating a definitive decrease in physiological loads is more straightforward and has been accomplished by casting, space flight, and hindlimb suspension. Hindlimb suspension was developed as a ground-based model for space flight, demonstrating similar skeletal effects. Compared to age-matched controls, suspended growing animals continue to grow, but at a reduced rate, with lower age-related increases in femur strength and cross-sectional area (Figure 7.4). Decreased bone formation occurs on the outer cortical surface, exactly the location of the greatest reduction in mechanical stimulus.

Although many experiments have been performed, quantitative relationships between mechanical loads and bone adaptation do not yet exist. *In vivo* strain gauge studies have found a remarkable similarity of peak surface strains: $-2000\mu\epsilon$ at the midshaft of different bones across different animals at maximum activity. Measuring strains in adaptation studies would allow us to relate *in vivo* load changes to altered surface strains to adapted bone mass and strength.

Applying loads directly to a skeletal site has the advantage that the load magnitudes, frequency, and duration are known or controllable. Loads

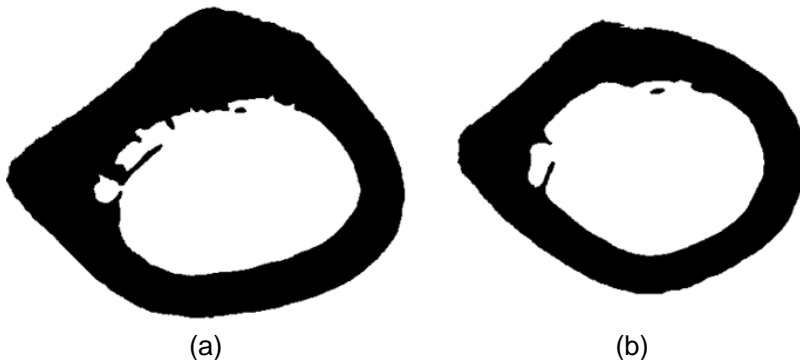


Figure 7.4. Digitised cross sections of femora from 67-day-old rats: (a) normal control and (b) four-week suspended animals. Cross-sectional area of (a) is 5.3 mm^2 and (b) is 3.8 mm^2 , a 29 per cent reduction due to the unloading during growth.

at sites or in directions that are not normally loaded have been demonstrated to induce a greater response than increasing physiological loads. Recent experimental models for noninvasive, controlled *in vivo* loading have been developed to test weight-bearing bones in the rat. These new *in vivo* approaches can be integrated with *in vitro* and *ex vivo* studies to acquire a more complete understanding of load-induced adaptation. These animal models can be used to examine loading parameters, to study gene expression, and to validate computer simulations. The mouse has recently become more relevant; our ability to manipulate the mouse genome has led to the development of mutations and new biological markers and assays. *In vivo* loading of mouse mutants will help identify critical genes and regulatory factors in the mechanical response pathway.

Adaptation around bone implants has received considerable attention clinically and experimentally. When a bone segment is replaced by a stiff metal prosthesis, the implant becomes the primary load bearing structure, reducing the mechanical stimulus to the surrounding bone. Severe bone loss is one of the impediments to the long-term success of orthopaedic joint replacements. Future developments will include active devices that stimulate the surrounding bone and, ultimately, artificial organs engineered in the laboratory.

7.3.2 Modelling

For a skeletal element to respond to its mechanical environment, the cells in the tissue must regulate their environment in response to the mechanical stimuli they receive. The regulatory process can be thought of as a feedback loop (Figure 7.5) in which the osteocyte senses the stimulus and

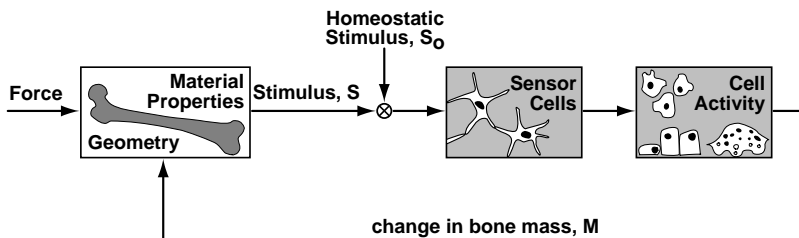


Figure 7.5. Feedback diagram for skeletal mechanical regulation. When forces are applied to a whole bone, the stimulus that results is sensed by the bone cells in the tissue. The sensor cells then signal bone-forming and -removing cells to change the geometry and material properties of the bone.

signals the osteoblasts and osteoclasts either to add or to resorb tissue to regain the physiological environment the cell requires. To maintain normal bone mass, the sensing cells require a desired or reference stimulus value. If the actual stimulus present in the tissue is less than the reference level, bone mass will be lost through resorption by osteoclasts, and if the actual stimulus is above the reference level, bone will be formed by osteoblasts. As a result of this adaptive response, the stimulus in the tissue will approach and ultimately equal the desired stimulus value. Since the sensory cells are distributed throughout the tissue, this model describes a spatially discrete process in which each cell regulates its mechanical stimuli by changing the mass or density of its extracellular environment. The driving mechanical stimulus is not known, and many biomechanical measures have been proposed, including strain, strain energy density, and fatigue microdamage. These approaches can be coupled to computational stress analysis procedures and have been used to predict bone adaptation around implants and simulate the influence of mechanics on long bone growth.

Recently, considerable interest has been centered on investigations of the nonlinear dynamics of bone adaptation. Finite element models have been used in iterative computer simulation procedures based on the feedback approach described above. The evolution of bone density and structure can be simulated for a given mechanical stimulus and initial density pattern (Figure 7.6). This phenomenon can be viewed as a self-organisational process operating within the bone: all elements will either become fully dense or resorb to zero density, creating a porous 'trabecular' structure. The evolution of this density depends on the initial conditions, so that a different initial density leads to a different trabecular structure, indicating a nonlinear process. Furthermore, the final configuration is metastable because a perturbation of the structure caused by the fracture of a trabecula, for example, will not be followed by a return to the former equilibrium. In reality, however, bone structures are stable because the inevitable trabecular microfractures that occur in aged osteoporotic bone do not lead to immediate degeneration, but rather the regulatory process leads to a completely new equilibrium. If this computer simulation does indeed capture the essence of bone adaptation, then adaptation is a far-from-equilibrium dynamic process generated by positive feedback. To date, these approaches have focused attention on the central role of mechanical factors in determining bone structure.

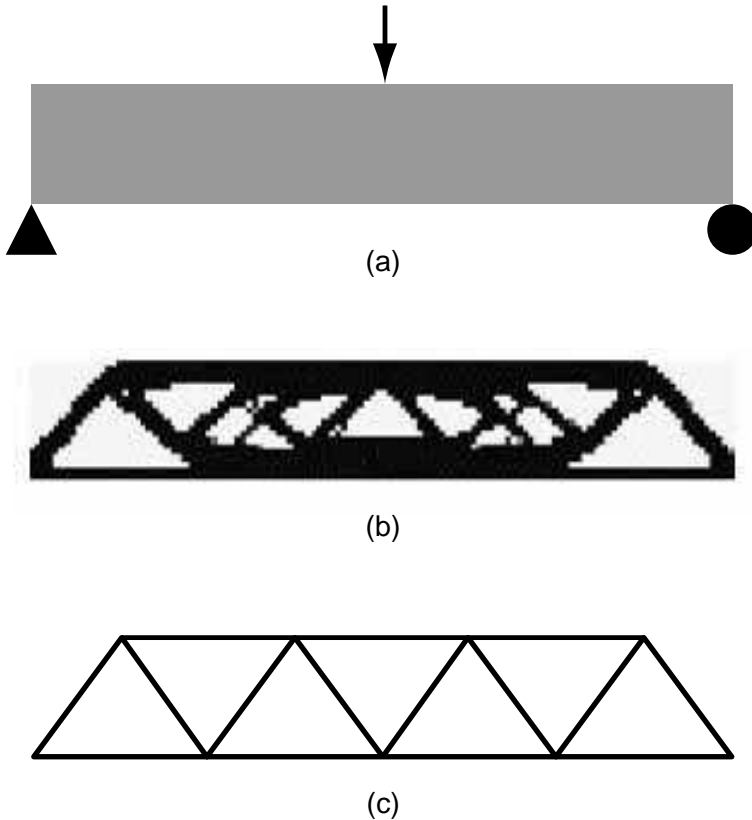


Figure 7.6. When a force is applied to a bone of uniform structure (a), the structure adapts by the feedback mechanism shown in Figure 7.5 and forms a nonuniform structure to carry the load efficiently (b). The resulting structure resembles the familiar design of bridges and other man-made trusses (c).

Mechanobiologic concepts have been applied to other skeletal tissues. Differentiation of stem cells to form cartilage, fibrous tissue, and bone is central to tissue growth and regeneration. Friedrich Pauwels proposed in 1941 that hydrostatic stresses stimulate differentiation to cartilage cells, whereas distortion stimulates differentiation into fibrous cells (Figure 7.3). Simulations based on Pauwels's ideas have correlated patterns of mechanical stimuli with skeletal tissue type during fracture healing. These models suggest that we will soon be able to simulate skeletal growth, adaptation, and degeneration over an individual's lifetime.

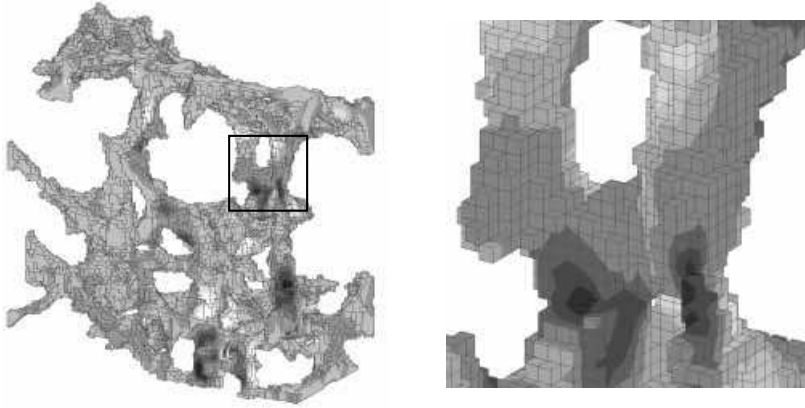


Figure 7.7. A finite element model of a bone specimen in compression. This model was created by converting the voxels from a microcomputed tomography scan into individual bone elements. Loads can then be applied to the model to understand the stresses that are created in the bone tissue.

7.3.3 Imaging

A key new tool in the validation of analytical models is high-resolution imaging coupled with computer analyses to calculate the material stresses, strains, and stimuli within cancellous bone. The average thickness of a trabecula is 100–150 μm , undetectable with conventional computed tomography resolution of 100–200 μm . Microcomputed tomography can image bone at 17 μm resolution, and the images can be converted directly into large-scale finite element models (Figure 7.7). These models can determine bone stiffness and strength without the need for a traditional mechanical test. These ‘virtual bone biopsies’ have the potential to revolutionise the clinical assessment of bone health, an increasingly important clinical objective in an aging population susceptible to osteoporosis. Although these tomography-based models simulate the architecture precisely, the magnitude and variation of tissue-level material properties still need to be determined.

Another imaging development is laser scanning confocal microscopy to image individual living cells noninvasively. The deformation of osteoblasts and chondrocytes has been observed using this method. Confocal microscopy has also been used to image microdamage in bone tissue showing modes of microcrack arrest within the complex microstructure of bone tissue.

7.4 Visions for the future

With these new tools and so many unanswered questions about tissue function and maintenance, the time for mechanobiology has truly arrived. High-resolution imaging systems will allow us to determine tissue structures from the highest hierarchy of the organ to the lowest of the genome. These digital images are ideally suited for analysing physical forces and linking continuum level tissue stresses to deformation-induced gene activation in the DNA molecule. Advances in dynamic systems theory and applied mathematics will play a critical role in explaining the behaviour of otherwise intractable models.

As the complete genomes of organisms become mapped, functional genomics will combine with biomechanics to answer questions such as: what is the regulatory role of mechanics in skeletal gene expression? How would organisms grow in the microgravity environment of space? Can we define the mechanical forces needed to culture complete skeletal organs in the laboratory? Are there genes that code for 'bone strength'? Orthopaedics and reconstructive surgery will be completely revolutionised.

The rapid growth of the field has produced an interdisciplinary community of engineers, biologists, mathematicians, and physicians who hope to answer scientific questions of the highest import. These questions will bridge the boundary between physics and biology – between forces and cells – to understand how organic forms are shaped by the mechanical world and how living systems actually 'extract order from their environment,' first posed by Erwin Schrödinger in 1943 in his famous lectures *What Is Life?*

7.5 Further reading

- Carter, D. R. & Beaupré, G. S. 2000 *Skeletal Function and Form*. Cambridge: Cambridge University Press.
- Currey, J. D. 1984 *Mechanical adaptations of bones*. Princeton: Princeton University Press.
- Martin, R. B., Burr, D. B. & Sharkey, N. A. 1989 *Skeletal tissue mechanics*. New York: Springer Verlag.
- Odgaard, A. & Weinans, H. (eds.) 1995 *Bone structure and remodeling*. Recent advances in human biology, Volume 2. Singapore: World Scientific Publishing Co.
- Thompson, D. W. 1917 *On growth and form*. Cambridge: Cambridge University Press.

van der Meulen, M. C. H. & Prendergast, P. J. 2000 Mechanics in skeletal development, adaptation and disease. *Phil. Trans. R. Soc. Lond. A* **358**, 565–578.



8

The making of the virtual heart

Peter Kohl,¹ Denis Noble,¹ Raimond L. Winslow² and Peter Hunter³

¹ *Laboratory of Physiology, University of Oxford, OX1 3PT, UK*

² *Department of Biomedical Engineering, JHU, Baltimore, MD 21205-2195, USA*

³ *Engineering Science Department, University of Auckland, New Zealand*

8.1 Introduction

This essay is about the making of the most comprehensive computer model of a human organ to date: the virtual heart. It will beat, ‘consume’ energy or experience the lack of it, respond to stress or drug administration, grow and age – in short, it will behave like the real thing. Or, let’s say, close to. Because the virtual heart may be stopped without harm at any point in time, and dissected, inspected, resurrected, etc. . . . We shall address this in more detail below, together with other enticing aspects of virtual organ development. In particular, we will try to:

- review the need for virtual organs in the context of contemporary biomedical research;
- introduce the ideas behind the ‘Physiome Project’ – a world-wide research effort, similar to the Genome Project, to describe human biological function using analytical computer models;
- provide insights into some of the more technical aspects of the virtual heart; and finally
- address the utility and benefit of this new tool for biomedical research, drug and device development, and the wider society.

In order to understand the dimensions of the making of the virtual heart – let’s stand back, for a minute, and consider the difficulties of studying and describing any unknown complex system.

8.1.1 Martians and the Highway Code

Imagine you are an alien. From Mars, to keep things simple. You are given the assignment, should you accept it, to report on the use of cars by humans. Please read on – this book will not self-destruct in a few seconds . . .



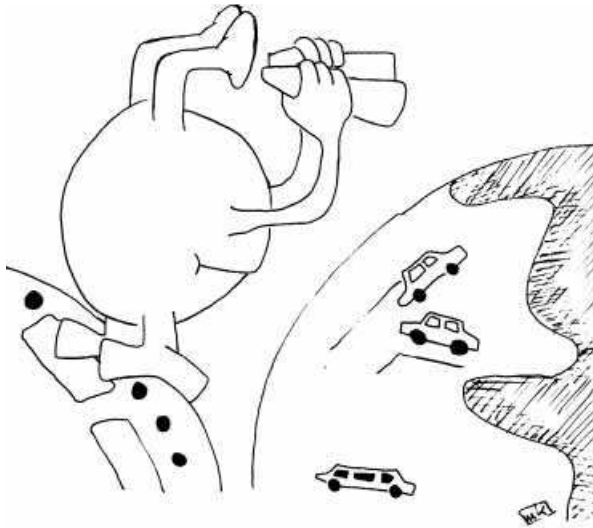
How would you go about it?

You could visit earth, hire a mechanical workshop in a remote area, car-jack a few specimens, and dissect them. You would observe that cars differ in their colour, shape, size and spec. Some may even contain a bar, cinema or swimming pool, but, perhaps, limousines are excluded from your exploration. On closer examination you would notice small ID-numbers imprinted on various strategic body parts. In short – you would find no two cars that are *exactly* the same.

Alternatively, you could focus on essential similarities between cars. For example that they *all* require one or the other kind of fuel to work.

Or, you could stay in orbit and look down at the *movement* and interactions of cars. You would soon find that in some parts of the planet cars stick to the left side of the road, while elsewhere they prefer the right. You would notice that most cars stop at red lights, but others don't. You would

also see that there are complicated rules of 'who goes first' at crossings, although they would not appear to be perfect. In short – you would discover a complex code for auto-motion.



Conversely, you might observe that – most of the time – all cars are *stationary!*

In your report you would summarise your observations. Your conclusions could range from 'cars are all different' to 'they are all the same', or from 'cars are made for driving' to 'they are for parking'.

