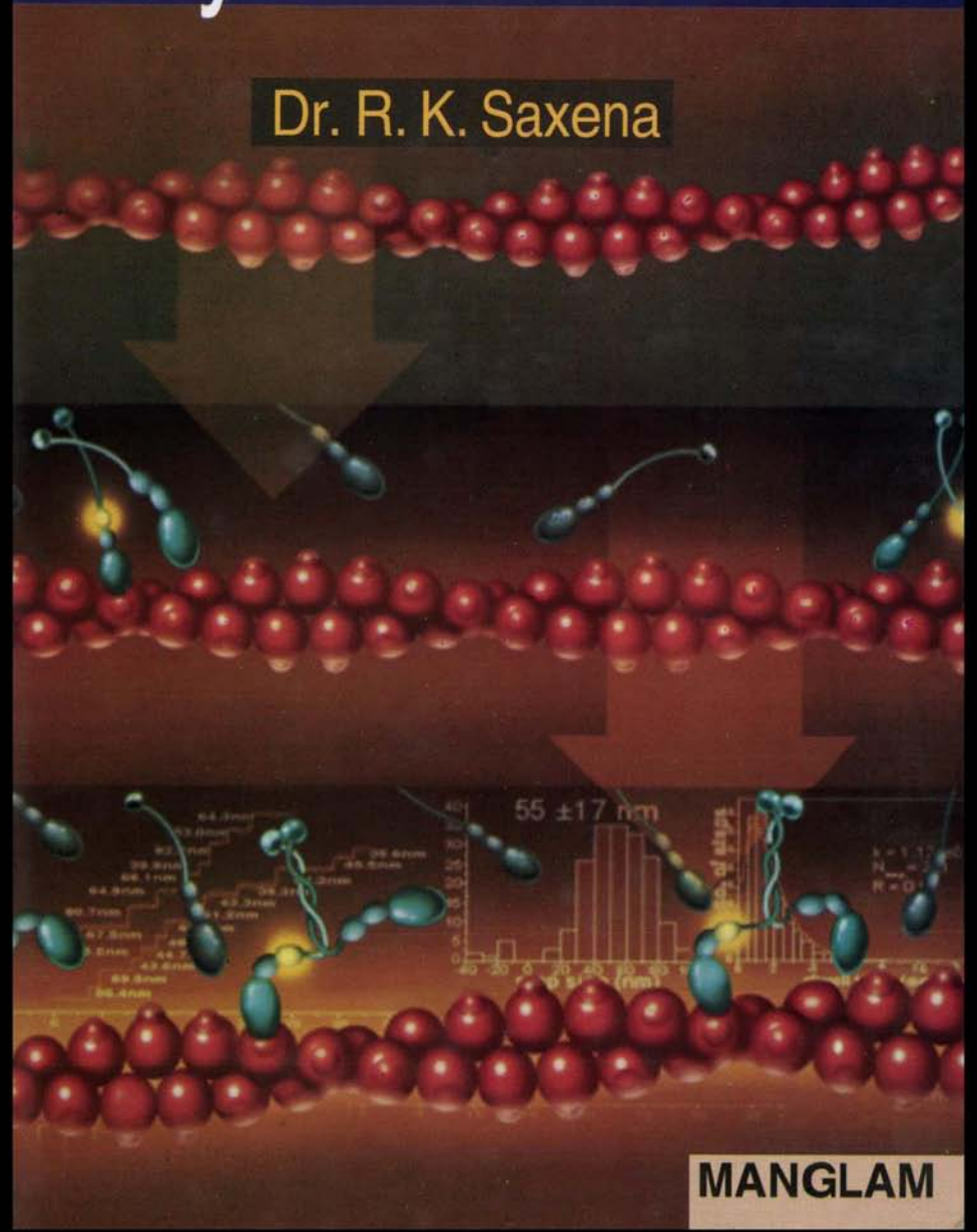
Biology of Physical Science

Dr. R. K. Saxena



MANGLAM

BIOLOGY OF PHYSICAL SCIENCE

"This page is Intentionally Left Blank"

BIOLOGY OF PHYSICAL SCIENCE

Dr. R.K. Saxena

MANGLAM PUBLISHERS & DISTRIBUTORS
DELHI - 110 053 (INDIA)

MANGLAM PUBLISHERS & DISTRIBUTORS
L-21/1, Street No. 5, Shivaji Marg, Near Kali Mandir,
J.P. Nagar, Kartar Nagar, West Ghonda, Delhi -53
Phone: 22945678, Cell.: 9868572512

Biology of Physical Science

© Resrved

First Published 2008

ISBN 978-81-89972-07-3

[No Part of this book may be reproduced in any form by photocopying or by any electronic or mechanical means, including information storage or retrieval systems, without permission in writting from publisher of this book.]

PRINTED IN INDIA

Published by D.P. Yadav for Manglam Publishers & Distributors, Delhi and Printed at Sachin Printers, Maujpur, Delhi-53.

PREFACE

The present title "Biology of Physical Science" is an attempt to provide an uptodate account of the physical processes used in biological sciences. It is basically designed for students, research scholars, teachers, biochemists, microbiologists, immunologists and pharmaco-logists at various universities, institutions and laboratories. This book will enable the user to find his way through the large and often bewildering array of methods currently available. Each chapter has been written in the light of modern development in a simple and lucid style. Sincere efforts have been made to include those topics which have been found to be common among the various university syllabi.

The author expresses his thanks to all those friends, colleagues, and research scholars whose continuous inspirations have initiated him to bring this title.

The author wishes to thank the publisher, printer and staff members for bringing out this book.

Constructive criticisms and suggestions for improvement of the book will be thankfully acknowledged.

Author

"This page is Intentionally Left Blank"

Contents

1

Introduction

There are many apparently different forms of radiation, *e.g.*, visible light, radio waves, infrared, x-rays, and gamma rays. According to the wave model, all of these kinds of radiation may be described as oscillating electric and magnetic fields. Radiation, traveling in the z direction for example, consists of electric and magnetic fields perpendicular to each other and to the z direction. Polarized radiation was selected for simplicity of representation, since all other components of the electric field except those in the x-z plane have been filtered out. The wave travels in the z direction with the velocity of light, $c(3 \times 10^{10}$ cm sec$^{-1})$. The intensity of the radiation is proportional to the amplitude of the wave given by the projection on the x- and y-axes. At any given time, the wave has different electric and magnetic field strengths at different points along the z-axis. The wavelength, λ, of the radiation, and the variation in the magnitude of this quantity accounts for the apparently different forms of radiation listed above. If the radiation consists of only one wavelength, it is said to be monochromatic. Polychromatic radiation can be separated into essentially monochromatic beams. For visible, UV, or IR radiation, prisms or gratings are employed for this purpose.

Radiation consists of energy packets called *photons*, which travel with the velocity of light. The different forms of radiation have different energies. In our discussion of rotational, vibrational, and electronic spectroscopy, our concern will be with the interaction of the electric field component of radiation with the molecular system. This interaction results in the absorption of radiation by the molecule.

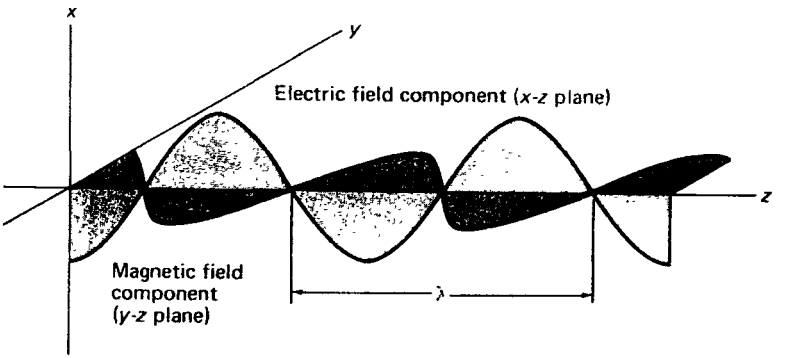


Figure 1.1 Electric and magnetic field components of plane polarized electromagnetic radiation.

In epr and nmr the concern is with the interaction with the magnetic component of radiation. In order for absorption to occur, the energy of the radiation must match the energy difference between the quantized energy levels that correspond to different states of the molecule. If the energy difference between two of these states is represented by ΔE, the wavelength of the radiation, λ, necessary for matching is given by the equation:

$$\Delta E = hc/\lambda \text{ or } \lambda = hc/\Delta E \qquad ...(1)$$

where h is Planck's constant, 6.623×10^{-27} erg sec molecule^{-1}, and c is the speed of light in cm sec^{-1}, giving ΔE in units of erg molecule^{-1}. Equation (1) relates the wave and corpuscular models for radiation. Absorption of one quantum of energy, hc/λ, will raise one molecule to the higher energy state.

As indicated by equation (1), the different forms of electromagnetic radiation (*i.e.*, different λ) differ in energy. By considering the energies corresponding to various kinds of radiation and comparing these with the energies corresponding to the different changes in state that a molecule can undergo, an appreciation can be obtained for the different kinds of spectroscopic methods.