What a muddle!

To shed more light on this, you might try to generalise all findings. You could develop a model concept of car use by humans, based on apparent traffic rules.

This would be a challenging task as you would have to understand the Highway Code from the observed behaviour of other road users! However, you might come up with a reasonably close model of national traffic rules. And you would not need to give detailed descriptions of individual components of a car to do so.

If, however, all cars would stop as a consequence of an oil crisis, governmental budget, or other major disaster – you would realise that there is no way of fully understanding the rules of auto-motion without addressing how an engine works.

The same applies to the heart.

No two cells in the heart are *exactly* the same, but they are *all* made of rather similar components. Also, it is possible to study and model the general ‘traffic rules’ for the spread of the electrical signal that controls cardiac contraction and pumping, without addressing the workings of the individual cells that produce the electrical wave. However, this knowledge alone would be of little help for diagnosis and treatment of major energy crises like myocardial ischaemia, or heart attack.

8.2 The need for computational modelling in bio-medical research

8.2.1 What can we learn from Martians?

Well, probably a lot. If they exist. What does exist for sure, though, is the challenge to understand in detail how the human heart works. And, similar to the above scenario, among the many different ways to advance this venture, there are at least two main directions: the top-down and the bottom-up route. Accordingly, bio-scientists tend to get pigeonholed into two schools of thought.

‘*Reductionism*’ is the direction that unites those guys who try to disassemble the parts of a biological system, and put them under a microscope (a laser-scanning quantum-leaping one, of course) to see the sparks of imagination hidden in the least of the components. The under-the-bonnet view, to stay with the Martian’s analogy.

‘*Integrationism*’, on the other hand, unites those who pride themselves for their holistic view of the complete system, without necessarily being burdened by a detailed understanding of structure and function of the minute components that make it work. The up-in-the-air perspective.

Reductionists might say that the division between the two schools of thought simply runs along the split between ‘thorough’ and ‘not-so-thorough’. Integrationists would probably claim that the divide is nearer the categories ‘geeky’ and ‘not-so-geeky’.

The two contrasting views were expressed at a higher level of sophistication during a recent Novartis Foundation meeting on *The limits of reductionism in biology* by Professor Lewis Wolpert and Professor Gabriel A. Dover, who said (respectively): ‘. . . there is no good science that doesn’t have a major element of reductionism in it . . .’, and ‘. . . we have imagined we have explained something merely by describing its parts, but all we

have done is create an excuse for not to think about it . . .' (Bock & Goode 1998).

This leaves us with the question of whether or not the two directions are irreconcilable.

We would like to think that the answer is a clear NO.

The logic of life will neither be recognised without precise understanding of the manifold of components that give rise to biological function, nor without a clear conception of the dynamic interactions between individual components. Likewise, the logic of life lies exclusively neither in the most incredible detail, nor in the most sweeping synopsis.

8.2.2 Combined opposites

This concept of a *natural interdependence* of opposites that seemingly exclude each other but, equally cannot survive without the other, is not new at all.

It is *the* central part of modern Dialectics – ‘the soul of all knowledge which is truly scientific’ – as taught by Hegel (*Encyclopaedia of the philosophical sciences*, 1830) and Engels (*Dialectics of nature*, 1879). And, to go back in time even further, ‘combined opposites’ – Yin and Yang – are central to old Chinese philosophy and ancient popular wisdom.

Thus, common sense would suggest that neither of the two – *Integrationism* and *Reductionism* (and this shall be the last time we affront the reader with an ‘-ism’) – is self-sufficient, and both are obligatory to the quest for knowledge.

This view lays the basis of probably the most exciting new development in bio-medical research – the Physiome Project.

8.3 The Physiome Project

8.3.1 The vision

The Physiome Project represents a world-wide effort to organise systematically the huge data mass on biological function into a ‘quantitative description of the physiological dynamics and functional behaviour of the intact organism’ (Bassingthwaighte). It was publicly initiated at the 33rd World Congress of the *International Union of Physiological Sciences*, 1997 in St. Petersburg (see <http://www.physiome.org>).

The Physiome Project sets a vision that will be much harder to accomplish than that of the Human Genome Project – formally begun in October

1990 as an international effort to sequence, by the year 2005, all the 60000 to 80000 human genes in an attempt to make them accessible for biomedical handling. By the time this essay is published, about a third of the human genome will have been accurately sequenced. A decade into the project, this may seem little, but at the current rate of increase it would appear that the Genome Project will be completed at least two years earlier than originally planned. The new target date in 2003 would fittingly coincide with the 50th anniversary of Watson and Crick's description of DNA, the fundamental structure of our genes.

The Physiome Project should be viewed as both a vision and a route. It has been portrayed as consisting of two parts (Bassingthwaite *et al.* 1998): (i) the databasing of biological information (the 'Mechanic's touch'), and (ii) the development of descriptive and, ultimately, analytical models of biological function (the 'Orbiter's view'). These are by no means sequential stages of the development.

The Physiome Project will undoubtedly benefit from lessons learned during the progress of the Genome Project, in particular, that big visions and small steps (at least initially) are not necessarily a contradiction. It will, however, have to develop a completely different approach to problem solving than that used for the Genome Project, as neither the total dimension of the task (there are 'only' 23 human chromosome pairs) nor the size of the smallest component that needs investigating (DNA bases) can be defined at the outset of the Physiome Project.

Another difference from the Genome Project is that there will not necessarily be a concerted effort along the whole breadth of the problem. Biological function may be modelled at any level of functioning – from protein folding to neuronal networks – and for any tissue, organ or organ system. Existing examples range from hepatocytes, and pancreatic beta cells, to muscle fibres, neurones, receptors, etc. Despite this breadth, the Physiome Project has developed its first significant foundation in the cardiovascular field.

The reasons for this are diverse and include the fact that models of cardiac cellular activity were among the first cell models ever developed. Analytical descriptions of virtually all cardiac cell types are now available. Also, the large-scale integration of cardiac organ activity is helped immensely by the high degree of spatial and temporal regularity of functionally relevant events and structures, as cells in the heart beat synchronously.

The Physiome Project will build on linking descriptions of biological function and structure. On a macroscopic level, this will benefit from another on-going large-scale research effort – the *Visible Human* Project. This is an expansion of the 1986 long-range plan of the US National Library for Medicine to create anatomically detailed, three-dimensional representations of the human anatomy. The project is based on collecting transverse computer tomography, magnetic resonance, and cryosection images at 0.5–1 mm intervals. This spatial resolution is sufficient to develop initial models of biological function, in particular where these are related to macro-mechanics or passive electrical properties. A finer resolution will, however, be required in the context of anatomico-functional modelling at tissue level and, almost certainly, when addressing inter-cellular or sub-cellular events.

8.3.2 The route

So much about the vision – what about the route? The Physiome Project will – like the Genome and Visible Human projects – crucially depend on the ability to develop the necessary tools for its own successful implementation. Apart from obtaining useful data and building representative databases, this primarily includes the capacity to devise appropriate algorithms to model physiological function.

But – why model?

The concise Oxford dictionary of current English defines a model as ‘*a simplified . . . description of a system etc., to assist calculations and predictions*’. One can apply this definition in its wider sense to any intellectual activity (or its product) that tries to make out the components of a system and to predict the outcome of their interaction. Thus, to think is to model (beware, though, that the reverse is not necessarily true).

To implement the Physiome Project, a lot of ‘good science’ (Wolpert) and ‘thinking’ (Dover) will be required. The tools that will ultimately define the success of the project are analytical models of biological processes that have *predictive* power – virtual cells, tissues, organs and systems.

This will extend, and partially replace, the traditional approach to biomedical research that is based on studying living cells or tissues *in vitro*, or on obtaining data from human volunteers *in vivo*, by introducing ‘*in silico*’ experiments (a term, derived from the currently prevailing silicon-based computer chips).

8.3.3 The tools

The Physiome Project's *in silico* models are based on and validated against solid experimental data. Much of the 'input' data is already available from many decades of bio-medical research. More will follow and, with the development of new experimental tools and technologies, the insight into sub-cellular, genetic and molecular levels of biological activity is becoming increasingly detailed. Virtual biological systems will be produced by describing in great detail the constituent parts *and* their interrelation according to the laws of conservation of energy, mass, and momentum.

Such models can be used to perform *in silico* experiments, for example by monitoring the response of a system or its components to a defined intervention. Model 'output' – predictions of biological behaviour – is then validated against *in vitro* or *in vivo* data from the real world.

A confirmation of the modelling-derived predictions would allow the performance of new *in silico* experiments, either with a higher degree of confidence or at a higher level of functional integration. Rejection of model output would help to pinpoint where the model needs refinement, either by providing new input data, or by direct model improvement. Subsequently, the *in silico* experiment could be repeated with a higher degree of confidence, until the model satisfactorily reflects the tested aspect of reality.

This is a steady iterative process between the virtual organ and the real thing. Its prime objectives are the development of our understanding of a biological system like the heart, and the improvement of its *in silico* description. Through this multiple iteration, virtual organ models mature towards a tool that can be used with a high degree of confidence for research, development or clinical applications by scientists and doctors who do not need to be specialists in model development or validation.

8.4 The virtual heart¹

8.4.1 Science or fiction?

. . . A patient who recently recovered from a minor heart attack is suffering from periods of ectopic ventricular beats, originating from what is believed to be a small area of post-ischaemic fibrosis in the free wall of the left ventricle.

¹ This section will contain some of the more technical aspects of bio-mathematical modelling (in-depth information can be found in Kohl *et al.* 2000). The subsequent section on 'The utility of virtual organs' will address more general aspects that can be appreciated without knowledge of the detail presented next.

The extent and localisation of the area is investigated and confirmed, using catheter impedance tracking of ventricular wall tissue properties and non-invasive monitoring of cardiac dimensions and relative catheter location. A small area of increased fibrosis is diagnosed and mapped in real time to a patient-specific 3D virtual heart model. The model is then used to assess treatment strategies. A decision is taken by the surgeons to ablate the area. Using the virtual heart, optimal pattern and localisation for the tissue ablation are established with the aim of maximising the anti-arrhythmic effect while minimising the energy levels of the procedure. Using the same catheter, a minimal tissue area is ablated, obliterating the ectopic focus and terminating the arrhythmia. The whole, minimally-invasive procedure took only 12 minutes, and the patient made – as typical for 97 per cent of cases – a full recovery . . .

Cardiac models are amongst the most advanced *in silico* tools for bio-medicine, and the above scenario is bound to become reality rather sooner than later. Both cellular and whole organ models have already ‘matured’ to a level where they have started to possess predictive power. We will now address some aspects of single cell model development (the ‘cars’), and then look at how virtual cells interact to simulate the spreading wave of electrical excitation in anatomically representative, virtual hearts (the ‘traffic’).

8.4.2 Single cell models

The most prominent expression of cardiac activity is the rhythmical contraction of the heart – its pumping action. Less well known is the fact, that this mechanical activity is tightly controlled by an electrical process called ‘excitation’.

In the normal heart, electrical excitation originates in specialised pacemaker cells and spreads as an electrical wave throughout the whole organ. This electrical signal determines the timing and, to a degree, the force of cardiac contraction. Thus, the heartbeat is a consequence of an electrical process (which does, however, go completely unnoticed in day-to-day life). Modelling of the heart’s electrical activity has a long history. In 1928, two Dutch engineers, van der Pol and van der Mark, described the heartbeat by comparing it to a simple oscillator. This approach, which was revolutionary at the time, gave rise to a whole family of models of the heartbeat and of the operation of other periodically active, electrically excitable cells (like neurones or skeletal muscle cells).

A common denominator of these models is the attempt to represent

cellular electrical activity by describing, with a very small number of equations, the time-course of changes in the electrical potential in the cells (Figure 8.1(a)), but not of the ionic currents that gave rise to it.

This approach is, at the same time, the great advantage and a major limitation of membrane potential models. As they are rather compact, models of this type were the first to be used in investigations of the spread of excitation in multi-dimensional 'tissue' representations consisting of relatively large numbers of interconnected excitable elements; their role in assessing biophysical behaviour like cardiac impulse propagation is undiminished.

The major drawback of these models, however, is their lack of a clear reference between model components and constituent parts of the biological system (e.g. structures like ion channels, transporter proteins, receptors, etc.). These models, therefore, do not permit the simulation of patho-physiological detail, such as the series of events that follows a reduction in oxygen supply to the cardiac muscle and, ultimately, causes serious disturbances in heart rhythm.

A breakthrough in cell modelling occurred with the work of the British scientists, Sir Alan L. Hodgkin and Sir Andrew F. Huxley, for which they were in 1963 (jointly with Sir John C. Eccles) awarded the Nobel prize. Their new electrical models calculated the changes in membrane potential on the basis of the underlying *ionic currents*.

In contrast to the pre-existing models that merely portrayed membrane potentials, the new generation of models *calculated* the ion fluxes that give rise to the changes in cell electrical potential. Thus, the new models provided the core foundation for a mechanistic description of cell function. Their concept was applied to cardiac cells by Denis Noble in 1960.

Since then, the study of cardiac cellular behaviour has made immense progress, as have the related 'ionic' mathematical models. There are various representations of all major cell types in the heart, descriptions of their metabolic activity, its relation to cell electrical and mechanical behaviour, etc. Drug-receptor interactions and even the effects of modifications in the genetic information on cardiac ion channel-forming proteins have begun to be computed. Principal components of cell models of this type are illustrated in Figure 8.1(b) on the example of work by the Oxford Cardiac Electrophysiology Group. As one can see, great attention is paid to the implementation of vital (sub)cellular mechanisms that determine function.

These detailed cell models can be used to study the development *in time* of processes like myocardial ischaemia (a reduction in coronary blood flow that causes under-supply of oxygen to the cardiac muscle), or effects of genetic mutations on cellular electrophysiology. They allow to predict the outcome of changes in the cell's environment, and may even be used to assess drug actions.

8.4.3 Organ models

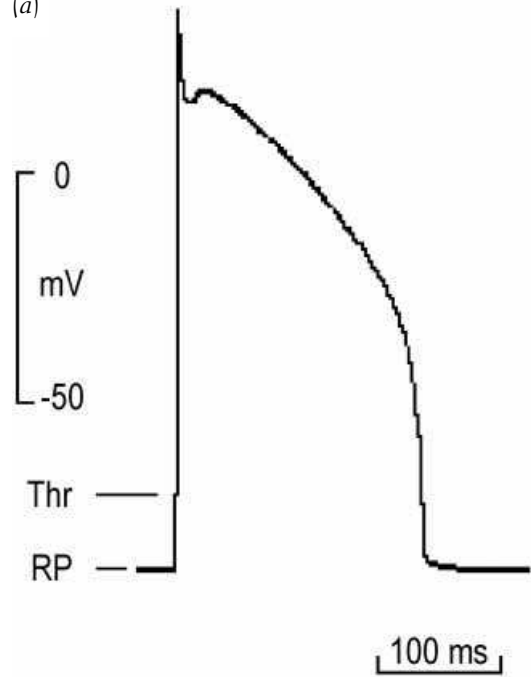
Clearly, cardiac function may not be addressed exclusively on the basis of describing the working mechanisms of single cells. Both normal and disturbed heart rhythms are based on a *spreading wave* of electrical excitation, the meaningful investigation of which requires conduction pathways of at least hundreds if not thousands of cells in length.

These may be produced by grouping together multiple cell models to form virtual tissue segments, or even the whole organ. The validity of such multi-cellular constructs crucially depends on whether or not they take into account the heart's fine architecture, as cardiac structure and function are tightly interrelated.

Modern representations of the virtual heart, therefore, describe structural aspects like fibre orientation in cardiac muscle, together with the distribution of various cell types, active and passive electrical and mechanical properties, as well as the coupling between cells. This then allows accurate reproduction of the spread of the electrical wave, subsequent contraction of the heart, and effects on blood pressure, coronary perfusion, etc. It is important to point out, here, that all these parameters are closely interrelated, and changes in any one of them influence the behaviour of all others. This makes for an exceedingly complex system.