1.1 ENERGIES CORRESPONDING TO VARIOUS KINDS OF RADIATION

Radiation can be characterized by its wavelength, λ, its wave number, $\bar{\nu}$, or its frequency, ν. The relationship between these quantities is given by equations (2a) and (2b):

$$\nu(\text{sec}^{-1}) = \frac{c(\text{cm sec}^{-1})}{\lambda(\text{cm})} \qquad ...(2a)$$

$$\bar{\nu}(\text{cm}^{-1}) = \frac{1}{\lambda(\text{cm})} \qquad \ldots(2b)$$

The quantity $\bar{\nu}$ has units of reciprocal centimeters, for which the official IUPAC nomenclature is a Kayser; 1000 cm^{-1} are equal to a kiloKayser (kK). From equations (1) and (2), the relationship of energy to frequency, wave number, and wavelength is:

$$\Delta E(\text{ergs molecule}^{-1}) = h\nu = hc/\lambda = hc\bar{\nu} \qquad \ldots(3)$$

In describing an absorption band, one commonly finds several different units being employed by different authors. Wave numbers, $\bar{\nu}$, which are most commonly employed, have units of cm^{-1} and are defined by equation (2b). Various units are employed for λ. These are related as follows: 1 cm = 10^8 Å (Ångstroms) = 10^7 nanometers = 10^4 μ (microns) = 10^7 mμ (millimicrons). The relationship to various common energy units is given by: 1 cm^{-1} = 2.858 cal/mole of particles = 1.986 × 10^{-16} erg/molecule = 1.24 × 10^{-4} eV / mole. These conversion units can be employed to relate energy and wavelength; or the equation

$$\Delta E(\text{kcal mole}^{-1}) \times (\text{Å}) = 2.858 \times 10^5 \qquad \ldots(4)$$

can be derived to simplify the calculation of energy from wavelength.

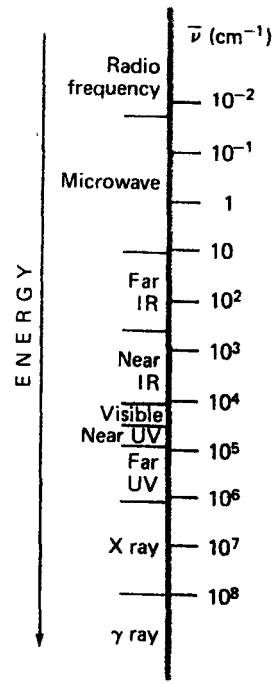


Figure 1.2 Wave numbers of various types of radiation.

Wave numbers corresponding to various types of radiation are indicated in Figure 1.2. The small region of the total spectrum occupied by the visible portion is demonstrated by this figure. The higher energy radiation has the smaller wavelength and the larger frequency and wave number [equation(3)]. The following sequence represents decreasing energy:

ultraviolet> visible> infrared> microwave> radio-frequency

1.2 ATOMIC AND MOLECULAR TRANSITIONS

In an atom, the change in state induced by the quantized absorption of radiation can be regarded as the excitation of an electron from one energy state to another. The change in state is

from the ground state to an excited state. In most cases, the energy required for such excitation is in the range from 60 to 150 kcal $mole^{-1}$. Calculation employing equation (3) readily shows that radiation in the ultraviolet and visible regions will be involved. Atomic spectra are often examined as emission spectra. Electrons are excited to higher states by thermal or electrical energy, and the energy emitted as the atoms return to the ground state is measured.

In molecular spectroscopy, absorption of energy is usually measured. Our concern will be with three types of molecular transitions induced by electromagnetic radiation: electronic, vibrational, and rotational. A change in *electronic state* of a molecule occurs when a bonding or non bonding electron of the molecule in the ground state is excited into a higher-energy empty molecular orbital. For example, an electron in a π-bonding orbital of a carbonyl group can be excited into a π^* orbital, producing an excited state with configuration $\sigma^2\pi^1\pi^{*1}$. The electron distributions in the two states (ground and excited) involved in an electronic transition are different.

The *vibrational energy states* are characterized by the directions, frequencies, and amplitudes of the motions that the atoms in a molecule undergo. As an example, two different kinds of vibrations for the SO_2 molecule are illustrated in Figure 1.3. The atoms in the molecule vibrate (relative to their center of mass) in the directions indicated by the arrows, and the two extremes in each vibrational mode are indicated. In the vibration indicated in (a), the sulfur-oxygen bond length is varying, and this is referred to as the *stretching vibration.* In (b), the motion is perpendicular to the bond axis and the bond length is essentially constant. This is referred to as a *bending vibration.* In these vibrations, the net effect of all atomic motion is to preserve the center of mass of the molecule so that there will be no net translational motion. The vibrations indicated in Figure 1.3 are drawn to satisfy this requirement. Certain vibrations in a molecule are referred to as *normal vibrations* or *normal modes.* These are independent, self-repeating displacements of the atoms that preserve the center of mass. In a normal vibration, all the atoms vibrate in

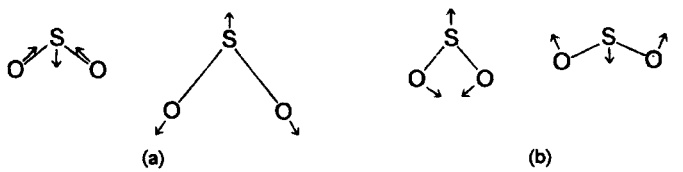


Figure 1.3 Two different vibrations for the SO_2 molecule.

phase and with the same frequency. It is possible to resolve the most complex molecular vibration into a relatively small number of such normal modes. For a non-linear molecule, there are $3N - 6$ such modes, where N is the number of atoms in the molecule. The normal modes can be considered as the $3N - 6$ internal degrees of freedom that (in the absence of anharmonicity) could take up energy independently of each other. The motion of the atoms of a molecule in the different normal modes can be described by a set of *normal coordinates*. These are a set of coordinates defined so as to describe the normal vibration most simply. They often are complicated functions of angles and distances.

The products of the normal modes of vibration are related to the total vibrational state, ψ_v, of a non-linear molecule as shown in equation (5).

$$\psi_v = \prod_{n=1}^{3N-6} \psi_n \qquad ...(5)$$

where Π indicates that the product of the n vibrational modes is to be taken and ψ_n is the wave function for a given normal mode. There exists for each normal vibration (ψ_n) a whole series of excited vibrational states, i, whose harmonic oscillator wave functions are given by:

$$\psi_i = N_i \exp\left[\left(-\frac{1}{2}\right)a_i q^2\right] H_i\left(\sqrt{a_i}q\right) \qquad ...(6)$$

where $i = 0, 1, 2 ...$; H_i is the Hermite polynomial of degree i; $a_i = 2\pi\nu_i/h$; $N_i = [\sqrt{a_1}/(2^i i!\sqrt{\pi})]^{1/2}$; and q is the normal coordinate.

From this, the following wave functions are obtained for the ground, first, and second excited states:

$$\psi_0 = \left(\frac{a_0}{\pi}\right)^{1/4} \exp\left[\left(-\frac{1}{2}\right)a_0 q^2\right] \qquad ...(7)$$

$$\psi_1 = 2\left(\frac{a_1}{\pi}\right)^{1/2} q \exp\left[\left(-\frac{1}{2}\right)a_1 q^2\right] \qquad ...(8)$$

$$\psi_2 = 2\left(\frac{a_2}{\pi}\right)^{1/2} (2a_2 q^2 - 1) \exp\left[\left(-\frac{1}{2}\right)a_1 q^2\right] \qquad ...(9)$$

When ψ_0 is plotted as a function of displacements about the normal coordinate, q, with zero taken as the equilibrium internuclear

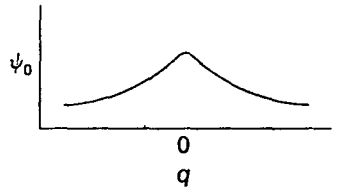


Figure 1.4 A plot of the ground vibrational wave function versus the normal coordinate.

distance, is obtained. This symmetric function results because q appears only as q^2. The same type of plot is obtained for the ground state of all normal modes. It is totally symmetric; accordingly, the total vibrational ground state of a molecule must belong to the totally symmetric irreducible representation, for it is the product of only totally symmetric vibrational wave functions.

The vibrational excited state function ψ_1 has a functional dependence on q that is not of an even power, and accordingly it does not necessarily have a_1 symmetry. When one normal mode is excited in a vibrational transition, the resulting total state is a product of all the other totally symmetric wave functions and the wave function for this first vibrational excited state for the excited normal mode. Thus, the total vibrational state has symmetry corresponding to the normal mode excited.

The *rotational states* correspond to quantized molecular rotation around an axis without any appreciable change in bond lengths or angles. Different rotational states correspond to different angular momenta of rotation or to rotations about different axes. Rotation about the C_2 axis in SO_2 is an example of rotational motion.

In the treatment of molecular spectra, the *Born-Oppenheimer approximation* is invoked. This approximation proposes that the total energy of a system may be regarded as the sum of three independent energies: electronic, vibrational, and rotational. For example, the electronic energy of the system does not change as vibration of the nuclei occurs. The wave function for a given molecular state can then be described by the product of three independent wave functions: ψ_{el}, ψ_{vib}, and ψ_{rot}. As we shall see later, this approximation is not absolutely valid.

The relative energies of these different molecular energy states in typical molecules. Rotational energy states are more closely spaced than are vibrational states, which, in turn, have smaller energy differences than electronic states. The letters ν_0, ν_1, etc., and ν_0',

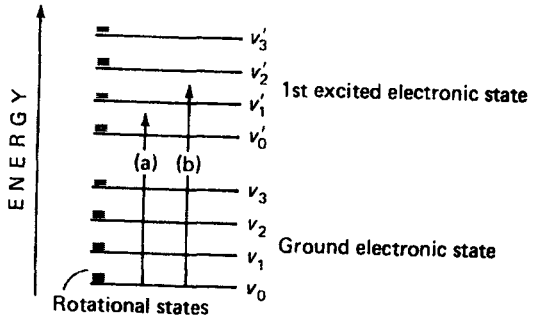


Figure 1.5 Energy states of a diatomic molecule.

v_1', etc., represent the vibrational levels of one vibrational mode in the ground and first electronic excited states, respectively. Ultraviolet or visible radiation is commonly required to excite the molecule into the excited electronic states. Lower-energy infrared suffices for vibrational transitions, while pure rotational transitions are observed in the still lower-energy microwave and radio-frequency regions.

Electronic transitions are usually accompanied by changes in vibration and rotation. Two such transitions are indicated by arrows (a) and (b) of Figure 1.5. In the vibrational spectrum, transitions to different rotational levels also occur. As a result, vibrational fine structure is often detected in electronic transitions. Rotational fine structure in electronic transitions can be detected in high resolution work in the gas phase. Rotational fine structure in vibrational transitions is sometimes observed in the liquid state and generally in the gaseous state.

The energy level diagram in Figure 1.5 is that for a diatomic molecule. For a polyatomic molecule, the individual observed transitions can often be described by diagrams of this type, each transition in effect being described by a different diagram.

1.3 SELECTION RULES

In order for matter to absorb the electric field component of radiation, another requirement in addition to energy matching must be met. The energy transition in the molecule must be accompanied by a change in the electrical center of the molecule in order that electrical work can be done on the molecule by the electromagnetic radiation field. Only if this condition is satisfied can absorption occur. Requirements for the absorption of light by matter are summarized in the *selection rules*. Transitions that are possible according to these rules are referred to as *allowed* transitions, and those not

possible as *forbidden* transitions. It should be noted, however, that the term "forbidden" refers to rules set up for a simple model and, while the model is a good one, "forbidden" transitions may occur by mechanisms not included in the simple model. The intensity of absorption or emission accompanying a transition is related to the probability of the transition, the more probable transitions giving rise to more intense absorption. Forbidden transitions have low probability and give absorptions of very low intensity. This topic and the symmetry aspects of this topic will be treated in detail when we discuss the various spectroscopic methods.

1.4 CHEMICAL PROCESSES AFFECTING THE NATURAL LINE WIDTH OF A SPECTRAL LINE

When two or more chemically distinct species coexist in rapid equilibrium (two conformations, or rapid chemical exchange, *e.g.*, proton exchange between NH_3 and NH_4^+, or other equilibria), one often sees absorption peaks corresponding to the individual species in some forms of spectroscopy; but with other methods, only a single average peak is detected. A process may cause a broadening of the spectral line in some spectral regions, while the same process in some other spectral region has no effect. This behaviour can be understood from the Uncertainty Principle, which states

$$\Delta E \Delta t \sim \hbar$$

or

$$\Delta \nu \Delta t \sim \frac{1}{2\pi} \qquad \ldots(10)$$

Consider the case in which two different sites give rise to two distinct peaks. As the rate of the chemical exchange comes close to the frequency $\Delta\nu$ of the spectroscopic method, the two peaks begin to broaden. As the rate becomes faster, they move together, merge, and then sharpen into a single peak. In subsequent chapters, we shall discuss procedures for extracting rate data over this entire range. Here we shall attempt to gain an appreciation for the time scales corresponding to the various methods by examining the rates that result in line broadening of a sharp single resonance and merging of the two distinct resonances.

First, we shall be concerned with the broadening of a resonance line of an individual species. If Δt is the lifetime of the excited state, we have:

$$\Delta \nu (\text{sec}^{-1}) \sim \frac{1}{2\pi \Delta t (\text{sec})}$$

If there is some chemical or physical process that is fast compared to the excited state lifetime, Δt can be shortened by this process and the line will be broadened. In infrared, for example, it is possible to resolve two bands corresponding to different sites that are separated by 0.1 cm^{-1}, so we can substitute this value into equation (10) and find that lifetimes, Δt, corresponding to this are given by:

$$\Delta\nu = 0.1\,\text{cm}^{-1} \times 3\times10^{10}\ \text{cm/sec} = 3\times10^{9}\,\text{sec}^{-1}$$

$$\Delta t = \frac{1}{(2)(\pi)(3\times10^{9}\,\text{sec}^{-1})} = 5\times10^{-11}\,\text{sec}$$

Therefore, we need a process that will give rise to lifetimes of $\sim 10^{-11}$ sec or less to cause broadening. Since lifetimes are the reciprocals of first order rate constants, this means we need a process whose rate constant is at least 10^{11} sec^{-1} in order to detect a broadening in the infrared band. Diffusion-controlled chemical reactions have rate constants of only 10^{10}, so for systems undergoing chemical exchange, the chemical process will have no influence on the infrared line shape. Rotational motion occurs on a time scale that causes broadening of an infrared or Raman line, and it is possible to obtain information about the motion from the line shape. The inversion doubling of ammonia is also fast enough to affect the Raman and microwave spectra.

In nmr, typical resolution is ~0.1 Hz (cycle per second), which, when substituted into the above equation, shows that a process with a lifetime of about 2 sec (or less) or a rate constant of about 0.5 sec^{-1} (or more) is needed to broaden the spectral line. This is in the range of many chemical exchange reactions.

Next, we shall consider the rate at which processes must occur to result in a spectrum in which only an average line is detected for two species. In order for this to occur, the species must be interconverting so fast that the lifetime in either one of the states is less than Δt, so that only an average line results. This is calculated by substituting the difference between the frequencies of the two states for $\Delta\nu$ in equation (10). In the infrared, a rate constant of 5 $\times$ 10^{13} sec^{-1} or greater would be required to merge two peaks that are separated by 300 cm^{-1}.

$$\Delta t = \frac{1}{(2)(\pi)(9.0\times10^{12}\,\text{sec}^{-1})} = 2\times10^{-14}\,\text{sec}$$

In nmr, for a system in which, for example, the two proton peaks are separated by 100 Hz, one obtains

$$\Delta t = \frac{1}{(2)(\pi)(100\,\text{sec}^{-1})} = 2\times10^{-3}\,\text{sec}$$

Thus, an exchange process involving this proton that had a rate constant of 5×10^2 sec^{-1} would cause these two resonances to appear as one broad line. As the rate constant becomes much larger than 5×10^2 *(e.g.,* by raising the temperature for a positive activation enthalpy process), the broadened single resonance begins to sharpen. Eventually, as the process becomes very fast, the line width is no longer influenced by the chemical process. The radiation is too slow to detect any chemical changes that are occurring and a sharp average line results. This is analogous to the eye being too slow to detect the electronic sweep that produces a television picture.

In x-ray diffraction, the frequency of the radiation is 10^{18} sec^{-1}. Since this is so fast, compared with molecular rearrangements, all we would detect for a dynamic system would be disorder, *i.e.,* contributions from each of the dominant conformations.

In Mossbauer spectroscopy, we observe a very high energy process in which a nucleus in the sample absorbs a γ-ray from the source, There are no chemical processes that affect the lifetime of the nuclear excited state, which for iron is 10^{-7} sec. Thus, if there is a chemical process occurring in an iron sample that equilibrates two iron atoms with a first order rate constant greater than 10^7 sec^{-1}, the Mossbauer spectrum will reveal only an average peak. In dichlorobisphenanthroline iron (III), the equilibrium mixture of high spin (five unpaired electrons) and low spin (one unpaired electron) iron complexes interconvert so rapidly (they have a lifetime $<10^{-7}$ sec) that only a single iron species is observed in the spectrum. For a ruthenium nucleus, the lifetime of the nuclear excited state is 10^{-9} sec; this determines the time scale for experiments with this nucleus. A similar situation pertains in x-ray photoelectron spectroscopy, in that the time scale for ejection of an electron cannot be influenced by chemical processes. The lifetime for the resulting excited state is about 10^{-18} sec.

1.5 GENERAL APPLICATIONS

The following general applications of spectroscopy are elementary and pertain to both vibrational and electronic spectroscopic methods: (1) determination of concentration, (2) "fingerprinting," and (3) determination of the number of species in solution by the use of isosbestic points.

1.5.1 Determination of Concentration

Measurement of the concentrations of species has several important applications. If the system is measured at equilibrium, equilibrium constants can be determined. By evaluating the equilibrium constant, K, at several temperatures, the enthalpy ΔH^o for the equilibrium reaction can be calculated from the van't Hoff equation:

$$\log K = \frac{\Delta H^0}{2.3RT} + C \qquad ...(11)$$

Determination of the change in concentration of materials with time is the basis of kinetic studies that give information about reaction mechanisms. In view of the contribution of results from equilibrium and kinetic studies to our understanding of chemical reactivity, the determination of concentrations by spectroscopic methods will be discussed.

The relationship between the amount of light absorbed by certain systems and the concentration of the absorbing species is expressed by the Beer-Lambert law:

$$A = \log_{10} \frac{I_0}{I} = \varepsilon cb \qquad ...(12)$$

where A is the absorbance, I_0 is the intensity of the incident light, I is the intensity of the transmitted light, ε is the molar absorptivity (sometimes called extinction coefficient) at a given wavelength and temperature, c is the concentration (the molarity if ε is the molar absorptivity), and b is the length of the absorbing system. The molar absorptivity varies with both wavelength and temperature, so these must be held constant when using equation (12). When matched cells are employed to eliminate scattering of the incident beam, there are no exceptions to the relationship between absorbance and b (the Lambert law). For a given concentration of a certain substance, the absorbance is always directly proportional to the length of the cell. The part of equation (12) relating absorbance and concentration ($\log_{10} I_o/I = \varepsilon c$, for a constant cell length) is referred to as Beer's law. Many systems have been found that do not obey Beer's law. The anomalies can be attributed to changes in the composition of the system with concentration *(e.g.,* different degrees of ionization or dissociation of a solute at different concentrations). For all systems, the Beer's law relationship must be demonstrated, rather than assumed, over the entire concentration range to be considered. If Beer's law is obeyed, it becomes a simple matter to use equation

(12) for the determination of the concentration of a known substance if it is the only material present that is absorbing in a particular region of the spectrum. The Beer's law relationship is tested and £ determined by measuring the absorbances of several solutions of different known concentrations covering the range to be considered. For each solution, ε can be calculated from:

$$\varepsilon = \frac{A(1\text{ cm cell})}{c, \text{molarity}} \qquad ...(13)$$

where the units of e are liters mole^{-1} cm^{-1}. For a solution of this material of unknown concentration, the absorbance is measured, ε is known (13), and c is calculated from equation (12).