The example in Figure 8.2 illustrates a combination of 'only' two subsystems to study the effects of cardiac contraction on coronary tree architecture and function. Models like this allow one to determine changes in coronary blood pressure during the cardiac cycle of contraction and relaxation. Like in the real heart, the coronary tree moves with the cardiac tissue, into which it is embedded, and the pressure inside the vessels changes with the external compression by the contracting muscle. This external pressure is calculated and shown with the deforming coronary vessel tree in Figure 8.2. Thus, current electro-mechanical models of ventricular anatomy and function allow one to describe coronary perfusion during the cardiac cycle. By linking this to the models of cell metabolism

(a)



(b)

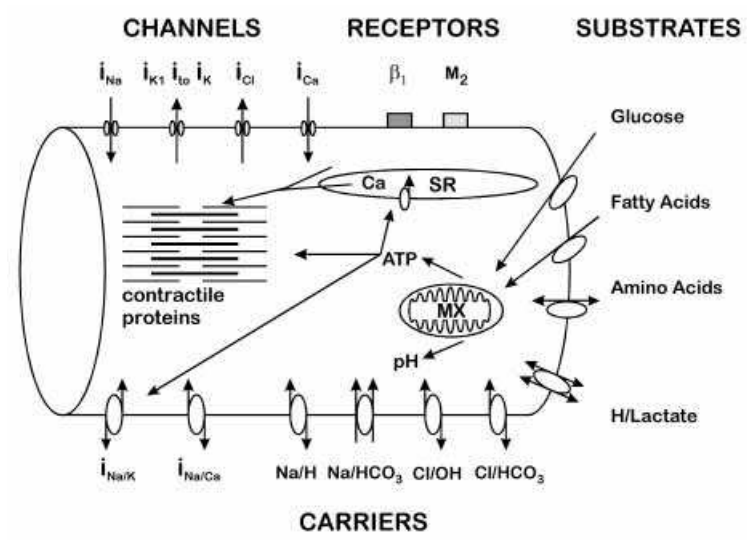


Figure 8.1. Scheme of a ventricular action potential (a) and its sub-cellular mechanisms (b). Membrane potential models simulate action potentials (a) with a deliberately small number of equations; ionic current models reproduce the action potential on the basis of calculating the sub-cellular ion movements that actually give rise to it (b). (a) Cardiac contraction is controlled by an electrical waveform called action potential. Action potentials are induced by change in cell voltage to less negative values. Cells are said to 'depolarise' from their resting potential (RP) towards a threshold (Thr), at which automatic excitation occurs: an action potential is initiated. The action potential is characterised by a swift upstroke to positive cell voltages, followed by a plateau and slower return to RP levels. The well-ordered spread of this waveform provides the basis to the regular contraction of the heart. (b) Example of major constituent parts of a detailed ionic current model (here Oxsoft Heart v4.8). The model incorporates essential intracellular structures like the contractile proteins, sarcoplasmic reticulum (SR, a calcium store) or mitochondria (MX, the powerhouse of the cell). It computes the action potential as a function of ion movements through channels (see a selection, top left), exchangers and pumps (bottom). This makes it possible to predict the cell's electrical and mechanical activity, and to account for effects of receptor stimulation (see selection at top right: adrenergic – β_1 , and cholinergic receptors – M_2 , that provide neural input), or changes in substrate transporter activity, cell metabolism and pH (right hand side). With this type of models, (patho-) physiological behaviour may be simulated as it develops in time. From Kohl *et al.* 2000.

and electro-mechanical function, the whole sequence of the natural heart-beat may be reproduced.

The same applies to pathologically-disturbed function. A simulated reduction in coronary blood flow (heart attack) would lead to reduced oxygen supply to the cells in the virtual heart, which would reduce efficiency of cardiac contraction and possibly give rise to heart rhythm disturbances. Ventricular pressure development would be compromised, as would the blood supply to all organs of the body, including the heart. All these implications can be studied in a virtual heart.

This possesses an immense potential, not only for bio-medical research, but also for clinical applications, including patient-specific modelling of therapeutic interventions. For example, dynamic changes in a patient's cardiac anatomy can already be modelled on the basis of a non-invasive technique called Magnetic Resonance Imaging. The location of coronary vessels for that patient may be determined by 3D coronary angiography, a common procedure in the context of coronary surgery, and implemented into the virtual heart. Integrated patient-specific virtual hearts of this kind may, therefore, already be constructed – however, not in real time yet. Once the interrelation between computational demand and computing power has sufficiently improved (and it does get better all the time), virtual hearts will be used for the prediction of optimal coronary bypass procedures, aid the surgeon's decision on the operative approach, and even predict the potential long-term consequences of various treatment strategies.

So – what about the prospects of powerful computational equipment? The calculations for Figure 8.2, for example, took about six hours on a 16-processor Silicon Graphics Power Challenge, which is a fairly powerful computer. Six hours to calculate a single heartbeat! However, there are already much more powerful systems that could cope with the same task in less than an hour. In December 1999, IBM announced the development of a new generation of super computers (Blue Gene), and other major systems manufacturers will have similar projects in the pipeline. The new computer generation is said to provide a 500-fold increase in maximum computing power over the next four to five years. Computation of the above model would then be possible in real time.

Such are the prospects.

Patient-specific and modelling-aided clinical treatment could, therefore, become a reality within the first decade of this millennium!

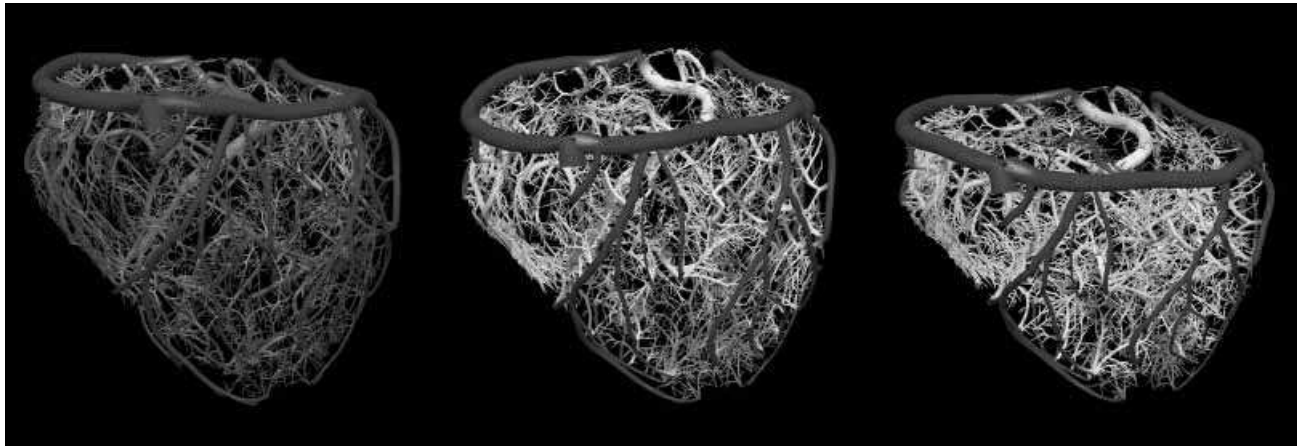


Figure 8.2. Coronary vasculature (shown) coupled to the deforming myocardium (transparent). The grey-level coding represents the pressure acting on the coronary vessels during the myocardial contraction (dark – zero pressure, light – peak pressure). The deformation states are (from left to right): resting state, early contraction, and peak contraction. Courtesy of Dr Nic Smith, University of Oxford.

8.4.4 Simulating the ECG

To date, the most common tool for clinical assessment of cardiac electrical function is the electro-cardiogram (ECG). It is a dynamic representation, usually obtained from the body surface, of the changes in cardiac electrical behaviour.

While the ECG is an invaluable tool for the observation of heart rate and rhythm, as well as for the diagnosis of conduction abnormalities, ischaemia, and infarcts, its detailed interpretation is not without pitfalls. One reason for this is that different changes in cardiac cellular behaviour may give rise to very similar effects on the ECG. This makes it difficult to draw conclusions from a patient's ECG to the underlying (sub-)cellular mechanisms. This issue is usually referred to as the 'inverse problem'.

Today's heart models do not yet possess the power to solve the inverse problem. They do, however, aid the understanding and interpretation of the ECG by repeatedly solving 'forward problems' to study the effects of cellular modifications on the calculated ECG. Model reconstruction of a normal ECG is therefore a necessary first step towards developing a better understanding of the information 'hidden' in it. Figure 8.3(a) illustrates this.

The same may be applied to simulations of the ECG in pathologies. Here, we illustrate work on simulating the typical ECG of patients with congestive heart failure (CHF), a disease that affects roughly 1 per cent of the population in Western countries and causes a reduction in cardiac output. While therapeutic advances have reduced mortality from pump failure, they have been relatively ineffective in reducing the incidence of sudden cardiac death caused by disturbances in the heart's electrical activity. The virtual heart can be used to identify promising targets for pharmacological interventions in CHF.

CHF is accompanied, at the cellular level, by changes in the content of proteins that govern electrical repolarisation and cellular calcium handling. This threatens orderly repolarisation of cardiac tissue by increasing the likelihood of spontaneous 'early after-depolarisations' that can initiate an irregular heartbeat. Figure 8.3(b) illustrates the effect of CHF-typical cellular changes on the computed electrical activity of the heart and the ECG. The virtual heart shows the characteristic pattern of rhythm disturbance observed in CHF patients, together with the distinctive saw-tooth like ECG. The circulating waves of electrical excitation prevent the heart from

relaxing between beats, which impedes the filling of the cardiac chambers and prevents effective pumping function.

Might this be reversed by pharmacological interventions? Figure 8.3(c) illustrates the application of a (virtual) drug that specifically activates one type of the potassium channels in the heart (the so-called ATP-modulated potassium channel, whose activity is increased in the model from 0 to 0.0002). This intervention leads to termination of the dangerous depolarisations at the cellular level and allows the whole heart to regain a stable resting state (compare last frames of the sequences in Figure 8.3(b) and (c)). Thus, while the pattern of impulse conduction in the heart has not entirely normalised, the development of fatal re-entry is terminated.

Thus, the virtual heart may be used to simulate cardiac pathologies, their effect on the ECG, and the consequences of drug administration. It can be seen that drug discovery and assessment will be among the first fields where *in silico* technologies could reform research and development in a whole industry.

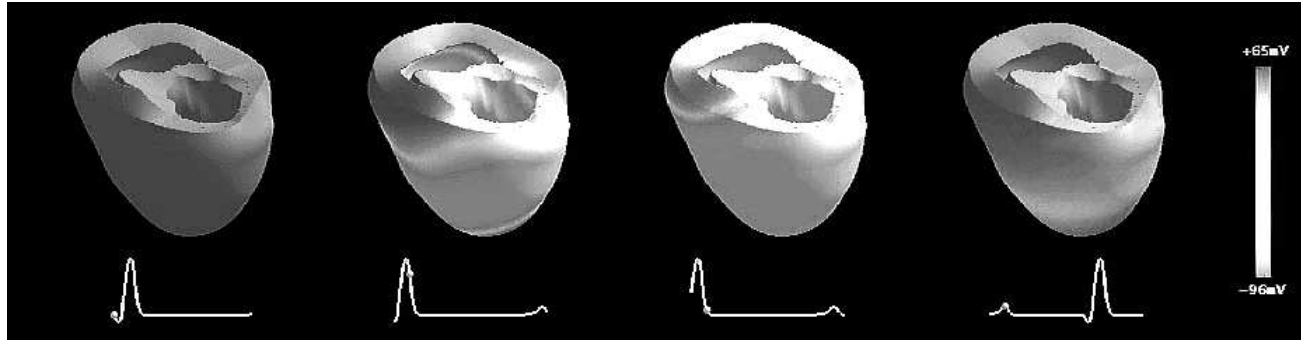
8.4.5 Summary: The virtual heart

Analytical models of the heart are a reality. They are based on detailed descriptions of cardiac tissue architecture and anatomy, including the coronary vasculature. *In silico* cardiac tissues possess realistic passive mechanical properties, and both electrical and mechanical activity can be simulated with high accuracy. Descriptions of key components of cellular metabolism have been introduced, as have models of drug-receptor interactions.

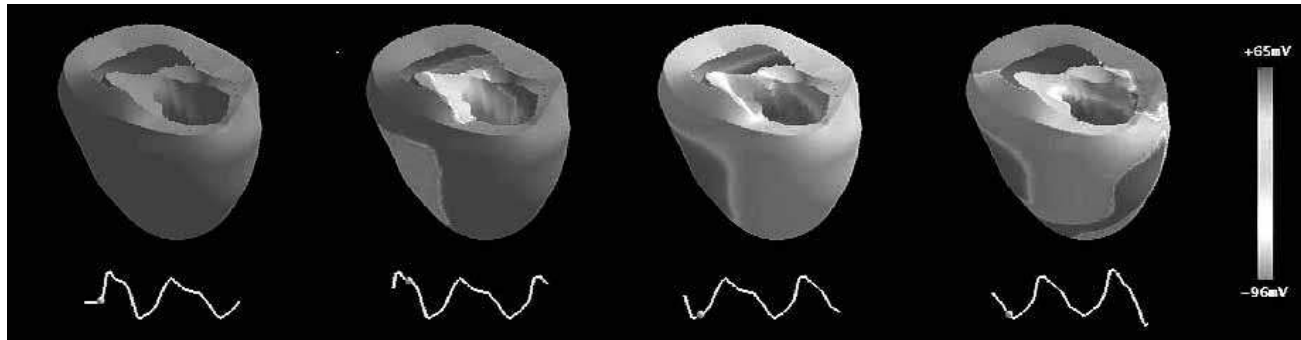
The individual modules of the *in situ* heart can be coupled together to compute a whole sequence from ventricular pressure development, coronary perfusion, tissue supply of metabolites, cell energy consumption, and electrophysiology, to contractile activity and ventricular pressure development in the subsequent beat. The 'starting point' (here chosen as ventricular pressure development) can be freely selected, and drug effects on the system can be simulated. 'Inserted' into a virtual torso, these models allow one to compute the spread of excitation, its cellular basis, and the consequences for an ECG under normal and pathological conditions.

Ongoing work is devoted to the accurate description of the origin and spread of excitation from the natural pacemaker to the rest of the heart. Computations of ventricular pressure development are being extended to account for blood flow dynamics in adjacent blood vessels. The thorax

(a)



(b)



(c)

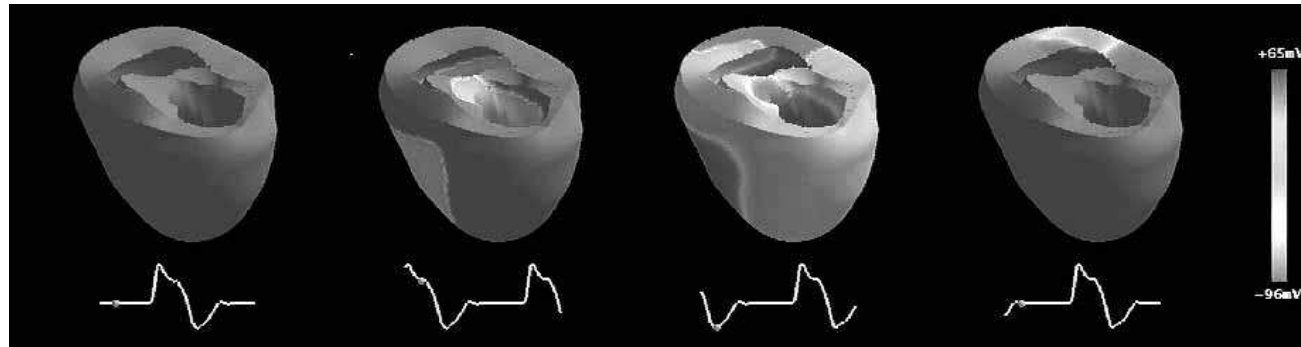


Figure 8.3. Simulation of the spread of excitation in canine ventricles. Ventricular cell models are based on a simplified version of the Oxsoft v.4.6 ionic models. Membrane potentials are grey-level coded (dark – resting potential, light – action potential) and ECG equivalents are computed (curves below the images). (a) Normal spread of excitation. Frames illustrate the normal sequence of excitation and repolarisation during one cardiac cycle (from left to right). (b) Spread of excitation in a congestive heart failure model. The initial activation sequence (frames 1 and 2) is followed by irregular re-entrant excitation (frames 3 and 4). Note the typical, for this pathology, saw-tooth shaped ECG. (c) Simulation of the effect of ATP-modulated potassium channel openers on the spread of excitation in the same congestive heart failure model. The first three frames are closely reminiscent of those leading to re-entrant excitation in (b), with the saw-tooth like ECG shape still apparent. Due to the drug effect, however, the heart does reach a resting state before a new cycle of cardiac excitation is triggered ('dark' cardiac chamber and 'flat' segment in the ECG, frame 4). This allows time for diastolic filling and permits pumping action of the heart. From Kohl *et al.* 2000.

representation is being developed to allow simulation of respiratory movement, and the computation of pulmonary ventilation and gas exchange is well underway. Thus, the stage for patient-specific models is set.

8.5 The utility of virtual organs

8.5.1 Added value for research

Virtual organs will increasingly determine bio-medical research. Advantages of *in silico* models include the following:

- Complex investigations, for example on the (sub)cellular level, can be performed in a fraction of the *time* required for 'wet' (*in vivo* or *in vitro*) studies.
- The *costs* involved are much smaller than for traditional research. This applies not only to direct financial aspects, but also to requirements in terms of human resources, and to ethical matters related, for example, to the origin of 'wet' tissue or organ samples.
- The *quality* of information benefits from the fact that interventions and observations can be specifically targeted at any component or mechanism represented in the model, and at any desired temporal and spatial resolution.
- While the first three points improve the quantity and quality of information, *in silico* models benefit further from their unrestricted potential for customised presentation of results. This allows addressing aspects like individual preferences in information gathering, remote usage of models, interactive teaching and training, etc.