There are some interesting variations on the application of Beer's law. If the absorption of two species should overlap, this overlap can be resolved mathematically and the concentrations determined. This is possible as long as the two ε values are not identical at all wavelengths. Consider the case in which the ε values for two compounds whose spectra overlap can be measured for the pure compounds. The concentration of each component in a mixture of the two compounds can be obtained by measuring the absorbance at two different wavelengths, one at which both compounds absorb strongly and a second at which there is a large difference in the absorptions. Both wavelengths should be selected at reasonably flat regions of the absorption curves of the pure compounds, if possible. Consider two such species B and C, and wavelengths λ_1 and λ_2. There are molar absorptivities of $\varepsilon_{B\lambda 1}$ for B at λ_1, $\varepsilon_{B\lambda 2}$ for B at λ_2, and similar quantities $\varepsilon_{C\lambda 1}$ and $\varepsilon_{C\lambda 2}$ for C. The total absorbance of the mixture at λ_1 is A_1 and that at λ_2 is A_2. It follows that:

$$A_1 = x\varepsilon_{B_{\lambda 1}} + y\varepsilon_{C_{\lambda 1}} \qquad ...(14)$$

and

$$A_2 = x\varepsilon_{B_{\lambda 2}} + y\varepsilon_{C_{\lambda 2}} \qquad ...(15)$$

where x = molarity of B and y = molarity of C. The two simultaneous equations (14) and (15) have only two unknowns, x and y, and can be solved.

A situation often encountered in practice involves an equilibrium of the type

$$D + E \rightleftharpoons DE$$

Here the spectra of, say, D and DE overlap but E does not absorb; if the equilibrium constant is small, pure DE cannot be obtained, so

its molar absorptivity cannot be determined directly. It is possible to solve this problem for the equilibrium concentrations of all species, i.e., to determined the equilibrium constant:

$$K = \frac{[DE]}{[D][E]} \qquad ...(16)$$

Let $[D]_0$ be the initial concentration of D. This can be measured, as can $[E]_0$, the initial concentration of E. The molar absorptivity of DE, ε_{DE}, cannot be determined directly, but it is assumed that Beer's law is obeyed. It can be seen from a material balance that:

$$[D] = [D]_0 - [DE] \qquad ...(17)$$

$$[E] = [E]_0 - [DE] \qquad ...(18)$$

so

$$K = \frac{[DE]}{([D]_0 - [DE])([E]_0 - [DE])} \qquad ...(19)$$

The total absorbance consists of contributions from [D] and [DE] according to

$$A = \varepsilon_D[D] + \varepsilon_{DE}[DE] \qquad ...(20)$$

Substituting equation (17) into (20) and solving for [DE], one obtains

$$[DE] = \frac{A - \varepsilon_D[D]_0}{\varepsilon_{DE} - \varepsilon_D}$$

but $\varepsilon_D[D]_0$ is the initial absorbance, A^0, of a solution of D with concentration $[D]_0$ without any E in it, so

$$[DE] = \frac{A - A^0}{\varepsilon_{DE} - \varepsilon_D} \qquad ...(21)$$

When equation (21) is substituted into the equilibrium constant expression, equation (19), and this is rearranged, we have

$$K^{-1} = \frac{A \cdots A^0}{\varepsilon_{DE} - \varepsilon_D} - [D]_0 - [E]_0 + \frac{[D]_0[E]_0(\varepsilon_{DE} - \varepsilon_D)}{A - A^0} \qquad ...(22)$$

The advantage of this equation is that it contains only two unknown quantities, ε_{DE} and K^{-1}. Furthermore, these unknowns are constant for any solution of different concentrations $[D]_0$ and $[E]_0$ that we care to make up. For two different sets of experimental conditions (different values of $[D]_0$ and $[E]_0$) two simultaneous equations can be solved; ε_{DE} is eliminated, and K is obtained. If several sets of experimental conditions are employed, all possible combinations of simultaneous equations can be considered by

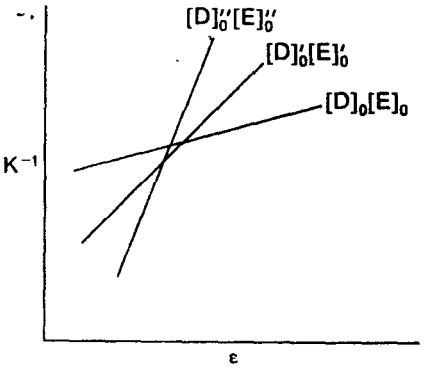


Figure 1.6 Graphical solution of equation (22).

employing a least-squares computer analysis that finds the best values of K^{-1} and ε_{DE} to reproduce the experimental absorbances. These least-squares procedures usually provide an error analysis that is most important to have (*vide infra*). There are several advantages to a graphical display of the simultaneous equations that are being analyzed. The graph is constructed by taking a solution of known $[D]_0$, $[E]_0$, and A, and calculating the values of K^{-1} that would result by selecting several different values of ε_{DE} near the expected value. These results are plotted, by the line $[D]_0[E]_0$*. The procedure is repeated for other initial concentrations, *e.g.*, $[D]_0'[E]_0'$ and $[D]_0"[E]_0"$. The intersection of any two of these lines is a graphical representation of the solution of two simultaneous equations. The intersection of all the calculated curves should occur at a point whose values of K^{-1} and ε satisfy all the experimental data. This common intersection justifies the Beer's law assumption used in the derivation because it indicates a unique value for ε or all concentrations. As a result of experimental error, a triangle usually results: stead of a point. The best K^{-1} and ε to fit the data are then found by the least-squares procedure. When the experiment described above is not designed properly, a set of concentrations of D and E are employed that result in a set of parallel lines for the K^{-1} *vs.* ε plots. The slope of one of these lines is given by taking the partial derivative of the K^{-1} expression in (22) with respect to $\varepsilon_{DE} - \varepsilon_D$:

$$\frac{\partial K^{-1}}{\partial_{DE} - \varepsilon_D} = -\frac{A - A^0}{(\varepsilon_{DE} - \varepsilon_D)^2} + \frac{[D]_0[E]_0}{A - A^0} \qquad ...(23)$$

Since the first term is generally small, we see from equation (23) that if experimental conditions are selected for a series of

experiments in which [DE] or the value of A nearly doubles every time $[D]_0$ or $[E]_0$ is doubled, then the various resulting K^{-1} and *vs.* ε plots will be nearly parallel. The values of K^{-1} and ε obtained from the computer analysis of this kind of data are thus highly correlated and should be considered undefined, for the simultaneous equations are essentially dependent ones. The K^{-1} *vs.* ε plots have been described in terms of the results from the error analysis of the least-squares calculation.

The approach described above for the evaluation of equilibrium constants is a general one that applies to any form of spectroscopy or any kind of measurement in which the measured quantity is linearly related to concentration (*i.e.*, in which a counterpart to equation (20) exists). Similar analyses have been described for calorimetric data and for nmr spectral data.

There have been many reports in the literature of attempts to solve simultaneous equations for equilibrium constants for two or more consecutive equilibria:

$$\begin{aligned} A + B &\rightleftharpoons AB \qquad \text{step 1} \\ AB + B &\rightleftharpoons AB_2 \qquad \text{step 2} \end{aligned}$$

Usually ε_{AB}, ε_{AB2}, and the stepwise equilibrium constants K_1 and K_2 are unknown. In most instances, even though the reported parameters fit the experimental data well, the system is undefined. Careful examination often shows that many other, very different combinations of parameters also fit the data. To solve this problem, one must find a region of the spectrum in which AB makes the main contribution to the absorbance and another in which AB_2 makes the predominant contribution. Only by working at both wavelengths can one solve for all the unknowns in a rigorous fashion. This is impossible in many nmr, uv-visible, and calorimetric experiments because properties related predominantly to the individual AB and AB_2 species cannot be monitored separately. This problem has been discussed in detail in the literature and several examples are given there.

1.5.2 Isosbestic Points

If two substances at equal concentrations have absorption bands that overlap, there will be some wavelength at which the molar absorptivities of the two species are equal. If the sum of the concentrations of these two materials in solution is held constant, there will be no change in absorbance at this wavelength as the ratio of the two materials is varied. For example, assume a reaction

A + B → AB, in which only A and AB absorb; the sum of [A] plus [AB] will be constant as long as the initial concentration of A is held constant as B is varied. Since the absorbance of the solution is given by

$$\text{abs} = \varepsilon_A[A] + \varepsilon_{AB}[AB] \qquad ...(24)$$

the absorbance will be constant when $\varepsilon_A = \varepsilon_{AB}$ and [A] plus [AB] is held constant. The invariant point obtained for this system is referred to as the *isosbestic point.* The existence of one or more isosbestic points in a system provides information regarding the number of species present. For example, the spectra were obtained by keeping the total iron concentration constant in a system consisting of:

Curve 1 2.1 moles of LiCl per $Fe(DMA)_6(ClO_4)_3$
Curve 2 2.6 moles of LiCl per $Fe(DMA)_6(ClO_4)_3$
Curve 3 3.1 moles of LiCl per $Fe(DMA)_6(ClO_4)_3$
Curve 4 4.1 moles of LiCl per $Fe(DMA)_6(ClO_4)_3$

where DMA is the abbreviation for *N,N*-dimethylacetamide. Points *A* and *B* are isosbestic points; their existence suggests that the absorption in this region is essentially accounted for by two species. Curve 4 is characteristic of $FeCl_4^-$. A study of solutions more dilute in chloride establishes the existence of a species $Fe(DMA)_4Cl_2^+$, which exists at a 2:1 ratio of Cl^- to Fe^{III} and absorbs in this region. The isosbestic points indicate that the system can be described by the species $Fe(DMA)_4Cl_2^+$ and $FeCl_4^-$ over the region from 2:1 to 4:1 ratios of Cl^- to Fe^{III}. The addition compound $FeCl_3 \cdot DMA$ probably does not exist in appreciable concentration in this system, *i.e.,* at total Fe^{III} concentrations of $2 \times 10^{-4}M$. Curve 4 in this spectrum does not pass through the isosbestic points. Small deviations *(e.g.,* curve 4 at point *A)* may be due to experimental inaccuracy, changes in solvent properties in different solutions, or a small concentration of a third species, probably $FeCl_3 \cdot DMA$, present in all systems except that represented by curve 4.

The conclusion that only two species are present in appreciable concentrations could be in error if $FeCl_3 \cdot DMA$ or other species were present that had molar absorptivities identical to that of the above two ions at the isosbestic points. However, if equilibrium constants for the equilibrium

$$Fe(DMA)_4Cl_2^+ + 2Cl^- \rightleftharpoons FeCl_4^- + 4DMA$$

are calculated from these different curves at different wavelengths, the possibility of a third species is eliminated if the constants agree.

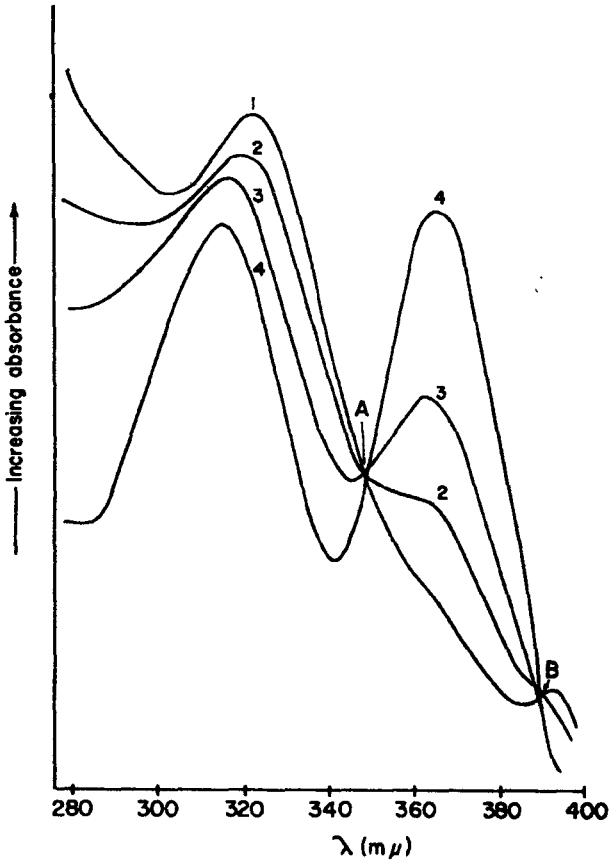


Figure 1.7 Spectra of the $Fe(DMA)_4Cl_2^+$–LiCl system in *N,N*-dimethylacetamide as solvent.

An interesting situation results in which an isosbestic point can be obtained in solution when more than two species with differing extinction coefficients exist if the base (or acid) employed has two donor (or acceptor) sites. For example, if DMA were to coordinate to an acid A to produces an oxygen-bound complex and a nitrogen-bound one, the mixture of complexes AN (nitrogen bound), AO (oxygen bound) and free A (an absorbing Lewis acid) will give rise to an isosbestic point. The absorbance for such a system is given by equation (25):

$$\text{abs} = \varepsilon_A[A] + \varepsilon_{AO}[AO] + \varepsilon_{AN}[AN] \qquad ...(25)$$

The equilibrium constant expressions are given by

$$K_0 = \frac{[AO]}{[A][B]} \text{ and } K_N = \frac{[AN]}{[A][B]}$$

$$K_0 + K_N = \frac{[AO] + [AN]}{[A][B]} = \frac{[AB]}{[A][B]} \quad ...(26)$$

where we define [AO] + [AN] = [AB]. The fraction of complex that is oxygen-coordinated, X_{AO}, is given by

$$x_{AO} = \frac{K_0}{K_0 + K_N} = \frac{\frac{[AO]}{[A][B]}}{\frac{[AB]}{[A][B]}} = \frac{[AO]}{[AB]} \quad ...(27)$$

The fraction that is nitrogen-coordinated is similarly derived as:

$$X_{AN} = \frac{[AN]}{[AB]} \quad ...(28)$$

Now the total absorbance, abs, becomes

$$\text{abs} = \varepsilon_A[A] + \varepsilon_{AO}X_{AO}[AB] + \varepsilon_{AN}X_{AN}[AB]$$

or

$$\text{abs} = \varepsilon_A[A] + (\varepsilon_{AO}X_{AO} + \varepsilon_{AN}X_{AN})[AB] = \varepsilon_A[A] + \varepsilon'[AB] \quad ...(29)$$

Since the sum of [A] and [AB] is constant, and there must be a point in the overlapping spectra at which $\varepsilon = \varepsilon'$, an isosbestic point will be obtained even though three absorbing species exist. This is true because

$$\frac{X_{AN}}{X_{AO}} = \frac{K_N}{K_O} = \text{a constant}$$

We have taken two absorbing species whose ratio is a constant, independent of the parameter being varied, and we have translated them into what is effectively a single absorbing species *via* equation (29). The general criterion is thus that 2 + N absorbing species will give rise to an isosbestic point if there are N *independent* equations of the form

$$\frac{[Y]}{[Z]} = k \quad ...(30)$$

where Y and Z are two of the absorbing species in the system and the value of k is independent of the parameter being varied.

As an example of the application of this criterion, consider an acid that can form two isomers with a base, AB and A*B *(e.g.,* $Cu(hfac)_2$ forming basal and apical adducts); let this acid form an

adduct with a base that can form two isomers with an acid, AB and AB* *(e.g., N*-methyl imidazole bound through the amine and imine nitrogens). Five absorbing species (A, AB, A*B, AB*, and A*B*) can exist,

$$A + B \rightarrow AB$$
$$A + B \rightarrow A^*B$$
$$A + B \rightarrow AB^*$$
$$A + B \rightarrow A^*B^*$$

so three constants are needed. The equilibrium constants are

$$K_1 = \frac{[AB]}{[A][B]} \quad K_2 = \frac{[A^*B]}{[A][B]} \quad K_3 = \frac{[AB^*]}{[A][B]} \quad K_4 = \frac{[A^*B^*]}{[A][B]}$$

We now note that three independent ratios exist, which are independent of [B]:

$$\frac{K_1}{K_2} = \frac{[AB]}{[A^*B]} \quad \frac{K_2}{K_3} = \frac{[A^*B]}{[AB^*]} \quad \frac{K_3}{K_4} = \frac{[AB^*]}{[A^*B^*]}$$

Therefore, an isosbestic point is expected if $[A]_0$ is held constant and $[B]_0$ is varied. Any other ratio of equilibrium constants, *e.g.*, K_1/K_3, is not independent of the three ratios written. The general rules presented here apply to a large number of systems. However, rote application of rules is no substitute for an understanding of the systems under consideration.

1.5.3 Job's Method of Isomolar Solutions

By examining the spectra of a series of solutions of widely varying mole ratios of A to B, but with the same total number of moles of A + B, the stoichiometry of complexes formed between A and B in solution can often be determined. Absorbance at a wavelength of maximum change is plotted against the mole ratio of A to B; the latter is usually used as the abscissa. The plots have at least one extremum, often a maximum. In simple cases, the extrema occur at mole fractions corresponding to the stoichiometry of the complexes that form in solution.

1.5.4 Fingerprinting

This technique is useful for the identification of an unknown compound that is suspected to be the same as a known compound. The spectra are compared with respect to ε values, wavelengths of maximum absorption, and band shapes. In addition to this direct comparison, certain functional groups have characteristic absorptions in various regions of the spectrum. For example, the carbonyl group

will generally absorb at certain wavelengths in the ultraviolet and infrared spectra. Its presence in an unknown compound can be determined from these absorptions. Often one can even determine whether or not the carbonyl group is in a conjugated system. These details will be considered later when the spectroscopic methods are discussed individually.

Spectroscopic methods provide a convenient way of detecting certain impurities in a sample. For example, the presence of water in a system can easily be detected by its characteristic infrared absorption. Similarly, a product can be tested for absence of starting material if the starting material has a functional group with a characteristic absorption that disappears during the reaction. Spectral procedures are speedier and less costly than elemental analyses. The presence of contaminants in small amounts can be detected if their molar absorptivities are large enough.

2

Electronic Absorption Spectroscopy

Prior to a discussion of electronic absorption spectroscopy, the information summarized by a potential energy curve for a diatomic molecule will be reviewed. Figure 2.1 is a plot of E, the total energy of the system, *versus* r, the internuclear distance, and is one of many types of potential functions referred to as a Morse potential. The curve is expressed mathematically by,

$$V = D\{1 - \exp[-\nu_0(2\pi^2\mu D)^{1/2}(r - r)]\}^2$$

All terms are defined in Figure 2.1 except μ, which is the reduced mass $(m_1m_2)/(m_1 + m_2)$.

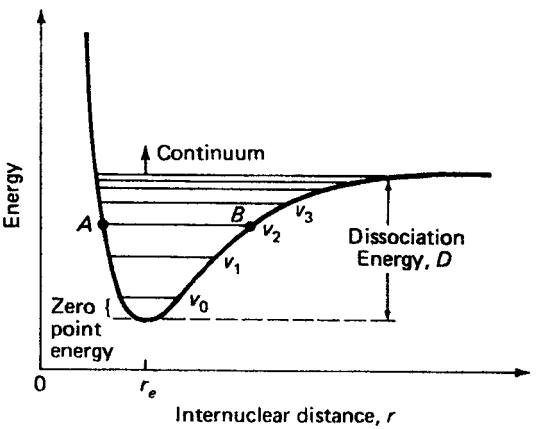


Figure 2.1 Morse energy curve for a diatomic molecule.