So much for the advantages. Virtual organs clearly have one major drawback: they are models only. While this very nature of *in silico* technology is *the* core foundation for the benefits listed above, it also calls for a word of caution. It is imperative for *in silico* tools to be seen in the context of a whole range of scientific and research tools, and to never neglect that theoretical considerations will continue to need experimental validation.

Thus, *in silico* models are by no means self-sufficient. They are irreplaceable for the future progress of bio-medicine. They do not aim, however, to substitute but to improve bio-medical research, which will remain indispensable, not the least for model development and validation.

8.5.2 Added value for drug and device development

Drug development is currently largely based on trial and error. This is an exceedingly time-consuming process, and some of the associated errors have proved quite costly for patients involved. Even if distressing consequences of clinical testing could be avoided, the economical costs of bringing a new drug to market are prohibitive: close to US\$0.5 billion.

Also, the fact that only an estimated 10 per cent of pre-clinically tested lead-compounds are likely to ever reach the market must discourage companies from investing into new drug development, in particular for pathologies that are not deemed to constitute a profitable market. Thus, from the point of view of a commercial drug developer, ideal targets are chronic and non-lethal complaints that affect people in the developed world at the prime of their financial viability. In other words, it is 'more economical' to come up with a treatment for obesity, baldness or impotence, rather than to tackle a rare but lethal disease that affects small patient groups or people in underdeveloped regions of the world.

Analytical computer models clearly have the potential to improve this situation, as they may help:

- to speed-up drug development by *in silico* screening for early identification of promising lead compounds;
- to simplify the assessment of complex pre-clinical data and predict (patho-)physiological (side-)effects of drugs;
- to cut the associated financial and ethical costs;
- to reduce the risk of clinical testing.

The above may not be enough, though, as it will be crucial to change the whole approach to drug development. What is needed is a method to identify the desired drug effect and (sub-)cellular target for pharmacological intervention *before* directed compound synthesis and testing commence.

Virtual organs will form the basis for this novel approach.

Similar concepts apply to the world of medical devices. In future, successful products will increasingly be tuned to flow with the stream of human physiological function, even to mimic it in fine detail. Modelling and computation are set to make major contributions, since:

- devices become sufficiently 'intelligent', with their on-board computing power, to use analytical descriptions of (patho-)physiological organ function;

- accurate and efficient modelling of body functions provides a test-bed for the development of devices that are energy-efficient and less invasive;
- specialist equipment is easier to (re)produce and more portable than the 'specialists' themselves.

Future medical training, diagnosis and – even surgical – treatment will increasingly be performed remotely. Thus, the combination of sophisticated sensory devices with advanced micro-manipulation equipment will, together with 3D 'interactive feedback' models, provide new tools and approaches for the medical profession.

8.5.3 Added value for society

The proper study of mankind is man

(Alexander Pope, *An essay on man*, 1733)

In silico technology is set to produce a quantum leap in our understanding of the nature of man, for it is only through the identification of useful information in the vast amount of data on 'man' that we will arrive at a genuine comprehension of our biological nature.

Analytical bio-modelling is also set to make major practical contributions and to transform the way society handles health-related matters. The 'added benefit' of *in silico* technologies for health care includes the following:

- New, interactive *in silico* teaching and educational tools will be available for doctors and the greater public. This will help to improve professional skills and general health awareness. Future health-related implications of an individual's behavioural patterns or of various treatment strategies can be assessed and compared on the basis of long-term case predictions.
- *In silico* technologies will help health care policy development and acute decision making. The latter will be based on improved access to expert information, statistics, case reports, etc. Medium-term decisions will benefit from the early recognition of epidemiological patterns, etc. Long-term policies can be based on detailed investigations into the cost-benefit-relation of restorative versus preventative strategies which, undoubtedly, will consolidate the case of preventative medicine.

- *In silico* models will aid both the standardisation and individualisation of medical care. Standardisation of diagnoses, drug and device descriptions, procedures, etc. will make relevant information more readily and more widely available. On the other hand, advanced models will allow development of patient-specific procedures for diagnosis and treatment. This will move the focus from the treatment of diseases to the curing of patients.
- All the above effects will ultimately lead to reduction in morbidity and mortality and an improvement in the quality of life.

On this optimistic note we shall finish.

8.6 Further reading

- Bassingthwaighte, J. B., Li, Z. & Qian, H. 1998 Blood flows and metabolic components of the cardiome. *Prog. Biophys. Molec. Biol.* **69**, 445–461.
- Bock, G. R. & Goode, J. A. (eds.) 1998 *The limits of reductionism in modern biology*. Novartis Foundation Symposium. Chichester: John Wiley & Sons.
- Kohl, P., Noble, D., Winslow, R. L. & Hunter, P. J. 2000 Computational modelling of biological systems: tools and visions. *Phil. Trans. R. Soc. Lond. A* **358**, 579–610.



9

Exploring human organs with computers

Paul J. Kolston

*MacKay Institute of Communication and Neuroscience, University of Keele,
Staffordshire ST5 5BG, UK*

9.1 Introduction

Your body is an extraordinarily complex piece of biological machinery. Even when you are sitting still in a quiet room your blood is being pumped to every tissue, your kidneys are filtering body fluids, your immune system is guarding against infection and your senses are continuously monitoring all kinds of information from their surroundings. Scientists are very good at studying single molecules and single cells but the challenge of understanding molecular and cellular processes in entire organs or body systems is a daunting one. Fortunately, although organs are complex, they do not rely on magic. Their behaviour is controlled by basic laws of physics that can be described by mathematical equations. By solving these equations on computers, scientists are able to investigate the operation of complete organs with a flexibility and level of detail that is impossible experimentally.

This article shows how computers have already illuminated the workings of a variety of biological systems, with a particular emphasis on the operation of the ear. Soon it will be possible to represent in a computer model every single cell contained within a whole organ. These models will lead to a profoundly deeper understanding of how biological organs work, whilst reducing dramatically the need for animal experimentation. Longer term, computer modelling will provide a valuable tool in the design of new, simpler cellular structures that would mimic the known operation of a biological organ to either replace their defective counterparts or to perform

entirely new tasks with efficiencies that are impossible using current technologies. Given the impressive specifications of such organs, these new devices – manufactured in carbon or silicon – could have numerous research, clinical and industrial applications.

9.2 Making cars

Computer modelling has been used extensively in manufacturing industries for many years. One familiar application is in the design of car crashworthiness. Cars must both protect the occupants from physical intrusions into the passenger compartment, and minimise the deceleration forces that act upon them. The first of these requirements could be achieved easily by making the car body rigid. Unfortunately, the deceleration forces would then be intolerably large, so instead the design aim is to make those parts of the vehicle that are outside the passenger compartment absorb as much of the impact energy as possible, by making them deform in the pre-defined time-dependent manner that minimizes peak deceleration levels.

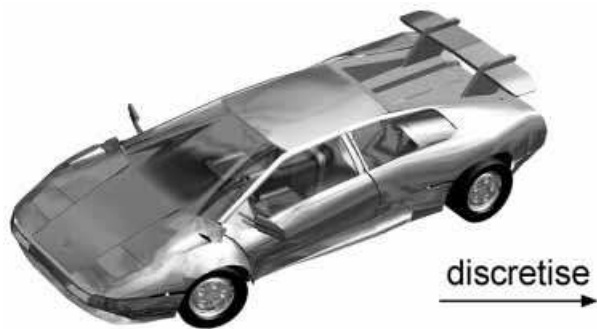
In the past when car crashworthiness was designed entirely experimentally, full-sized prototypes were subjected to the crash scenarios required by the relevant authorities. If the performance was unacceptable, the shape deformations of the components making up the prototype were examined. A new prototype was engineered empirically to overcome the identified weaknesses before being built and then destroyed in a subsequent test. These tests would be repeated many times before an appropriate design was found. The cost of the process was enormous.

Nowadays, car manufacturers cannot afford to destroy thousands of prototypes when designing crashworthiness into their vehicles. Instead, they spend most of their time building and analysing models on computers, using a technique known as finite-element analysis (Figure 9.1). Only once a computer model is found to be consistent with statutory requirements do they resort to expensive and time-consuming physical testing. This beneficial relationship between modelling and experimentation is still in its infancy in biological research, thanks partly to the great complexity of biological organs.

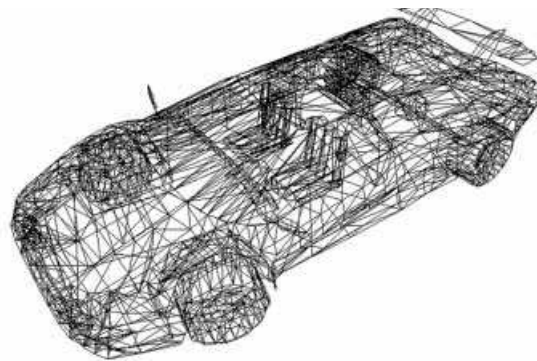
9.3 Designing drugs

Car crashworthiness design involves the manufacture of a new system, but each stage of the process requires an understanding of the operation of an existing system. This is analogous to most research in biology. However, in contrast to crashworthiness design, investigations into the operation of biological organs are still dominated by experimental approaches. There are some exceptions, such as in the development of therapeutic drugs to combat disease. In the past this was performed purely in a brute-force trial-and-error manner. Cell cultures, animals or humans were subjected to many variations of a likely candidate for a drug, with the final choice being chosen on the basis of best performance with the minimum adverse side effects. This is analogous to building thousands of car prototypes simultaneously, each with slight differences in design, and then subjecting them all to the rigors of experimental crash testing. The best model is that which, largely by chance, survives best. If car crashworthiness was still designed in this way, only the very rich would be able to afford the end product. Fortunately, computer models are now being used to 'experiment' with the effects that changes in structure will have on the potency of the drug, with corresponding reductions in production costs.

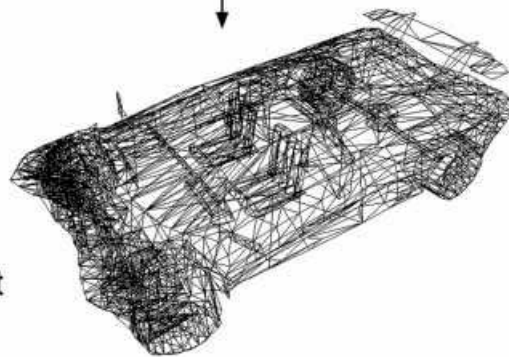
Each organ in the human body plays a crucial life-sustaining role, and understanding how each works is of profound interest for many reasons, from the possibility of widespread treatment, or even the prevention of disease, to the possible engineering applications of the unique types of signal processing employed by each organ. It is natural to describe the function of an organ, and hence model its behaviours, in terms of the components at the next level down in the biological hierarchy, the cell (Figure 9.2). Both the behaviour of the cells in isolation and all of the interactions between them must be considered. A finite-element computer model that represents an organ at the level of the cell allows us to observe the individual interactions between tens of thousands of cells simultaneously. Such experiments are impossible to perform on the real system. Finite-element modelling of biological systems has already begun in a number of areas, including bone, skin and brain mechanics, intercellular communication within tissues, and heart contraction.



discretise
→



simulate
↓



reconstruct
←



Figure 9.1. One important factor in the design of a car is protecting its occupants during collisions, known as crashworthiness. One aim is to minimise the deceleration forces that act upon the occupants, which is achieved by making the car body absorb impact energy by deforming in a particular way. But large deformations will result in physical intrusions into the passenger compartment, which are also undesirable. The trick is to balance these two opposing factors, rigidity and deformability, in an efficient way. The most cost-effective way of doing this is to use finite-element modelling on a computer. The intact car is first divided into a large number of small elements (discretisation). The relatively simple equations that describe the physical interactions between adjacent elements are then formulated, taking into account the material properties of the elements. This set of equations is then solved whilst an appropriate stimulus is applied. In this example, the stimulus is an obstacle at the front of the vehicle (simulation). The resulting deformations and deceleration forces are then investigated in detail, in order to predict how best to modify the design in order to improve its performance (reconstruction).

The beauty of finite-element modelling is that it is very flexible. The system of interest may be continuous, as in a fluid, or it may comprise separate, discrete components, such as the pieces of metal in this example. The basic principle of finite-element modelling, to simulate the operation of a system by deriving equations only on a local scale, mimics the physical reality by which interactions within most systems are the result of a large number of localised interactions between adjacent elements. These interactions are often bi-directional, in that the behaviour of each element is also affected by the system of which it forms a part. The finite-element method is particularly powerful because with the appropriate choice of elements it is easy to accurately model complex interactions in very large systems because the physical behaviour of each element has a simple mathematical description.

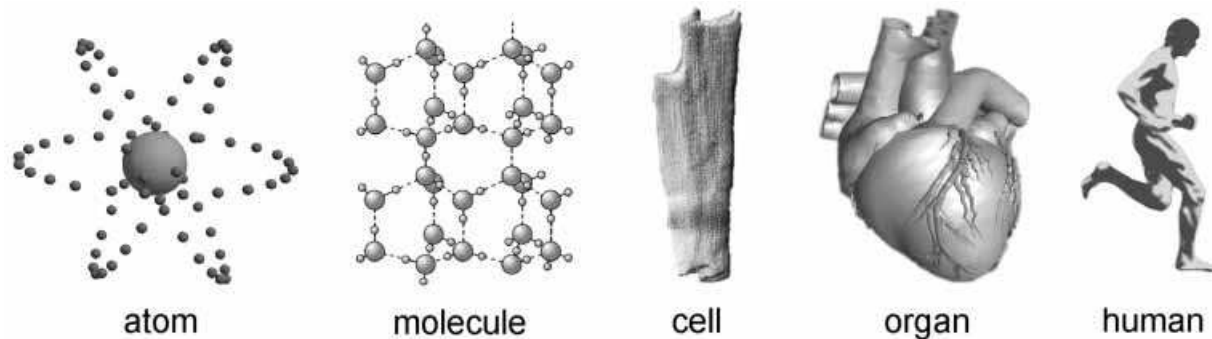


Figure 9.2. The biological hierarchy. Assemblies of carbon, hydrogen and oxygen atoms combine to form molecules that, when linked together, form proteins, carbohydrates and fats. These three fundamental types of biological molecule combine to form cells that are typically 0.01 mm in diameter and have the mass of 30 000 000 000 000 hydrogen atoms. Cells can be independent organisms, as in a bacterium, or, by co-operating with other cells, form tissues. By acquiring specialised functions, assemblies of tissues form the next distinct structural and functional unit, the organ. At the highest level, a human comprises 75 000 000 000 000 cells divided into ten major organ systems.

It is natural to describe the function at each level in the biological hierarchy in terms of the components at the next level down. Sometimes it is necessary to consider processes occurring two levels down, but further subdivision is seldom beneficial. Schrödinger's equation, for example, is useful when modelling the behaviour of atoms in a molecule, but it would be absurd to model car crashworthiness using this level of detail. When we are interested in the operation of a complete organ, a description at the level of the cell is the natural choice. The model must incorporate both the operation of the cell in isolation *and* the interactions between cells since, by analogy, we could not predict the load-bearing capacity of the Forth Rail Bridge by considering only the strength of the individual cantilevers in isolation.

9.4 Bone and skin

Perhaps the most obvious biological application of finite-element modelling, given the popularity of the technique in mechanical engineering, is in bone mechanics. The structural properties of bone are determined by non-cellular organic and inorganic components. It is only these components that are included in the simplest models. The potential exists to assess quantitatively an individual patient's risk of bone fracture, which has significant clinical implications in an ageing population. Currently, estimates of this risk are limited by the inability to allow for complex structural features within the bone. However, if the internal structure of a bone was determined *in vivo*, using X-ray-based computed tomography, an accurate finite-element model could be built to estimate the maximum load that can be borne before fracture. Finite-element models can aid in surgical spine-stabilisation procedures, thanks to their ability to cope well with the irregular geometry and composite nature of the vertebrae and intervertebral discs.

The acellular structure of real bone is modified continuously according to the internal stresses caused by applied loads. This process, which represents an attempt to optimize the strength-to-weight ratio in a biological structure, is achieved by the interaction between two types of cell, one that absorbs bone and the other that synthesises new bone. New bone is added where internal stresses are high, and bone is removed where stresses are low. An accurate finite-element model of this combined process could be used clinically to determine the course of traction that will maximise bone strength after recovery from a fracture.

Another well-established area of mechanical finite-element analysis is in the motion of the structures of the human middle ear (Figure 9.3). Of particular interest are comparisons between the vibration pattern of the eardrum, and the mode of vibration of the middle-ear bones under normal and diseased conditions. Serious middle-ear infections and blows to the head can cause partial or complete detachment of the bones, and can restrict their motion. Draining of the middle ear, to remove these products, is usually achieved by cutting a hole in the eardrum. This invariably results in the formation of scar tissue. Finite-element models of the dynamic motion of the eardrum can help in the determination of the best ways of achieving drainage without affecting significantly the motion of the eardrum. Finite-element models can also be used to optimise prostheses when replacement of the middle-ear bones is necessary.