As the bond distance is varied in a given vibrational state, e.g., along *A-B* in the υ_2 level, the molecule is in a constant vibrational energy level. At *A* and *B* we have, respectively, the minimum and maximum values for the bond distance in this vibrational level. At these points the atoms are changing direction, so the vibrational kinetic energy is zero and the total vibrational energy of the system is potential. At r_e, the equilibrium internuclear distance, the vibrational kinetic energy is a maximum and the vibrational potential energy is zero. Each horizontal line represents a different vibrational energy state. The ground state is υ_0 and excited states are represented as υ_1 υ_2 etc. Eventually, if enough energy can be absorbed in vibrational modes, the molecule is excited into the continuum and it dissociates. For most compounds, nearly all the molecules are in the υ_0 level at room temperature because the energy difference $\upsilon_1 - \upsilon_0$ is usually much larger than *kT* (thermal energy), which has a value of 200 cm^{-1} at 300°K.

Each excited electronic state also contains a series of different vibrational energy levels and may be represented by a potential energy curve. The ground electronic state and one of the many excited electronic states for a typical diatomic molecule are illustrated in Figure 2.2. Each vibrational level, υ_n, is described by a vibrational wave function, ψ_{vib}. For simplicity only four levels are indicated. The square of a wave function gives the probability distribution, and in this case ψ_{vib}^2 indicates probable internuclear distances for a particular vibrational state. This function, ψ_{vib}^2, is indicated for the various levels by the dotted lines in Figure 2.2. The dotted line is not related to the energy axis. The higher this line, the more probable the corresponding internuclear distance. The most probable distance for a molecule in the ground state is r_e, while there are two most probable distances corresponding to the two maxima in the next vibrational energy level of the ground electronic state, three in the third, etc. In the excited vibrational levels of the ground and excited electronic states, there is high probability of the molecule having an internuclear distance at the ends of the potential function.

2.1 RELATIONSHIP OF POTENTIAL ENERGY CURVES TO ELECTRONIC SPECTRA

An understanding of electronic absorption spectroscopy requires consideration of three additional principles:

1. In the very short time required for an electronic transition to take place (about 10^{-15} sec), the atoms in a molecule do not

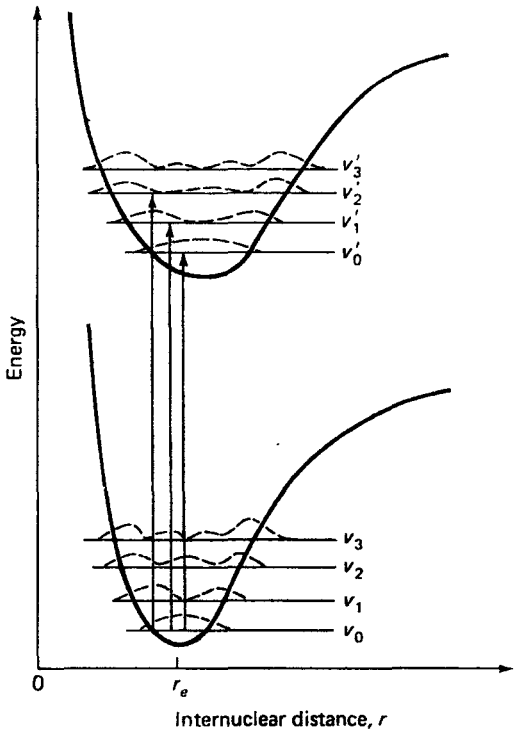


Figure 2.2 Morse curves for a ground and excited state of a diatomic molecule, showing vibrational probability functions, ψ_{vib}^2, as dotted lines.

have time to change position appreciably. This statement is referred to as the *Franck-Condon principle*. Since the electronic transition is rapid, the molecule will find itself with the same molecular on figuration and vibrational kinetic energy in the excited state that it had in the ground state at the moment of absorption of the photon. As a result, all electronic transitions are indicated by a vertical line on the Morse potential energy diagram of the ground and excited states; i.e., there is no change in internuclear distance during the transition.

2. There is no general selection rule that restricts the change in vibrational state accompanying an electronic transition. Frequently transitions occur from the ground vibrational level of the ground electronic state to many different vibrational levels of a particular excited electronic state. Such transitions may give rise to vibrational fine structure in the main peak of the electronic transition.

The three transitions indicated by arrows could give rise to three peaks. Since nearly all of the molecules are present in the ground vibrational level, nearly all transitions that give rise to a peak in the absorption spectrum will arise from υ_0. Transitions from this ground level (υ_0) to υ_0', υ_1', or υ_2' are referred to as $0 \rightarrow 0$, $0 \rightarrow 1$, or $0 \rightarrow 2$ transitions, respectively. It can be shown that the relative intensity of the various vibrational sub-bands depends upon the vibrational wave function for the various levels. A transition is favoured if the probabilities of the ground and excited states of the molecule are both large for the same internuclear distance. The spectrum could result from a substance in solution undergoing the three transitions. The $0 \rightarrow 0$ transition is the lowest energy-longest wavelength transition. The differences in wavelength at which the peaks occur represent the energy differences of the vibrational levels in the excited state of the molecule. Much information about the structure and configuration of the excited state can be obtained from the fine structure.

Electronic transitions from bonding to antibonding molecular orbitals are often encountered. In this case the potential energy curve for the ground state will be quite different from that of the excited state because there is less bonding electron density in the excited state. As a result, the equilibrium internuclear distance will be greater and the potential energy curve will be broader for the excited state. Because of this displacement of the excited state potential energy curve, the $0 \rightarrow 0$ and transitions to other low vibrational levels may not be observed.

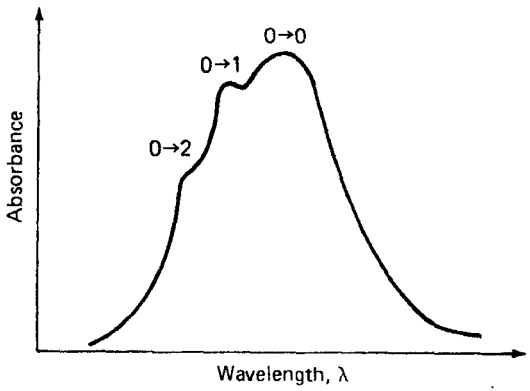


Figure 2.3 Spectrum corresponding to the potential energy curves.

A transition to a higher vibrational level becomes more probable. This can be visualized by broadening and displacing the excited state.

3. There is an additional *symmetry requirement*, which was neglected for the sake of simplicity in the above discussion. It has been assumed that this symmetry requirement is satisfied for the transitions involved in this discussion. This will be discussed subsequently in the section on selection rules.

The above discussion pertains to a diatomic molecule, but the general principles also apply to a polyatomic molecule. Often the functional group in a polyatomic molecule can be treated as a diatomic molecule. The electronic transition may occur in the functional group between orbitals that are approximated by a combination of atomic orbitals of the two atoms as in a diatomic molecule. The actual energies of the resulting molecular orbitals of the functional group will, of course, be affected by electronic, conjugative, and steric effects arising from the other atoms. This situation can be understood qualitatively in terms of potential energy curves similar to those discussed for the diatomic molecules. For more complex cases in which several atoms in the molecule are involved (i.e., a delocalized system), a polydimensional surface is required to represent the potential energy curves.

The energies required for electronic transitions generally occur in the far uv, uv, visible, and near infrared regions of the spectrum, depending upon the energies of the molecular orbitals in a molecule. For molecules that contain only strong sigma bonds (e.g., CH_4 and H_2O), the energy required for electronic transitions occurs in the far uv region of the spectrum and requires specialized instrumentation for detection. In fact, in the selection of suitable transparent solvents for study of the u.v. spectrum of the dissolved solute, such transitions come into play in determining the cutoff point of the solvent. On the other hand, the coloured dyes used in clothing have extensively conjugated pi-systems and undergo electronic transitions in the visible region of the spectrum. Standard instruments cover the region from 50,000 cm^{-1} to 5,000 cm^{-1}. This spectral region is subdivided as follows:

Ultraviolet	50,000 cm^{-1} to 26,300 cm^{-1} (2000 to 3800 Å)
Visible	26,300 cm^{-1} to 12,800 cm^{-1} (3800 to 7800 Å)
Near infrared	12,800 cm^{-1} to 5,000 cm^{-1} (7800 to 20,000 Å)

2.2 NOMENCLATURE

In our previous discussion we were concerned only with transitions of an electron from a given ground state to a given excited state. In an actual molecule there are electrons in different kinds of orbitals (σ bonding, nonbonding, π bonding) with different energies in the ground state. Electrons from these different orbitals can be excited to higher-energy molecular orbitals, giving rise to many possible excited states. Thus, many transitions from the ground state to different excited states (each of which can be described by a different potential energy curve) are possible in one molecule.

There are several conventions used to designate these different electronic transitions. A simple representation introduced by Kasha will be illustrated for the carbonyl group in formaldehyde. A molecular orbital description of the valence electrons in this molecule is:

$$\sigma^2 \quad \pi^2 \quad n_a^2 \quad n_b^2 \quad (\pi^*)^0 \quad (\sigma^*)^0$$

The n_a and n_b orbitals are the two non-bonding molecular orbitals containing the lone pairs on oxygen. Symmetry considerations do not require the lone pairs to be degenerate, for there are no doubly degenerate irreducible representations in C_{2v}. They are not accidentally degenerate either, but differ in energy. The ordering of these orbitals can often be arrived at by intelligent guesses and by looking at the spectra of analogous compounds. Also indicated with arrows are some transitions chosen to illustrate the nomenclature. The transitions (1), (2), (3), and (4) are referred to as $n \rightarrow \pi^*$, $n \rightarrow \sigma^*$, $\pi \rightarrow \pi^*$, and $\sigma \rightarrow \sigma^*$, respectively. The $n \rightarrow \pi^*$ transition is the lowest energy-highest wavelength transition that occurs in formaldehyde and most carbonyl compounds.