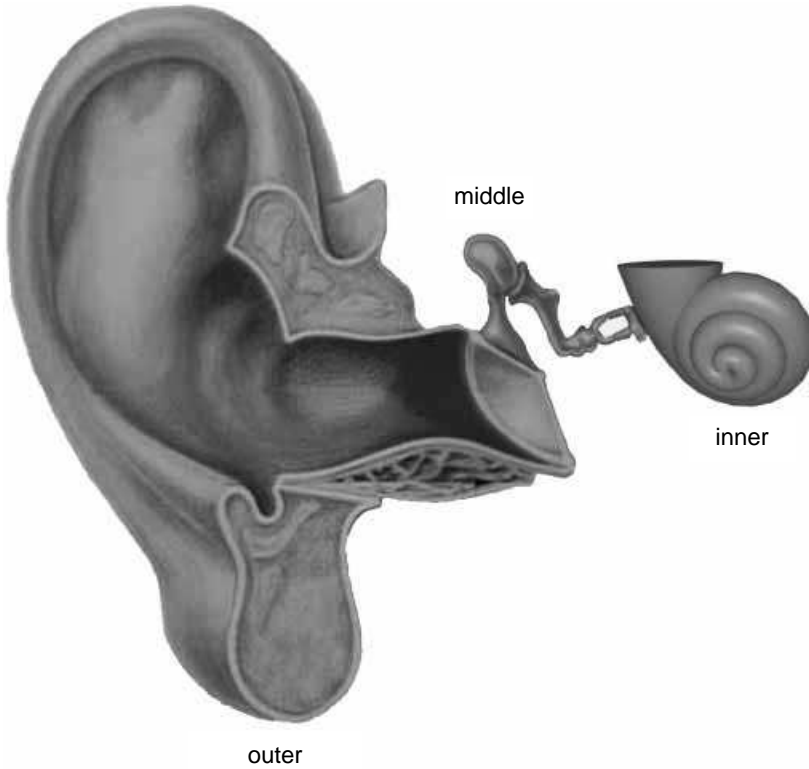


Figure 9.3. The human ear is divided into three main parts. The outer ear collects sound and directs it down the ear canal towards the eardrum. The size of the eardrum, combined with the lever action of the three bones of the middle ear, ensures the efficient conduction of sound from the ear canal, which is filled with air, to the inner ear, which is filled with a liquid. Very small muscles, not shown here, are connected to these bones to protect the ear from very loud sounds. The inner ear consists of two parts. Only the cochlea is shown, which is the part of the human ear that is responsible for converting sound into electrical signals in the auditory nerve. The other part of the inner ear, the vestibular organ, is involved in balance.

Finite-element techniques can cope with large, highly non-linear deformations, making it possible to model soft tissues such as skin. When relatively large areas of skin are replaced during plastic surgery, there is a problem that excessive distortion of the applied skin will prevent adequate adhesion. Finite-element models can be used to determine, either by rapid trial-and-error modelling or by mathematical optimisation, the best way of

covering a lesion with the available skin graft. The brain is another organ that is mechanically soft. Certain brain disorders are associated with variations in pressure in the cerebrospinal fluid that protects the brain from the hard skull. Imaging techniques can provide information about the resulting changes in brain shape, but finite-element models of the fluid-structure interactions have the potential to provide quantitative information about the forces exerted on the tissue itself.

9.5 Cell interactions

Of growing interest world-wide is the possible carcinogenic effect of low-frequency non-ionising electromagnetic radiation, such as that emitted from power lines. Possible candidates for explaining sensitivity to electromagnetic fields are the gap junctions that exist between cells in many types of tissue. These junctions are similar to the protein-based channels that enable ions to pass across cell membranes, except that they span the extracellular space between adjacent cells. Gap junctions provide cells with a direct means of intercellular communication to co-ordinate the physiology of large populations of cells.

The physical properties of gap junctions also influence the regulation of cell growth, and the cell's membrane voltage and frequency response. There is some experimental evidence to suggest that the properties of gap junctions change in the presence of electromagnetic fields. Finite-element models provide a flexible and accurate way of assessing the effects of such changes on the operation of large systems of cells.

9.6 The heart

Given that heart disease is the single largest cause of death in North America and Europe, finite-element models of the human heart have great potential clinical significance. The heart wall consists mostly of muscle, comprising millions of electrically activated contractile cells that are typically 0.1 mm long and 0.015 mm wide. Note that the cell in Figure 9.2 is from the heart. Heart contraction is activated by an electrical impulse that is generated by cells in the heart's pacemaker. This impulse spreads rapidly through the tissue due to the high degree of electrical coupling between the heart cells via gap junctions, ensuring that the whole organ contracts in a synchronised fashion.

A number of factors, including electric shock, deprivation of oxygen, or abnormally high levels of potassium or low levels of calcium in the blood, can cause malfunction of the conduction system. The resulting irregular contraction of the heart wall can be stopped by applying controlled electric shocks to the heart, either internally or externally. Patients at risk may be fitted with internal devices for supplying such shocks when they are needed. One very important design aim is to maximise battery life, thereby reducing the frequency of invasive surgery. Purely electrical finite-element models of the heart can aid the optimisation of the type of stimulation and positioning of such devices, so as to minimise the energy required to arrest irregular contractions.

Whole-body finite-element models can be used to optimise the delivery of external electric shocks to the control of irregular beating. These models can also be used in reverse, to aid in the interpretation of the skin-surface voltages induced by heart activity. Unfortunately, the body does not behave simply as a salty solution in a leathery container since, for example, the resistivity of bone is 100 times greater than that of blood. This huge variation in the resistance of the intervening tissues greatly influences how energy passes between the heart and the skin. Performing subject-specific analyses could reduce existing discrepancies between models and experiments in both types of whole-body model.

Purely electrical models of the heart are only a start. Combined electromechanical finite-element models of the heart take into account the close relationship that exists between the electrical and mechanical properties of individual heart cells. The mechanical operation of the heart is also influenced by the fluid–structure interactions between the blood and the blood vessels, heart walls, and valves. All of these interactions would need to be included in a complete description of heart contraction.

9.7 An ear model

Whilst finite-element modelling of gap junctions occurs at a sub-cellular level, these models do not consider the operation of intact organs. Conversely, in models of the complete heart the discretisation is usually on a millimetre scale. However, the cochlea (see Figure 9.3) is already being simulated on a 0.01 mm, or cellular, scale. Although cochlear malfunction is not life threatening, damage to it does adversely affect the ability of almost 1 000 000 000 people to communicate.

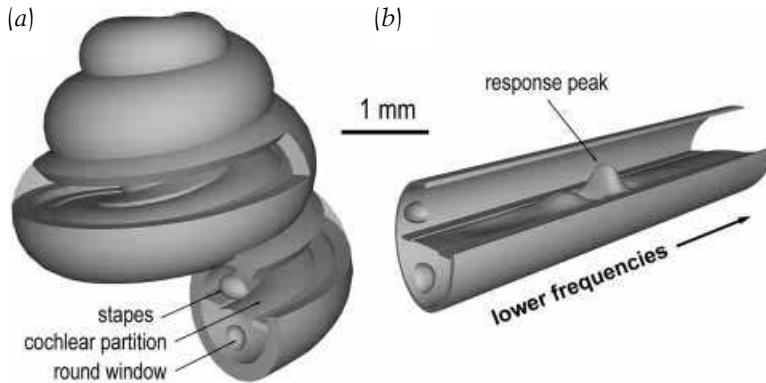


Figure 9.4. (a) The cochlea is a liquid-filled duct, typically 30 mm long and 1 mm in diameter, that is coiled into a spiral. The duct is divided lengthways into two chambers by the flexible cochlear partition. Mechanical stimulation is provided by piston-like motion of the stapes bone, which is the last bone in the middle-ear chain and is the smallest bone in your body. Stapes motion sets up a pressure differential across the partition which causes it to move. The incompressibility of the cochlear liquid means that when the stapes moves out the round window moves in, and vice versa. (b) Most models of the cochlea consider it to be straight, since the coiling is not thought to influence its operation. This figure shows the displacement of the partition at one instant in time during sound stimulation at a single frequency. Note that the displacement is shown greatly exaggerated, as near the threshold of hearing the maximum displacement is only one-millionth of a millimetre. The stiffness of the partition decreases away from the stapes, so when it is excited by a fluid pressure difference, it moves first near the stapes, since stiffer things respond more rapidly. This is followed by motion at positions progressively further away, thereby giving the appearance of a travelling wave propagating away from the stapes. As this wave travels along the length of the cochlea it increases in amplitude before reaching a peak and then dying away rapidly. The stiffness gradient also means that the position of the peak is dependent upon the frequency, so that lower frequency stimuli produce a peak further from the stapes. The cochlea thereby divides the incoming sound into its constituent pure tones. The sensory cells that convert partition motion into electrical signals in the auditory nerve are present along the entire length of the cochlea.

The human cochlea, which derives its name from the Greek word *kochlias*, consists of approximately three turns of a circular duct, wound in a manner similar to that of a snail shell (Figure 9.4). The ability to resolve different tones is determined by the pattern of vibration of the flexible cochlear partition. This partition has three main components (Figure 9.5). When

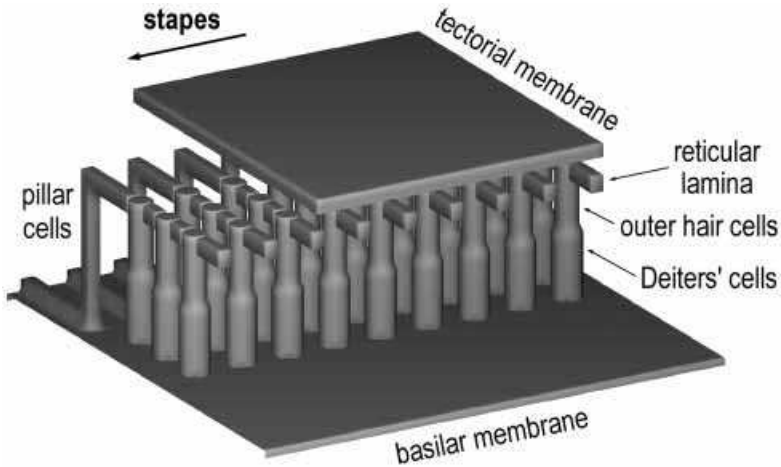


Figure 9.5. An oblique, simplified view of a 150- μm -long portion of the cochlear partition, approximately to scale. The partition has three main components, namely the basilar and tectorial membranes, and a collection of cells known as the organ of Corti. The tectorial membrane has been partially removed to reveal the tops of the outer hair cells and the reticular lamina. The organ of Corti contains two types of sensory hair cell: outer and inner, and a number of supporting cells. The inner hair cells, which are not shown here, are positioned adjacent to the pillar cells. At each position along the cochlea, the pillar cells couple vertical motion of the basilar membrane to a shearing motion between the tectorial membrane and the reticular lamina. This bends the stereociliary bundles that project from the top of both types of hair cell. These bundles contain ion channels whose electrical resistance changes with mechanical stimulation. The mechanical motion of the bundle thereby modulates the flow of ions into the cell, causing changes in the voltage across the cell membrane. This in turn modulates the release of neurotransmitter onto the nerve fibres that innervate the cell. Due to the pattern of innervation, it is the inner hair cells that are primarily responsible for providing the higher auditory centres with information about basilar membrane motion, whereby different frequencies are encoded onto different nerve fibres.

The outer hair cells have a quite different role. Experiments on isolated cells have shown that their length is proportional to the voltage across the cell membrane. As in inner hair cells, this voltage is modulated by mechanical motion of the bundle. But the resulting length changes influence the mechanics of the basilar membrane. When the basilar membrane moves down, during one half-cycle of a pure-tone stimulus, the outer hair cells increase their length so as to push downwards on the membrane, thereby increasing its displacement. And

the cochlea is functioning normally, the motion of the basilar membrane near the peak is boosted 1000-fold by forces exerted on it by the organ of Corti. This process makes it possible for us to hear very quiet sounds, and it improves our ability to resolve different tones. Furthermore, damage to it is responsible for 80 per cent of hearing losses. The forces driving cochlear amplification most probably come from one of the sensory cells inside the organ of Corti, the outer hair cell. Like heart cells, outer hair cells change their length in accordance with the voltage across the cell membrane. But outer hair cells are extra special, in that they are much faster than heart cells, operating on a timescale of one-millionth of a second, and they work in both directions, in that they both shorten and lengthen. Furthermore, outer hair cells are extremely sensitive, generating forces in response to displacements of one-millionth of a millimetre.

When developing a model we must decide what simplifications to use to retain as much structural realism as possible whilst ensuring that the model is solvable on present-day computers. In comparison with the heart, the development of structurally realistic finite-element models of cochlear mechanics is in its infancy. Most current models reduce the complex structure of the cochlea to just a handful of independent variables, which is a bit like simulating car crashworthiness using a Duplo model consisting of four wheels and a handful of blocks. My approach is to embed an orthogonal organ of Corti into the cochlear fluids, and to restrict the stimuli to pure tones, which happens to be consistent with most experimental investigations. These simplifications have made it possible to divide the complete cochlea into 0.01 mm pieces. The properties of the individual model structures in the resulting 1 000 000 system equations are based on recent experimental measurements.

The computer model allows use to predict what is going in the real organ of Corti (Figure 9.6). However, most experimental data currently relates only to the motion of the basilar membrane. By comparing the model response under different experimental conditions (Figure 9.7), we can get valuable insight into how the cochlear amplifier operates. The

when the basilar membrane is moving upwards, the hair cells contract. This process, known as cochlear amplification, increases the sensitivity and frequency sensitivity of the auditory system. An accurate understanding of cochlear amplification requires the characterisation of the interactions between the outer hair cells and the other structures of the cochlear partition, whilst taking into account loading by the fluids that surround them.



Figure 9.6. The displacement of the model organ of Corti during sound stimulation at 30kHz, at one instant in time near the position of the response peak using the normal set of parameters. The viewing angle is different from that in Figure 9.4. Here we are looking at the organ from behind the pillar cells. Scale varies in the figure, with only a 1.5-mm length of the model shown in the vicinity of the response peak, and many of the rows of cells have been removed to aid clarity (there are actually 1000 cells present in this region of the model). At different positions the outer hair cells can be seen to be lengthening and contracting, thereby modifying the displacement pattern of the basilar membrane. The bottom of each outer hair cell moves more than the top, indicating that the basilar membrane is moving considerably more than the tectorial membrane. The length of each Deiters' and pillar cell is constant throughout the model, due to their high axial stiffnesses. Looking in detail at animations of motion within the organ of Corti from all possible viewpoints gives us a deeper understanding of the operation of the cochlear amplifier.

behaviour of the cochlear model suggests that behaviour at organ level is impossible to predict from that of individual cells in isolation. This reinforces the view that finite-element models can provide insights into the operation of biological organs that are impossible to obtain any other way.

9.8 The next 10 years

We are at the threshold of a new era in biological research. Finite-element computer models are transforming our understanding of complete organs. Some organs, such as the cochlea, are already being modelled at a cellular level. Other organs, such as the heart, are represented by models that are more structurally accurate, and they incorporate interactions between different forms of energy. These different strategies for balancing structural realism against spatial resolution will continue to be driven by the processing power available from computers.

The maximum size and complexity of a finite-element model is

limited mainly by acceptable analysis times. The speed of inexpensive commodity microprocessors has increased exponentially since their introduction three decades ago, doubling every 18 months. If this were to continue, by 2010 they would be 100 times more powerful than today's, and they would be cheaper in real terms. Unfortunately, physical limitations to both transistor density and switching speed will almost certainly limit increases in the power of individual microprocessors.

An alternative is to look to the Internet, whose growth is sure to continue unabated, driven by factors as diverse as minimising the drudgery of grocery shopping to the widespread adoption of working from home, as people strive to avoid the damaging social and environmental effects associated with commuting. This leads naturally to the concept of distributed parallel-processing techniques, which divide the task of analysing the finite-element model between several processors that are housed in separate computers in different locations. By utilising commodity computers we benefit from the economies of mass production that are associated with sales of tens of millions of units annually.

Distributed parallel processing also provides the potential to utilize a wasted resource. Many people have a computer in their office or in their home that spends more than 99 per cent of its time doing little more than providing low levels of background heating and noise. It makes sense to give them something to do when they are not being used as expensive typewriters or handheld calculators. The utilisation of only 50 commodity computers would, with virtually no capital investment, provide a distributed parallel application with the processing performance that a single-processor computer will not be able to match within the next 10 years. And, of course, as individuals computers are upgraded the distributed application will have immediate access to the increased power.

9.9 The year 2020

It is difficult to predict developments beyond the first decade of the twenty-first century. However, there are two arenas in which modelling and biology may converge even further, namely developmental biology and carbon-based computing. Developmental biology is an area of experimental research that is expanding rapidly. Current tissue-based work on cochlear regeneration highlights the difficulties of artificially controlling the

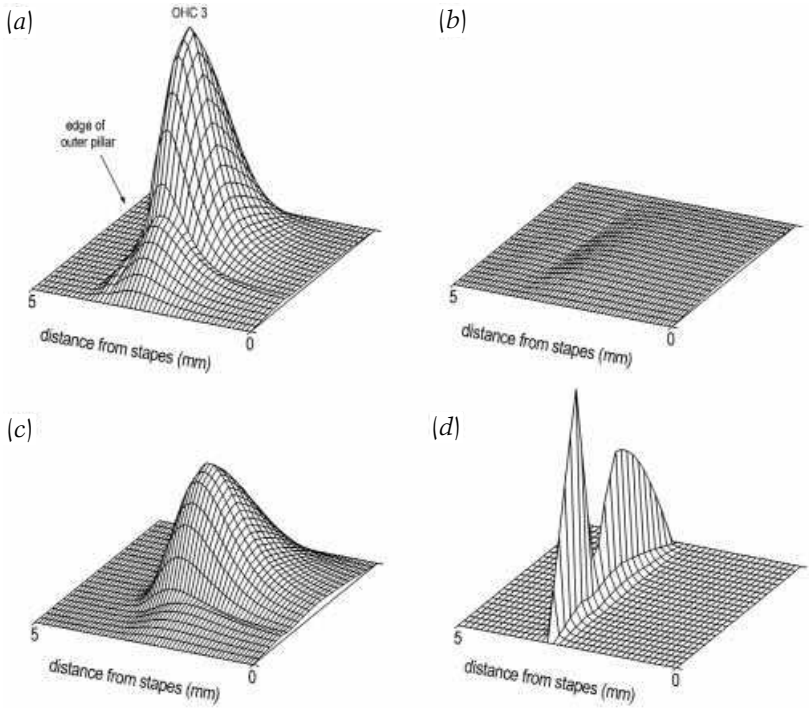


Figure 9.7. The amplitude of the basilar membrane displacement under four different experimental conditions, for pure-tone stimulation. Only the first 5 mm of the model is shown, since no significant motion occurred elsewhere. For these views, all the components of the organ of Corti plus the tectorial membrane have been peeled away. The viewing angle is similar to that in Figure 9.6, but the viewing point is now more elevated. The vertical scale, which is the same in all four panels, is greatly exaggerated (as in Figure 9.4(b)). (a) Simulating the response of a normally functioning cochlea, by using the normal amount of force generation by the outer hair cells. The motion of the basilar membrane has a peak that is localised along the length of the cochlea. At this position, the maximum displacement occurs beneath the outer hair cells, and there is a slight inflexion point at the outer edge of the outer pillar cell. (b) Simulating the response of a severely damaged or dead cochlea, which is achieved by switching off outer hair cell force generation. The maximum displacement of the basilar membrane is now much lower, indicating a profound loss of sensitivity. The change shown is consistent with experimental data. A person with no outer hair cell force generation would suffer a profound hearing loss. (c) Simulating the response when the outer hair cells are stimulated by electrical signals direct from the brain. Experiments on isolated cells have shown that such stimulation increases the

development of structurally complex biological systems. Rather than attempting to replicate the existing organ of Corti, we could use finite-element models to predict the degree of mechanical amplification that could occur in regenerated hair-cell-based sensory epithelia, whose structure and properties are quite different from those of the normal organ of Corti. Biological, carbon-based implementations of the simplified organ could be constructed using genetic techniques, both by manipulating the function of individual cells and controlling the way in which the developing cells form the structure. The development process itself is amenable to finite-element analysis since it is driven mainly by local effects. The replacement organ could be constructed from cells obtained from the eventual organ recipient, bypassing the problems associated with tissue rejection during transplantation. Conversely, silicon-based implementations of the simplified model could be used in signal processing applications. For example, a silicon cochlea could form the front-end of a speech recognition system with a performance superior to any designed by an electrical engineer.

It is highly likely that by the second decade of the new millennium silicon-based computing will have reached fundamental technological or physical limits. Computers will therefore be based on substrates that exhibit superior performance characteristics. One possibility is the photon. Optoelectronic devices, which use substrates such as gallium arsenide, permit the interconversion of electrons and photons. Hybrid computers, which may already be available commercially by 2010, would use silicon for computation and photons for data transfer. The coherent modulation of very-high-frequency light beams enables many high-capacity

amount of force generation. But in the model, increased force generation leads to less basilar membrane motion. This paradoxical observation is the first that is consistent with the experimental observations that an increased amount of brain stimulation causes a decrease in cochlear amplification. The model behaviour is the direct result of the inflexion point at the outer edge of the outer pillar cell becoming much more pronounced. (d) Simulating the response of the cochlea when individual outer hair cells are stimulated in the absence of sound. There are two motion peaks at the position of stimulation, one beneath the outer hair cells and the other at the outer edge of the outer pillar cell. This model response is consistent with experiments in which the cochlea is electrically stimulated, and comparison with Figure 9.7(a) shows that the response to normal sound cannot be predicted from these sorts of experiments.

signals to be combined onto a single beam, taking up little space and not interfering with cooling air. In, say, 20 years a fully optical computer would integrate lasers with optical modulators and photodetectors, and could be 1000 times faster than today's computers.

9.10 The year 2050

Within 50 years we may see the ultimate combination of biology and modelling, with finite-element models being implemented on carbon-based computing platforms. Carbon shares silicon's electron valency, making it a viable semiconductor. But carbon's real potential lies in its unrivalled ability to form compounds of very high molecular weight, which has made it suitable for the encoding and processing of the huge amount of information required to construct a human being. It is a logical step to consider utilizing DNA code and associated enzymes, which have been developed and refined over billions of years of evolution, to construct a carbon-based computer. Such a device could exist in a test-tube, into which DNA-like molecules would be placed containing the input data, and recombinant DNA techniques used to perform the processing function. The output would be the resulting new 'genetic' combinations. A carbon-based computer would have several attractive characteristics:

- **Fast:** trillions of strands of DNA would be processed in a single biochemical operation, so that a computation that would currently take one year to perform could be completed in one second.
- **Compact:** if a grain of sand represented each unit of information on DNA, the information contained in the sand on all the beaches on Earth would fit under a fingernail.
- **Efficient:** a computation that would currently consume all the output from a large nuclear power station could be performed using the output of a single 1 cm² solar cell.

Taken together, these performance levels represent a million-fold improvement over present-day computers. This means that the current rate of exponential growth in computing power will be sustained for another half century if carbon-based computers were to become commodity items by 2050. It may then be feasible to implement a finite-element model of a complete human at a cellular level.

9.11 Further reading

Encyclopedia Britannica (www.britannica.com) contains a wealth of information, both specialist and general, on many of the topics discussed here.

The *Scientific American* website (www.sciam.com) has an interesting 'ask the experts' section, including further information on DNA computing.

The potential power of distributed computing is well demonstrated at the website www.distributed.net

For more information about hearing, see *Springer handbook of auditory research: the cochlea* 1996 (eds. P. Dallos, A. N. Pooper & R. R. Fay). New York: Springer.

The Intel website (www.intel.com) contains an article ([/pressroom/archive/backgrnd/cn71898a.htm](http://pressroom/archive/backgrnd/cn71898a.htm)) that describes the full history of the commodity microprocessor since its launch in 1968.

For a more technical presentation of all the topics discussed here, please refer to: Kolston, P. J. 2000 Finite-element modelling: a new tool for the biologist. *Phil. Trans. R. Soc. Lond. A* **358**, 611–631.



10

Reverse engineering the human mind

Vincent Walsh

*Dept of Experimental Psychology, University of Oxford, South Parks Road,
Oxford OX1 3UD, UK*

Reverse engineering: *‘The process of analysing an existing system to identify its components and their interrelationships and create representations of the system in another form or at a higher level of abstraction’.*

On-line dictionary of computing.

In a letter to the Danish physicist Hans Oersted (1777–1851) in 1850 Michael Faraday remarked that, concerning scientific discoveries, ‘we have little idea at present of the importance they may have ten or twenty years hence’. It is of course a view with which no research scientist will disagree but even Faraday may have been surprised at the life span and biography of one his most widely known discoveries. In 1831 Faraday demonstrated that a moving magnetic field could induce an electric current in a nearby circuit, a discovery he believed ‘may probably have great influence in some of the most important effects of electric currents’. At the time of this discovery it was already known from Luigi Galvani’s (1737–1798) experiments, showing that electrical currents could produce muscle contractions, that nervous tissue had *something* to do with electricity; and in 1838 Carlo Matteucci (1811–1868) had introduced the term ‘muscle current’ to describe the activity of muscle tissue previously referred to as ‘animal electricity’. Ten years later Emil Du Bois-Reymond (1818–1896) demonstrated a direct relationship between electric current and nerve cell activity. Even so, it took until 1939 and the work of Alan Hodgkin and Andrew Huxley to show that brain activity *depends* upon electrical activity: the brain, then, is a machine that runs on electricity.

These discoveries ushered in the first wave of stimulation studies as a means of reverse engineering brain function. Physiologists began to apply electrical stimulation to the cerebral cortex (the outer surface of the brain), and in doing so were able to produce movements in muscles on the contralateral side of the body (the movements of one side of your body are controlled by the opposite side of your brain, so magnetic or electrical stimulation of, say, the left half of your brain will cause movements on the right side of your body). Working on dogs and monkeys, David Ferrier used magnetically induced currents to produce a map of cortical function (Figure 10.1) and the technique of direct stimulation to map function was later extended to human subjects. Progress in brain stimulation was rapid and reached its first peak when Wilder Penfield and his colleagues applied electrical stimulation to the cortex of patients undergoing neurosurgery and were able to work out the way in which body movements were represented in the brain (Figure 10.2). They also confirmed the location of speech reception and production areas, identified a third speech-related area and stimulated areas that produced specifically tactile or visual sensations. One patient, identified as Case J.V. (patients are usually referred to by their initials for confidentiality), experienced seeing familiar people and familiar scenes when stimulated in the temporal lobe.

There were several limitations to these methods of investigating brain function. The invasive nature of the experiments meant that they could only be carried out in patients who were awaiting surgery and of course this restricts the kinds of experiments one can do. Another important limit was the specificity of the movements or perceptions produced. The motor cortex is required for fine control and important skills such as giving complex hand signals to other road-users, but Penfield's stimulation elicited actions which were 'not more complicated than those a newborn infant is able to perform'. Some brain regions, however, which Penfield and Rasmussen referred to as 'elaboration areas' apparently did not respond to electrical stimulation because the brain does not only produce perceptual and motor outputs but also transforms them: it would be difficult to imagine how stimulation would elicit awareness of a transformation. For example, at some stage in reading, your brain is able to translate printed letters into sounds but stimulation never caused a subject to report anything like this. Reading probably seems so automatic that you may even have difficulty imagining that a written word is translated into a sound. The closest you might get is to read something like 'the door slammed

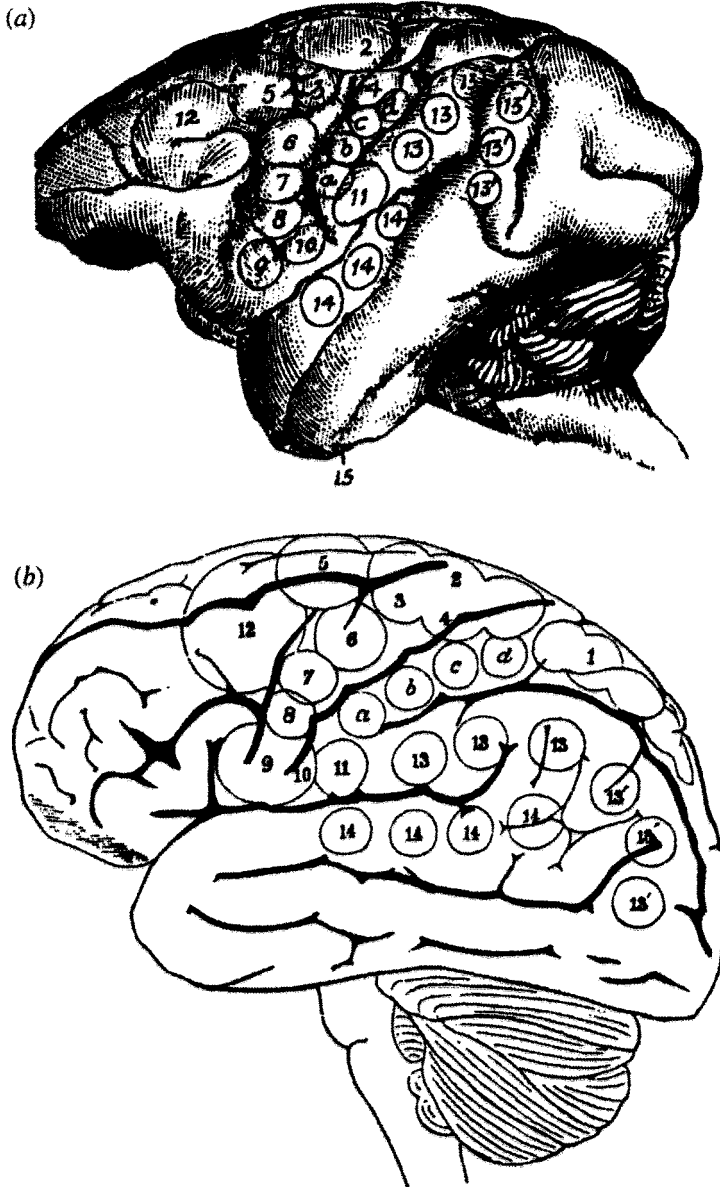


Figure 10.1. Ferrier (1876) mapped the different functions of the macaque brain (a) by direct stimulation of the cortex and transposed the functions to the human cortex (b).

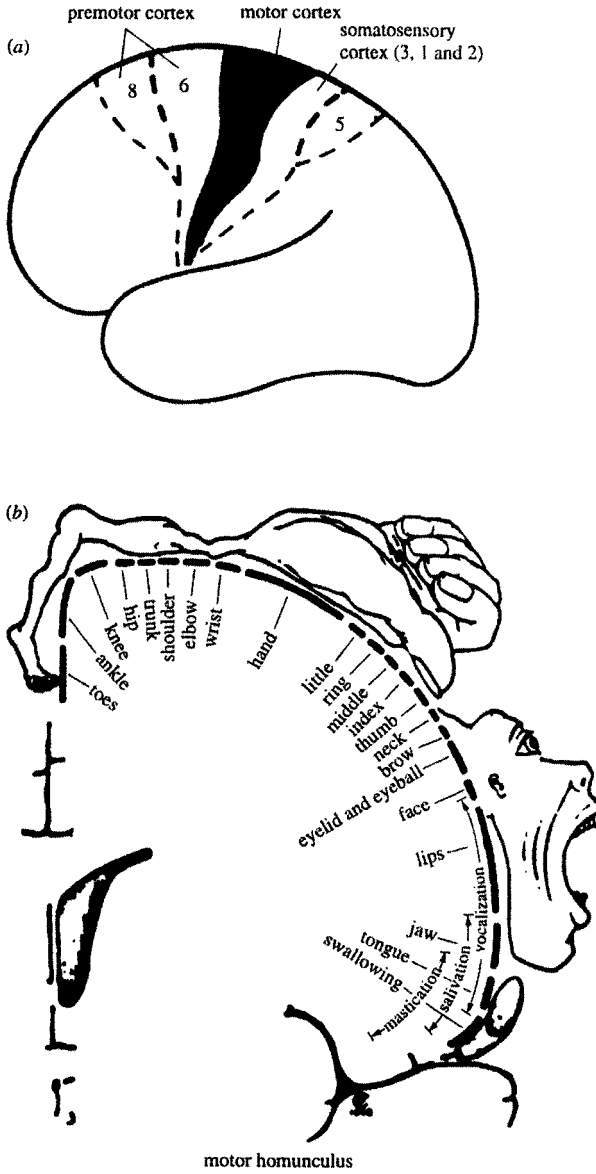


Figure 10.2. The motor homunculus produced by Penfield and Rasmussen from direct stimulation studies. Note that the body is distorted and those areas which produce fine motor actions and manipulations (the hand and the mouth) are disproportionately represented.

shut' – I know you heard the door, but that was the image not the translation itself. Penfield and Rasmussen were aware of this problem and concluded that in these cases stimulation 'sheds no light upon the function of an area unless the patient is making use of that area at the moment'. What is needed, then, is some way of reverse engineering the brain *in action* – a means of catching the brain in the act.