Electron excitations can occur with or without a change in the spin of the electron. If the spin is not changed in a molecule

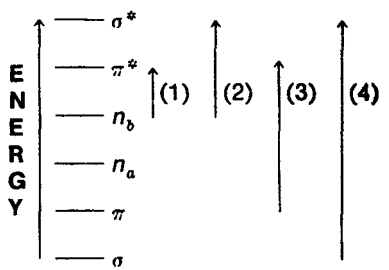


Figure 2.4 Relative energies of the carbonyl molecular orbitals in H_2CO.

containing no unpaired electrons, both the excited state and ground state have a multiplicity of one; these states are referred to as singlets. The *multiplicity* is given by two times the sum of the individual spins, m_s' plus one: $2S+1 = 2\Sigma m_s + 1$. If the spin of the electron is changed in the transition, the excited state contains two unpaired electrons with identical magnetic spin quantum numbers, has a multiplicity of 3, and is referred to as a triplet state.

There are some shortcomings of this simple nomenclature for electronic transitions. It has been assumed that these transitions involve a simple transfer of an electron from the ground state level to an empty excited state level of our ground state wave function. For many applications, this description is precise enough. In actual fact, such transitions occur between states, and the excited state is not actually described by moving an electron into an empty molecular orbital of the ground state. The excited state has, among other things, different electron-electron repulsions than those in the system represented by simple excitation of an electron into an empty ground state orbital. In most molecules, various kinds of electron-electron interactions in the excited state complicate the problem; in addition to affecting the energy, they give rise to many more transitions than predicted by the simple picture of electron promotion because levels that would otherwise be degenerate are split by electronic interactions. This problem is particularly important in transition metal ion complexes. We shall return to discuss this problem more fully in the section on configuration interaction.

In a more accurate system of nomenclature the symmetry, configuration, and multiplicity of the states involved in the transition are utilized in describing the transition. This system of nomenclature can be briefly demonstrated by again considering the molecular orbitals of formaldehyde. The diagrams qualitatively represent the boundary contours of the molecular orbitals. The solid line encloses the positive lobe and the dashed line the negative lobe. The larger π and π^* lobes indicate lobes above the plane of the paper, and the smaller ones represent those below the plane; the two lobes actually have identical sizes. To classify these orbitals it is necessary first to determine the overall symmetry of the molecule, which is C_{2v}. The next step is to consult the C_{2v} character table. By convention, the *yz* plane is selected to contain the four atoms of formaldehyde. The symmetry operations E, C_2, $\sigma_{v(xz)}$, and $\sigma'_{v(yz)}$ performed on the π orbital produce the result $+1$, -1, $+1$, -1. This result is identical to that listed for the irreducible representation B_1 in the table. The

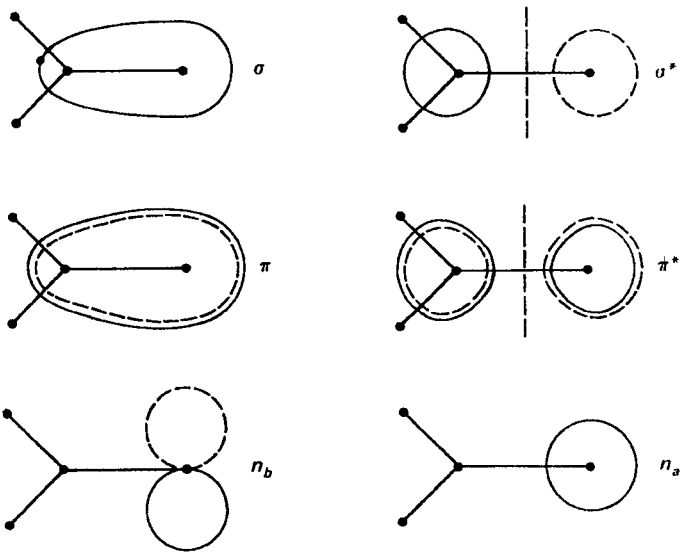


Figure 2.5 Shapes of the molecular orbitals of formaldehyde.

orbital is said to belong to (or to transform as) the symmetry species b_1, the lower case letter being employed for an orbital and the upper case letters being reserved to describe the symmetry of the entire ground or excited state. Similarly, if the n_a, n_b, π^*, σ, and σ^* orbitals of formaldehyde are subjected to the above symmetry operations, it can be shown that these orbitals belong to the irreducible representations a_1, b_2, b_1, a_1, and a_1, respectively. The two n orbitals can be viewed as s and p_y orbitals (p_z is used in the σ bond and p_x in the π). As a result, they lie in the yz plane and possess a_1 and b_2 symmetry. (The a_1 s orbital can mix with p_z in the sigma bonding.) The difference in s orbital character causes the energies of the a_1 and b_2 lone pairs to differ. Molecular orbital calculations are consistent with these ideas. As mentioned, a and b indicate single degeneracy. The a representation does not change sign with rotation about the n-fold axis, but b does.

Table 2.1 Character table for the C_{2v} point group

	E	C_2	$\sigma_{v(xz)}$	$\sigma'_v(yz)$	
A_1	1	1	1	1	z
A_2	1	1	−1	−1	R_z
B_1	1	−1	1	−1	x, R_y
B_2	1	−1	−1	1	y, R_x

The symmetry species of a state is the product of the symmetry species of each of the odd electron orbitals. In the state that results from the $n \rightarrow \pi^*$ transition in formaldehyde, there is one unpaired electron in the n orbital with b_2 symmetry and one in the π^* orbital with b_1 symmetry. The direct product is given by:

	E	C_2	$\sigma_{v(xz)}$	$\sigma_{v'(yz)}$	
$b_1 \times b_2 =$	(1)(1)	(−1)(−1)	(1)(−1)	(−1)(+1)	
result	+1	+1	−1	−1	$= A_2$

The resulting irreducible representation is A_2. The excited state from this transition is thus described as A_2 and the transition as $A_1 \rightarrow A_2$. A common convention involves writing the high energy state first and labeling the transition $A_2 \leftarrow A_1$. The spin multiplicity is usually included, so the complete designation becomes $^1A_2 \leftarrow {}^1A_1$. The ground state is A_1 because there is a pair of electrons in each orbital. Commonly, the orbitals involved in the transitions are indicated, and the symbol becomes $^1A_2(n, \pi^*) \leftarrow {}^1A_1$. If a general symbol is needed for a state symmetry species, Γ is employed.

Instead of representing the orbitals of formaldehyde symbolically, we could simply have been given the wave functions. We can deduce the symmetry from ψ by converting the wave functions into a physical picture. The following equations describe the formaldehyde π and π^* orbitals:

$$\psi_\pi = a\varphi_{p^o} + b\varphi_{p^c}$$

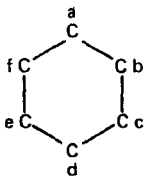
$$\psi_{\pi^*} = b'\varphi_{p^o} + a'\varphi_{p^c}$$

where φ_{p^o} and φ_{p^c} are the wave functions for the atomic oxygen and carbon p orbitals, respectively. The atomic orbitals are mathematically combined to produce the π and π^* orbitals. Since oxygen is more electronegative (i.e., $a > b$), it becomes clear why it is often stated that an electron is transferred from oxygen to carbon in the $\pi \rightarrow \pi^*$ transition. We shall return to a discussion of the experimental spectrum of formaldehyde shortly.

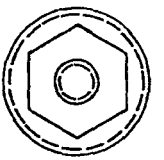
Some of the molecular orbitals for benzene are represented mathematically and pictorially. Notice that a difference in sign between adjacent atomic orbitals of the wave function represents a node (point of zero probability) in the molecular orbital. Using the D_{6h} character table, the symmetries of the orbitals can be shown to be a_{2u} e_{1g}, e_{2u}' and b_{2g}, for ψ_1, $\psi_2 + \psi_3$, $\psi_4 + \psi_5$, and ψ_6, respectively.

In addition to the above conventions used to label ground and excited states, a convention used for diatomic molecules will be

Lettering sequence for a.o.'s



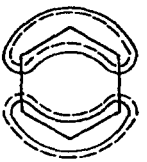
a_{2u} $\psi_1 = 1/\sqrt{6}\,(\varphi_a + \varphi_b + \varphi_c + \varphi_d + \varphi_e + \varphi_f)$



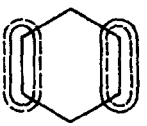
zero nodes

e_{1g}

$\psi_2 = 1/\sqrt{12}(2\varphi_a + \varphi_b - \varphi_c - 2\varphi_d - \varphi_e + \varphi_f)$



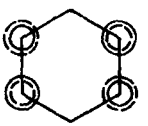
$\psi_3 = 1/2\ (\varphi_b + \varphi_c - \varphi_e - \varphi_f)$



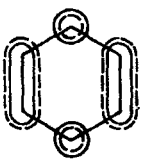
Degenerate set — one node

e_{2u}

$\psi_4 = 1/2\ (\varphi_b - \varphi_c + \varphi_e - \varphi_f)$

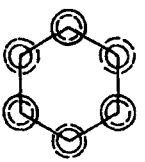


$\psi_5 = 1/\sqrt{12}(2\varphi_a - \varphi_b - \varphi_e + 2\varphi_d - \varphi_e - \varphi_f)$



Degenerate set — two nodes

b_{2g} $\psi_6 = 1/\sqrt{6}\,(\varphi_a - \varphi_b + \varphi_c - \varphi_d + \varphi_e - \varphi_f)$



Three nodes

Figure 2.6 Shape of the benzene molecular orbitals.

described. With this terminology the electronic arrangement is indicated by summing up the contributions of the separate atoms to obtain the net orbital angular momentum. If all electrons are paired, the sum is zero. Contributions are counted as follows: one unpaired electron in a σ orbital is zero, one in a π orbital one, and one in a δ orbital two. For more than one electron, the total is $|\Sigma m_l|$. If the total is zero, the state is described as Σ, one as Π, and two as Δ. The multiplicity is indicated by a superscript, e.g., the ground state of NO is $^2\Pi$. A plus or minus sign often follows the symbol to illustrate, respectively, whether the molecular orbitals are symmetric or antisymmetric to a plane through the molecular axis.