Another wave of reverse engineering, neuropsychology, began soon after the first and got into full flight with the report by Pierre Paul Broca (1824–1888) that damage to a part of the lower left frontal lobe rendered patients unable to produce speech. The approach taken in neuropsychology is to investigate the abilities of patients who have suffered brain damage and from the pattern of their deficits to infer something about the function of that region or about the general organisation of the system under investigation. The study of patients with focal brain damage formed perhaps the most important source of knowledge about the organisation and function of the brain for the best part of the twentieth century and the kinds of dissociations demonstrated were both informative and intellectually seductive. For example, one patient, called L.M., has been reported to be severely impaired in the perception of movement but less so in the perception of shapes and textures and not at all in the perception of colours. Another patient perceived the world totally devoid of colour without suffering any marked reductions in movement and form perception. Other such functional dissociations abound: patient D.F. has very poor visual perception of orientation but can nevertheless *use* information about orientation to grasp objects or put objects through differently oriented holes. Other specific and curious deficits include the loss of awareness of one half of the body, or of objects, or an inability to name an object presented to the right hand side of the brain when it is disconnected from the left side. All of these examples suggest that the brain is organised into groups of relatively specialised areas.

In many respects the classic findings of neuropsychology have formed the bedrock of much of what we know about how we see, hear, speak, move and even feel. Nonetheless, neuropsychology has not always influenced theories about how the intact brain carries out tasks. This is partly because nature is a poor surgeon: accidental brain damage is usually spatially diffuse, interrupts several functions, is irreversible and the time of its occurrence cannot be predicted. Another problem with the lesion method in general, even when specific areas can be removed from animals, is that

it doesn't possess any degree of temporal resolution. Temporal resolution simply refers to the window of time which can be used to look at a function and it is critical when one considers the nature of psychological models of brain function. Our models always contain stages of processing that are part parallel and part serial. In other words, to understand brain processes means understanding them in time as well as space. Knowledge of precisely *when* the brain carries out specific functions is fundamental to any accurate description of how the brain performs many complex tasks. And it's not just a matter of running a clock against brain functions. Indeed the brain may invent some aspects of what you think of as real time. You might think you experience a unified world in which objects have shape and colour and movement – but you are deluded. The brain areas that deal with the different attributes of an object all operate at different paces, perhaps several milliseconds apart (several milliseconds is a long time in the brain – while you're larding about the brain is doing some impressive housekeeping) and we don't know how they are brought together in synchrony.

The stimulation method could not address the role of the elaboration areas and the study of brain damaged patients or lesion studies of animals is hampered by the lack of temporal resolution. What is needed for another wave of reverse engineering, then, is the ability to stimulate the brain while it is doing something, or to be able to reversibly disrupt its functioning to give the lesion method a temporal dimension. The story of how we are able to achieve both of these takes us back to Faraday. . . .

Recall that Faraday discovered electromagnetic induction and we know the brain is a conductor of electricity. It follows that exposing the brain to a changing magnetic field will result in an induced electrical field and therefore neural activity in the brain. This was soon appreciated and as the nineteenth century drew to its close Arsene d'Arsonval (1896) reported the first production of visual percepts (spots or flashes of light called phosphenes) induced by magnetic stimulation (Figure 10.3). The subject also reported feelings of vertigo and under some conditions muscle contractions as well.

One might have thought that d'Arsonval's discovery would be sufficient to generate further studies of brain function by magnetic stimulation, but the technical solutions to this had to wait for the best part of the twentieth century until 1985 when Anthony Barker and colleagues at the University of Sheffield successfully stimulated the motor cortex and pro-

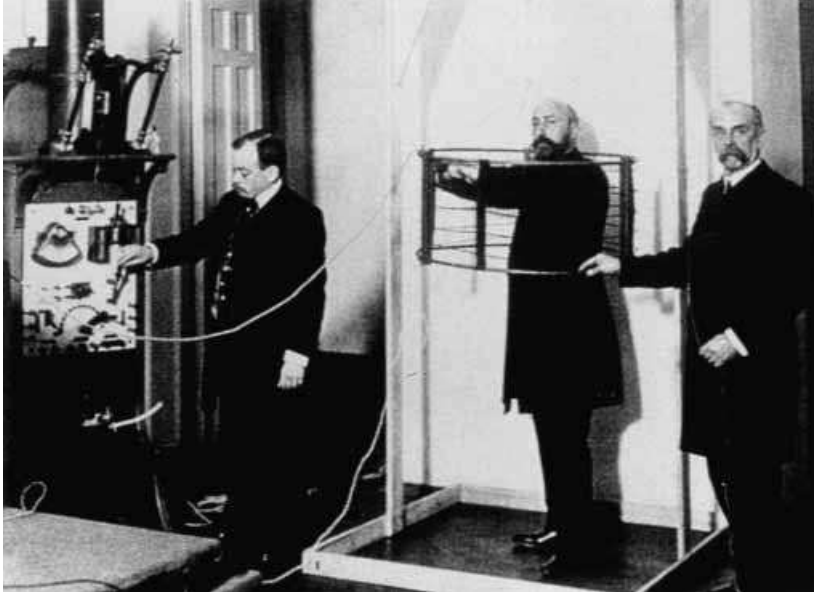


Figure 10.3. d'Arsonval and colleagues showing the apparatus used to induce visual phosphenes. This photo was taken in 1911.

duced movements of the hands without causing the subjects any discomfort. The magnetic pulse was generated by current (up to 8000A) flowing through a small coil held above the subject's head. The current is discharged over a period of 1 ms, reaching its peak in as little as $200\ \mu\text{s}$ and this produces an intense magnetic pulse (approx. 2T) which in turn induces current flow in the underlying cortical tissue. The technique is painless and safe as long as ethical and safety guidelines are followed.

The clinical neuroscience community was quick to pick up on the importance of this discovery and Barker's Transcranial Magnetic Stimulation (TMS) was soon widely used to measure nerve conduction velocities in clinical and surgical settings. However, it is not in the clinical domain that magnetic stimulation provides the most excitement; magnetic stimulation is a tool with which to discover new facts about brain function and it has already delivered in many areas.

I noted above that two of the problems with the lesion technique in patients and non-human primates were that the process could not be reversed and information about time was lost. With magnetic stimulation,

however, one can apply a single pulse (which lasts for less than 1 ms) at any time while a subject performs a task. The effect of the TMS is to cause neurons to discharge at random in and around the area stimulated and thus to impede the normal functioning of that area. Thus the subject 'suffers' from a temporary 'lesion effect' which lasts for a few tens of milliseconds. Theoretically we are now able to disrupt information transmission in specific circuits at specific moments in time in the same way as a debugger needs to be able to access parts of a computer program at a particular point in its execution: a reverse engineer's dream. This has become known as the creation of 'Virtual Patients' and takes us into the realms that Penfield and Rasmussen could not enter – those elaboration areas. But the first challenge for magnetic stimulation is to show that it can recreate the effects seen in real brain damaged patients.

The patient L.M., mentioned above, suffered brain damage, caused by a thrombosis, which affected those regions of her brain, known as the V5 complex, that are important for the perception of movement. According to the rationale of the virtual patient approach, magnetic stimulation applied to the visual motion areas of the brain should make subjects experience the same difficulties as L.M. Indeed several laboratories have now shown that magnetic stimulation over human area V5 specifically impairs the perception of movement. So magnetic stimulation has the face validity conferred by replication of others' findings (an important step in science) but it needs also to be able to extend the findings of others.

In their investigations of patients, Penfield and Rasmussen observed that stimulation of the brain regions responsible for seeing led patients to experience phosphenes which they described in terms such as 'I saw just one star', 'Silver things to the left of me' or 'red and blue wheels'. Penfield and Rasmussen were aware that seizures of the occipital lobe were associated with blindness in the parts of the visual field represented therein and they surmised that with their more localised electrical stimulation the patient 'may be blind only in that portion of the field where he seems to see the light'. This kind of focal blindness is known as a scotoma and magnetic stimulation has since been able to show that the prediction was correct. Thomas Kammer, working at the Max Planck Institute in Teubingen, applied magnetic stimulation to the visual cortex and gave subjects a task in which they were required to detect the presence of a target in different parts of the visual field. He found that the location of the transient scotoma coincided with the location of the phosphene produced by

magnetic stimulation at the same site. Thus, in the case of the visual cortex we now know that the mechanism of suppression appears to be excitatory. This is an important step forward because the production of a deficit does not of itself say anything about the neural mechanism.

Penfield called areas from which they could not elicit a response 'elaboration areas' and surmised that these could only be studied in action. In a recent series of experiments in Oxford, Matthew Rushworth has not only shown this to be true but has demonstrated the temporal structure of interactions between the motor cortex (which Penfield and Rasmussen could study) and the premotor cortex (an elaboration area which could not be studied by direct stimulation). Subjects were required to carry out a simple visual discrimination task (discriminating between large and small rectangles and circles) and to press an appropriate button. Magnetic stimulation was applied to one of three cortical areas at different times after the stimuli were presented. If TMS was applied to the motor cortex around 300ms after the stimuli were presented, subjects were slower to make their responses; if magnetic stimulation was applied to the pre-motor cortex around 100ms after stimulus onset the subjects were slower to make their response; and if an area between these two sites was stimulated, the time to respond was slower when the TMS arrived around 180ms after the visual stimuli were presented. Here we have an example of three links in a chain of motor signals being segregated by magnetic stimulation across a gap less than one fifth of a second. This millisecond-level power shows that the pre-motor elaboration area is important for selecting which movements to make over 100ms before the lower level motor cortex is instructed to execute the movement.

Correlating excitation with temporary blindness, recreating the effects of brain damage and elaborating the fine temporal structure of the interactions between different areas within a system all seem to be reasons for brain engineers to be cheerful. But the brain cheats. Like no other machine it changes the way it performs a task over time. One may have given a detailed and even accurate account of the function of an area, but the details of the function can change: an area which is crucial to learning a task may not be necessary once the task has been learned and even if it is, its role may have changed. Studies using magnetic stimulation have approached the issue of plasticity by either measuring the functional correlates of it or by actually manipulating it. A particularly pleasing example of charting the changing functions of the nervous system is the work of

Janet Eyre in Newcastle who stimulated the motor cortex in over 300 subjects between the ages of 32 weeks and 52 years while recording electrical activity in the biceps and the hand muscles. Eyre took notice of the time between applying stimulation and the arrival of signals at the muscle recording sites (a measure of the speed of nerve conduction) and also of the magnetic stimulation power required to produce muscle activity. There was a sharp decrease in both delay time and power required during the first two years of life and by the time the children had reached five years of age their delay time had reached the same level as that of adults. The importance of this is that the results correlate with the time taken for the muscle nerve fibres involved to reach their maximum diameter and, because diameter is a determinant of speed, their maximum conduction velocities. The magnetic stimulation data also correlate with the time at which children develop good fine finger and prehension skills.

Recording change is impressive enough but change can also be produced. A recent study by Alvaro Pascual-Leone at the Beth Israel Hospital in Boston, MA, has shown that TMS applied at different temporal rates can either impede or enhance one's ability to learn certain kinds of tasks. Remarkably low levels of stimulation (1 pulse per second) over the motor cortex slowed down learning on a visuomotor association task but learning on the same task was faster than normal when magnetic stimulation was applied at 10 pulses per second. Similar results have also been obtained in the visual system and also in studies of language. The implications of this kind of manipulation of learning function are far reaching and attempts to apply this in the clinic are already underway: can we speed up learning? Can we kick start the brain?

What will happen in the twenty-first century? As you no doubt remember from all the 'end of century' pundits who soiled magazines and newspapers as we entered the year 2000, prediction is no more than a veil for predilection, so I'll come clean and say what it is I would like to see happen in the near future with magnetic stimulation. The emergence of magnetic stimulation as a tool in neuropsychology has been slower than it should have been. Other techniques, such as functional magnetic resonance imaging, multi channel electroencephalography and magnetoencephalography have all attracted more attention. They are in themselves exciting developments and we have learned much about the human brain from them. However, they all record brain activity in one form or another and thus cannot reveal how the brain would function in the absence of a

certain component. Magnetic stimulation offers a unique combination of component removal and timing and for these reasons has a special role in addressing psychological problems. So prediction 1 is that every Psychology Department in the world will have a magnetic stimulation lab. My second prediction concerns the ability of magnetic stimulation to influence cortical activity. Already we are seeing signs it may be able to replace electroconvulsive therapy in the treatment of depression and one can only hope for an acceleration in the development of this program. In addition there is potential for magnetic stimulation to be used to influence the progress of those recovering from stroke, if the ability to influence learning turns out to have real potential. In these kinds of cases magnetic stimulation plays the role of treatment and tester because one can chart progress by interfering with functions as well trying to enhance them. My final prediction is that magnetic stimulation will be used in conjunction with the other imaging techniques to obtain a picture of the brain in action when it has been used to either impede or enhance processing. Indeed there has already been some success in this area. Using PET (Positron Emission Tomography) scanning Tomas Paus in Montreal measured changes in cerebral blood flow after subjects had received magnetic stimulation. The pattern of brain activation was not random: the areas activated by magnetic pulses included the site beneath the stimulating coil and several regions to which that area was anatomically connected. From hereon magnetic stimulation will be used to assess which of those activations have a functional meaning by applying it and recording brain blood flow when subjects are performing a task. It may even lead to crossing one of the longest bridges in cognitive neuroscience: how do the functionally specialised regions of the brain act together to produce our experience of the world? The upshot of all this will be what science always aims for – counterintuitive insights into a piece of the natural world.

Whatever happens there is only one route scientists can take and for a reminder we can go back to Faraday. In 1859, while trying to devise a means of measuring gravitational forces, he wrote in his diary, 'Let the imagination go, guiding it by judgement and principle, but holding it in and directing it by *experiment*' – good advice for the next millennium of science.

10.1 Further reading

- Cantor, G. 1991 *Michael Faraday: Sandemanian and scientist*. Macmillan Press.
- Collins, P. 2000 Field Workers. *New Scientist* **165** (2224), 36–39.
- Rothwell, J. C. 1993 Evoked potentials, magnetic stimulation studies and event-related potentials. *Curr. Opin. Neurol.* **6**, 715–723.
- Shallice, T. 1988 *From neuropsychology to mental structure*. Cambridge University Press.
- Walsh, V. 2000 Reverse engineering the human brain. *Phil. Trans. R. Soc. Lond. A* **358**, 497–511
- Walsh, V. & Cowey, A. 1998 Magnetic stimulation studies of visual cognition. *Trends Cognitive Sci.* **2**, 103–109.

Contributor biographies



Michael Thompson was born in Cottingham, Yorkshire, on 7 June 1937, studied at Cambridge, where he graduated with first class honours in Mechanical Sciences in 1958 and obtained his PhD in 1962 and his ScD in 1977. He was a Fulbright researcher in aeronautics at Stanford University and joined University College London (UCL) in 1964. He has published four books on instabilities, bifurcations, catastrophe theory and chaos and

was appointed professor at UCL in 1977. Michael was elected a fellow of the Royal Society in 1985 and was awarded the Ewing Medal of the Institution of Civil Engineers. He was a senior SERC fellow and served on the IMA Council. In 1991 he was appointed director of the Centre for Nonlinear Dynamics at UCL. He is currently editor of the Royal Society's *Philosophical Transactions* (Series A) which is the world's longest running scientific journal. His scientific interests include nonlinear dynamics and their applications. His recreations include walking, tennis and astronomy with his grandson Ben, shown above.

G. Roberts



Gareth Roberts' research interests are centred on the quantum dynamics of ultrafast laser–molecule interactions and molecular collisions. He was brought up in South Wales and holds degrees from the Universities of London and Cambridge. His interest in ultrafast phenomena was triggered in 1989 by an inspirational stay as a NATO Postdoctoral Fellow in the laboratory of Professor A. H. Zewail at the California Institute of Technology. He is 36 years old and is currently a Royal Society University Research Fellow at Cambridge University and a Fellow of Churchill College, Cambridge.

M. J. Sutcliffe



Born in Rochdale, Lancashire, Michael Sutcliffe (left) studied at Bristol, where he graduated with first class honours in chemical physics in 1985, and at Birkbeck College, London, where he obtained his PhD in protein modelling in 1988. Aged 35, he was a SERC/NATO Fellow at Oxford University and Junior Research Fellow at Linacre College, a Royal Society University Fellow at Leicester University, and joined Leicester University as a Lecturer in 1998, where he is currently Reader. His research involves the development and use of computational methods to address one of the major challenges in the biomolecular sciences, understanding the relationship between protein structure and function. He has over 60 publications, including specialised reviews, and was elected a Fellow of the Royal Society of Chemistry in 1999. His recreational interests include hill walking, cycling and canoeing.

N. S. Scrutton

Nigel Scrutton (right) was born in Cleckheaton, Yorkshire. He graduated from King's College, London, with first class honours in 1985 and was awarded the Sir William Robson prize. Nigel obtained his PhD at Cambridge as a Benefactors' Scholar. In 1988, he was elected a Research

Fellow at St John's College and was awarded the Humphreys Research prize. At Cambridge, Nigel was a Research Fellow of the Royal Commission for the Exhibition of 1851 and Royal Society University Research Fellow. He was elected a Fellow of the Royal Society of Chemistry in 1997. Aged 36, Nigel is now Professor at Leicester University and Lister Institute Research Fellow. He is a recipient of the Colworth Medal of the Biochemical Society. His scientific interests include mechanistic and quantum enzymology; his recreational interests include Victorian and College philately.

J. M. Goodman



Jonathan Goodman studied chemistry at Cambridge, graduating with a BA in 1986, and with a PhD in organic chemistry in 1990. He then worked at Columbia University, New York, with Professor Clark Still, before returning to Cambridge as a Research Fellow at Clare College. He is now a Royal Society University Research Fellow in the Department of Chemistry, and uses both computational and experimental techniques to study organic chemistry. He is aged 35, and has recently published a book with the Royal Society of Chemistry, *Chemical Applications of Molecular Modelling*, which introduces experimental organic chemists to computational techniques.

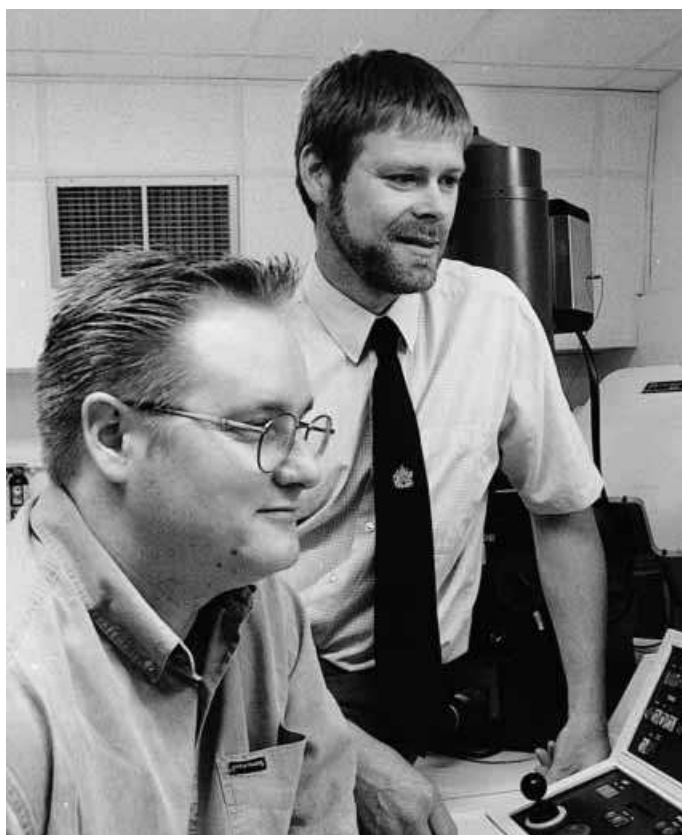
D. J. Macquarrie

Born in Oban, Argyll, in 1960, Duncan Macquarrie studied Pure and Applied Chemistry at the University of Strathclyde, graduating with a first class degree in 1982 and a PhD in 1985. He then moved to York, where he carried out research in Phase Transfer Catalysis. He subsequently spent time in industry, where he worked in the UK and abroad, mostly in synthetic chemistry, but always with an interest in method development and catalysis. He returned to York in 1995 to take up a Royal Society University Research Fellowship, and has developed a range of novel catalysts for green chemistry. He is Associate Editor of *Green Chemistry*, and a National Member of Council with the Royal Society of Chemistry.

Paul W. May



Born in London, Paul May grew up in Redditch, Worcestershire. He went on to study at Bristol University, where he graduated with a first class honours in chemistry in 1985. He then joined GEC Hirst Research Centre in Wembley where he worked on semiconductor processing for three years, before returning to Bristol to study for a PhD in plasma etching of semiconductors. His PhD was awarded in 1991, and he then remained at Bristol to co-found the CVD diamond research group. In 1992 he was awarded a Ramsay Memorial Fellowship to continue the diamond work, and after that a Royal Society University Fellowship. In October 1999 he became a full-time lecturer in the School of Chemistry at Bristol. He is currently 36 years old. His scientific interests include diamond films, plasma chemistry, interstellar space dust, the internet and web technology. His recreational interests include table-tennis, science fiction, and heavy metal music.

A. R. Hemsley

Alan Hemsley studied botany at Bedford College, London (1982–85) and Reading (1986–87) before returning to Royal Holloway, University of London, to study Palaeopalynology for his PhD (1990). Research into heterospory in fossil plants took him to Montpellier, France, and eventually to Cardiff where in 1995 he was awarded a Royal Society University Research Fellowship. Aged 36, Alan has recently co-authored a revision of a popular botany textbook on green plant diversity. His interests still lie principally in the study of fossil spores, particularly their evolution, wall development, structure and chemistry. In the photograph he stands behind his co-author Peter Griffiths.

P. C. Griffiths

Originating from Cornwall, Peter Griffiths studied initially at University College, North Wales (1985–88), and subsequently the University of Bristol (PhD, 1991). After post-doctoral positions in Bristol and Stockholm, he moved to a lectureship at Cardiff in 1995. Aged 32, his research interests centre around colloidal systems, in particular polymer/surfactant interactions.

Marjolein C. H. van der Meulen



Born in 1965 in Utrecht, the Netherlands, Marjolein van der Meulen received her Bachelors' degree in mechanical engineering from the Massachusetts Institute of Technology in 1987. Thereafter, she received her MS (1989) and PhD (1993) from Stanford University. She spent three years as a biomedical engineer at the Rehabilitation R&D Center of the Department of Veterans Affairs in Palo Alto, CA. In 1996, Marjolein joined the faculty of Cornell University as an Assistant Professor in the Sibley School of Mechanical and Aerospace Engineering. She is also an Assistant Scientist at the Hospital for Special Surgery, New York. She received a FIRST Award from the National Institutes of Health in 1995 and a Faculty Early Career Development Award from the National Science Foundation in 1999. Her scientific interests include skeletal mechanobiology and bone structural behavior.

Patrick J. Prendergast



Born in Enniscorthy, Ireland, in 1966, Patrick Prendergast studied at Trinity College Dublin (TCD) where he graduated with a BAI in engineering in 1987 and a PhD in 1991. He was a Council-of-Europe Scholar at the University of Bologna (Istituti Ortopedici Rizzoli) and a Marie Curie Fellow at the University of Nijmegen before being appointed a lecturer in TCD in 1995. He was elected to Fellowship of the College in 1998. He won the European Society of Biomechanics Research Award in 1996. He is on the editorial board of the *Journal of Biomechanics and Clinical Biomechanics*. He was President of the Section of Bioengineering of the Royal Academy of Medicine in Ireland from 1998 to 2000. Scientific interests include computer simulation of tissue differentiation and bone remodelling, and the design of medical devices.

Peter Kohl



Peter Kohl is a Royal Society University Research Fellow and, aged 38, leads the Mechano-Electric Feedback laboratory at the Physiology Department of Oxford University. He studied medicine and biophysics at the Russian Medical University, Moscow, and obtained his PhD at the Berlin Charité. In 1992, Peter joined the Physiology Department at Oxford to continue his studies on the effects of mechanical stimulation on heart rate and rhythm. His work uses a variety of techniques, ranging from experiments on single cells and tissues to analytical models of cardiac mechano-electrical interactions. An unusual facet of his work is devoted to the investigation of the role of connective tissue in the regulation of electrophysiological behaviour of the heart. Peter likes to travel, preferably with his growing family, and enjoys water sports and landscape photography. His favourite – and by far most regular – recreational activity, though, is cooking.

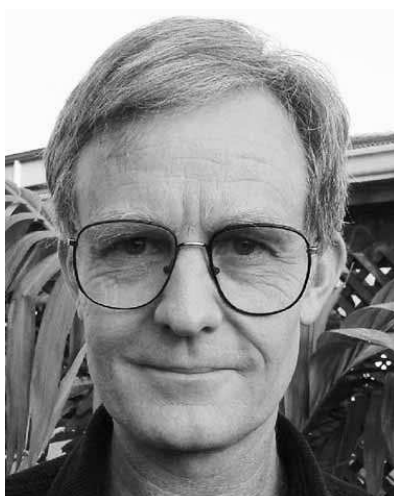
Denis Noble

Denis Noble, 64, is the British Heart Foundation Burdon Sanderson Professor of Cardiovascular Physiology at the University of Oxford and a Fellow of Balliol College. In the early 1960s, he developed the first 'ionic' cell models of cardiac excitation and rhythm generation and has been at the forefront of computational biology ever since. As the Secretary-General of the International Union of Physiological Sciences, he has been pivotal to the initiation of a world-wide effort to describe human physiology by analytical models – the Physiome Project. In 1998 he was honoured by the Queen for his services to Science with a CBE. Denis Noble enjoys playing classical guitar, communicating with people all over the world in their mother-tongue, and converting the preparation of a meal into a gastronomic celebration.

Raimond L. Winslow



Raimond L Winslow, 45, is Associate Professor of Biomedical Engineering, with joint appointment in the Department of Computer Science, at the Johns Hopkins University School of Medicine and Whiting School of Engineering. He is co-Director of the Center for Computational Medicine and Biology, Associate Director of the Whitaker Biomedical Engineering Institute at Johns Hopkins University, and a member of the Institute for Molecular Cardiobiology. His work is aimed at understanding the origins of cardiac arrhythmias through the use of biophysically detailed computer models. These models span levels of analysis ranging from that of individual ion channels, to cells, tissue, and whole heart.

Peter Hunter

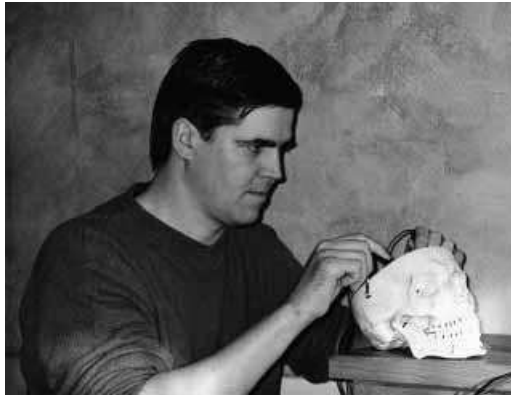
Peter Hunter, 52, is a NZ Royal Society James Cook Fellow and Chair of the Physiome Commission of the International Union of Physiological Sciences. He founded the Biomedical Engineering Group at Auckland University which, in close collaboration with the Auckland Physiology Department, uses a combination of mathematical modelling techniques and experimental measurements to reveal the relationship between the electrical, mechanical and biochemical properties of cardiac muscle cells and the performance of the intact heart. A similar approach is also being used by the Auckland group to analyse gas transport, soft tissue mechanics and blood flow in the lungs with the aim of producing an anatomically detailed, biophysically based coupled heart–lung model for use in drug discovery and the clinical diagnosis and treatment of cardiopulmonary disease.

Paul J. Kolston

Born in Wellington, New Zealand, Paul Kolston studied at Canterbury University (NZ) where he graduated in 1985 with first class honours in Electrical and Electronic Engineering. He obtained his PhD there in 1989, although he spent one year of his PhD studies at the Delft University of Technology, The Netherlands. After a one-year post-doctoral position at

the University Hospital Utrecht, The Netherlands, he moved to Bristol University (UK). In 1995 Paul was awarded a Royal Society University Research Fellowship, which he transferred to Keele University in 1999. Aged 37, he is married with four infant children. Paul's favourite scientific interest is computer modelling of biological systems; his favourite recreational pursuit is body-surfing.

V. Walsh



Born in Oldham, Greater Manchester, Vincent Walsh graduated in Psychology from the University of Sheffield and studied at UMIST for his PhD in Visual Neuroscience. His early interests were in the brain mechanisms for the construction and perception of form and colour. He currently holds a Royal Society University Research Fellowship in the Department of Experimental Psychology, Oxford, where his work is now concentrated on perceptual learning and mechanisms of brain plasticity. His recent initiatives using magnetic stimulation have been the first series of experiments to use the technique to study cognitive processes. Recent additions to the subjects of research in his laboratory include the brain processes involved in mathematics, music and language.

Index

- algorithm, genetic, 56
- animal electricity, 171
- antichaos, 100
- applications of diamond films, 88
- Astronomy and Earth Science*,
 - companion book, preface
- atom–laser interactions, 2–3
- atom–molecule collisions, ultrafast
 - dynamics of, 11–14
- atomic motions, optical control over,
 - 19–20
- attosecond lasers, 18

- bicontinuous mixtures, 98
- bone adaptation
 - computer simulation of, 121
 - experiments, 119
- bone and skin, mechanics of, 157
- bone cells and matrix, 116
- bone growth and maintenance, 118
- brain, 171
 - damage, 175, 178
 - electrical stimulation, 172, 176
 - magnetic stimulation, 176, 179, 180

- cardiac
 - cell models, 135
 - organ models 137, 143
- cars, making of, 155
- catalysis
 - enzyme, 21, 40
 - protein dynamics, 27
- cell models, cardiac, 135

- cells
 - nerve, 171
 - of the bone, 116
- chaos and antichaos, 100
- chemical vapour deposition (CVD), 77
- chemistry
 - green, 59
 - supramolecular, 64
- chess, analogy with organic synthesis,
 - 46
- chiral catalysis, 72
- clusters, ultrafast dynamics of, 15–16
- colloidal particle interaction, 102
- colloidal stability, 101
- computer simulation of bone adaptation,
 - 121
- computing
 - on the Internet, 165
 - using DNA, 169
 - using light beams, 167
- CVD
 - hot filament, 78
 - microwave plasma, 79

- designer materials, 72
- detergents, made-to-measure silicas,
 - 64–68
- diamond
 - films, 87
 - gemstones, 75
 - growth chemistry, 80
 - industrial, 77
- Dirac, Paul, 47

- DNA
 genetic information, 100
 use in computing, 169
- Dr Who, 46
- drugs, designing, 153
- duality, wave–particle, 22, 28
- ear
 models, 160
 structure of inner, 162
- Earth Science*, companion book, preface
- electrical brain stimulation 172, 176
- electricity, animal, 171
- electron transfer, 29
- Electronics*, companion book, preface
- enzyme catalysis, 21, 40
- enzyme mimics, chiral catalysis, 72
- femtosecond, definition, 1
- femtosecond lasers, principles of
 operation, 4–7
- force fields, 49
- fracture healing, 119
- genetic algorithm, 56
- genetic information, DNA, 100
- green chemistry, 59
- growth and maintenance of bone, 118
- growth chemistry of diamond, 80
- healing of fractures, 119
- heart, virtual, 127, 134
- heart models, 159
- homunculus, 174
- hot filament CVD, 78
- human organs, computer models of, 137,
 143, 151
- hydrogen tunnelling
 dynamic barrier models, 34
 experimental evidence, 37
 tunnelling, static barrier models, 32
- industrial diamonds, 77
- inorganic microarchitecture, 98
- integrationisms, 130
- Internet, computing on, 165
- ionization, by femtosecond lasers 3, 10,
 11, 18
- isotope, kinetic effect, 26
- kinetic isotope effect, 26
- laser–atom interactions, 2–3
- lasers
 attosecond, 18
 femtosecond, principles of operation,
 4–7
- light beams, use in computing, 167
- magnetic brain stimulation, 176, 179, 180
- mechanics, bone and skin, 157
- microarchitecture
 inorganic, 98
 organic, 106
 synthetic, 106
- microwave plasma CVD, 79
- Millennium Issues of the *Philosophical
 Transactions*, preface
- modelling
 approach to, 134
 of unknown systems, 128 ff
 purpose of, 133
 utility of, 146 ff
- molecule–atom collisions, ultrafast
 dynamics of, 11–14
- molecules, ultrafast fragmentation of,
 7–11
- nerve cells, 171
- neuropsychology, 175
- neurosurgery, 172
- nonlinear dynamics of bone adaptation,
 122
- nucleation, 86
- optical control over atomic motions,
 19–20
- organ models
 cardiac, 137, 143, 159
 computer of human, 151

- organic microarchitecture, 106
- organic synthesis, 46

- patients, virtual, 178
- Pauling, Linus, 52
- Philosophical Transactions* of the Royal Society, preface
- phosphenes, 176, 178
- Physics and Electronics*, companion book, preface
- Physiome Project, 127, 131
- potential energy curves, 7–8
- protein dynamics, in classical catalysis, 27
- purification of water, 71

- quantum mechanics, 47
- quantum tunnelling, 23

- reductionism, 130
- Royal Society, preface

- silicas
 - detergents as templates, 64
 - made-to-measure, 65
- skeletal adaptation, history of, 113–114
- skeletal form and function, 114

- skeletal imaging, 124
- skeletal structure, 115
- skin, mechanics of, 157
- substrate materials, 84
- supramolecular chemistry, 64
- synthesis, organic, 46
- synthetic microarchitecture, 106

- transition state theory, 22, 25, 26
- tunnelling
 - of hydrogen, 32, 34, 37
 - quantum, 23

- ultrafast dynamics
 - of atom–molecule collisions, 11–14
 - of clusters, 15–16

- virtual heart, 127, 134
- virtual patients, 178
- vitamin K3, clean synthesis of, 70

- water purification, 71
- wave–particle duality, 22, 28
- World Wide Web, 54

- zeolites, as shape selective materials, 60