If the energies of electronic transitions were related to the ground state molecular orbital energies, the assignment of transitions to the observed bands would be simple. In formaldehyde the $n \rightarrow \pi^*$ would be lower in energy than the $\pi \rightarrow \pi^*$, which in turn would be lower in energy than $\sigma \rightarrow \pi^*$. In addition to different electron-electron repulsions in different states, two other effects complicate the picture by affecting the energies and the degeneracy of the various excited states. These effects are spin-orbit coupling and higher state mixing.

2.3 SPIN-ORBIT COUPLING

There is a magnetic interaction between the electron spin magnetic moment (signified by quantum number $m_s = \pm 1/2$) and the magnetic moment due to the orbital motion of an electron. To understand the nature of this effect, consider the nucleus as though it were moving about the electron (this is equivalent to being on earth and thinking of the sun moving across the sky). We consider the motion from this reference point because we are interested in effects at the electron. The charged nucleus circles the electron, and this is equivalent in effect to placing the electron in the middle of a coil of wire carrying current. As moving charge in a solenoid creates a magnetic field in the center, the orbital motion described above causes a magnetic field at the electron position. This magnetic field can interact with the spin magnetic moment of the electron, giving rise to spin-orbit interaction. The orbital moment may either complement or oppose the spin moment, giving rise to two different energy states. The doubly degenerate energy state of the electron (previously designated by the spin quantum numbers $\pm 1/2$) is split, lowering the energy of one and raising the energy of the other. Whenever an electron can occupy a set of degenerate orbitals that permit circulation about the nucleus, this interaction is possible. For

example, if an electron can occupy the d_{yz} and d_{xz} orbitals of a metal ion, it can circle the nucleus around the z axis.

2.4 CONFIGURATION INTERACTION

As mentioned before, electronic transitions do not occur between empty molecular orbitals of the ground state configuration, but between states. The energies of these states are different from those of configurations derived by placing electrons in the empty orbitals of the ground state, because electron-electron repulsions in the excited state differ from those in the simplified "ground state orbital description" of the excited state. A further complication arises from configuration interaction. One could attempt to account for the different electron-electron repulsions in the excited state by doing a molecular orbital calculation on a molecule having the ground state geometry but with the electron arrangement of the excited configuration. This would not give the correct energy of the state because this configuration can mix with all of the other configurations in the molecule of the same symmetry by configuration interaction.

This mixing is similar in a mathematical sense (though much smaller in magnitude) to the interaction of two hydrogen atoms in forming the H_2 molecule. Thus, we could write a secular determinant to account for the mixing of two B_1 states, B'_{1a} and B'_{1b}, as:

$$\begin{vmatrix} E_1^0 - E & H_{12} \\ H_{12} & E_2^0 - E \end{vmatrix} = 0$$

where H_{12} is the difficult-to-solve integral $\int \psi(B_{1a}) \hat{H} \psi(B_{1b}) d\tau$, whose value is dependent upon the interelectron repulsions in the various states. The closer in energy the initial states E_1^0 and E_2^0, the more mixing occurs. Solution of the secular determinant gives us the two new energies after mixing:

$$E_1 = \frac{1}{2}[E_1^0 + E_2^0 + ((E_1^0)^2 + (E_2^0)^2 - 2E_1^0 E_2^0 + 4H_{12}^2)^{1/2}]$$

$$E_2 = \frac{1}{2}[E_1^0 + E_2^0 - ((E_1^0)^2 + (E_2^0)^2 - 2E_1^0 E_2^0 + 4H_{12}^2)^{1/2}]$$

The energies of the initial (B'_{1a}, B'_{1b}) and final (B_{1a}, B_{1b}) states along with the new wave functions to describe the final state. The wave function is seen to be a linear combination of the two initial molecular orbitals. Interactions of this sort can occur with all the molecular orbitals of B_1 symmetry in the molecule.

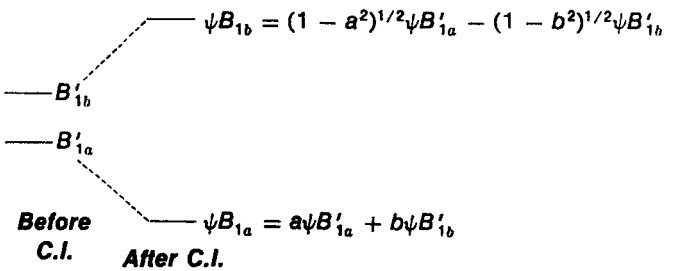


Figure 2.7 Energy levels before and after configuration interaction.

2.5 CRITERIA TO AID IN BAND ASSIGNMENT

An appreciation for some of the difficulties encountered in assigning transitions can be obtained from reading the literature and noting the changes in the assignments that have been made over the years. Accordingly, many independent criteria are used in making the assignment. These include the intensity of the transition and the behaviour of the absorption band when polarized radiation is employed. Both of these topics will be considered in detail shortly. Next we shall describe some simple observations that aid in the assignment of the $n \rightarrow \pi^*$ and $\pi \rightarrow \pi^*$ transitions.

For $n \rightarrow \pi^*$ transitions, one observes the following characteristics:

1. The *molar absorptivity* of the transition is generally less than 2000. An explanation for this is offered in the section on intensity.
2. A *blue shift* (hypsochromic shift, or shift toward shorter wavelengths) is observed for this transition in high dielectric or hydrogen-bonding solvents. This indicates that the energy difference between the ground and excited state is increased in a high dielectric or hydrogen-bonding solvent. In general, for solvent shifts it is often difficult to ascertain whether the excited state is raised in energy or the ground state lowered. A blue shift may result from a greater lowering of the ground state relative to the excited state or a greater elevation of the excited state relative to the ground state. It is thought that the solvent shift in the $n \rightarrow \pi^*$ transition results from a lowering of the energy of the ground state and an elevation of the energy of the excited state. In a high dielectric solvent the molecules arrange themselves about the absorbing solute so that the dipoles are properly oriented for maximum interaction (i.e., solvation that lowers the energy of the ground state). When the excited state is produced, its dipole will have an orientation different from that of the ground state. Since solvent molecules cannot

rearrange to solvate the excited state during the time of a transition, the excited state energy is raised in a high dielectric solvent.

Hydrogen bonding solvents cause pronounced blue shifts. This is reported to be due to hydrogen bonding of the solvent hydrogen with the lone pair of electrons in the n orbital undergoing the transition. In the excited state there is only one electron in the n orbital, the hydrogen bond is weaker and, as a result, the solvent does not lower the energy of this state nearly as much as that of the ground state. In these hydrogen bonding systems an adduct is formed, and this specific solute-solvent interaction is the main cause of the blue shift. If there is more than one lone pair of electrons on the donor, the shift can be accounted for by the inductive effect of hydrogen bonding to one electron pair on the energy of the other pair.

3. The $n \rightarrow \pi^*$ band often disappears in acidic media owing to protonation or upon formation of an adduct that ties up the lone pair, e.g., $BCH_3{}^+I^-$ (as in $C_5H_5NCH_3{}^+I^-$), where B is the base molecule containing the n electrons. This behaviour is very characteristic if there is only one pair of n electrons on B.
4. Blue shifts occur upon the attachment of an electron-donating group to the chromophore represents an increasing blue shift in the carbonyl absorption band). A molecular orbital treatment indicates that this shift results from raising the excited π^* level relative to the n level.
5. The absorption band corresponding to the $n \rightarrow \pi^*$ transition is absent in the hydrocarbon analogue of the compound. This would involve, for example, comparison of the spectra of benzene and pyridine or of $H_2C{=}O$ and $H_2C{=}CH_2$.
6. Usually, but not always, the $n \rightarrow \pi^*$ transition gives rise to the lowest energy singlet-singlet transition.

In contrast to the above behaviour, $\pi \rightarrow \pi^*$ transitions have a high intensity. A slight red (bathochromic) shift is observed in high dielectric solvents and upon introduction of an electron-donating group. It should be emphasized that in the above systems only the difference in energy between the ground and excited states can be measured from the frequency of the transition, so only the relative energies of the two levels can be measured. Other considerations must be invoked to determine the actual change in energy of an individual state.

2.6 INTENSITY OF ELECTRONIC TRANSITIONS

2.6.1 Oscillator Strengths

As described already, the intensity of an absorption band can be indicated by the molar absorptivity, commonly called the extinction coefficient. A parameter of greater theoretical significance is f, the *oscillator strength* of integrated intensity, often simply called the *integrated intensity*:

$$f = 4.315 \times 10^{-9} \int \varepsilon d\bar{\nu} \qquad \ldots(1)$$

In equation (1), ε is the molar absorptivity and $\bar{\nu}$ is the frequency expressed in wave numbers. The concept of oscillator strength is based on a simple classical model for an electronic transition. The derivation indicates that $f = 1$ for a fully allowed transition. The quantity f is evaluated graphically from equation (1) by plotting ε, on a linear scale, versus the wavenumber $\bar{\nu}$ in cm^{-1}, and calculating the area of the band. Values of f from 0.1 to 1 correspond to molar absorptivities in the range from 10,000 to 100,000, depending on the width of the peak.

For a single, symmetrical peak, f can be approximated by the expression:

$$f \approx (4.6 \times 10^{-9}) \varepsilon_{max} \Delta\nu_{1/2} \qquad \ldots(2)$$

where ε_{max} is the molar absorptivity of the peak maximum and $\Delta\nu_{1/2}$ is the half intensity band width, i.e., the width at $\varepsilon_{max}/2$.

2.6.2 Transition Moment Integral

The integrated intensity, f, of an absorption band is related to the transition moment integral as follows:

$$f \propto \left| \int_{-\infty}^{+\infty} \psi_{e1} \hat{M} \psi_{e1}^{ex} d\upsilon \right|^2 = D \qquad \ldots(3)$$

where D is called the dipole strength, ψ_{e1} and ψ_{e1}^{ex} are electronic wave functions for the ground and excited states respectively, $\hat{M}$ is the electric dipole moment operator (*vide infra*), and the entire integral is referred to as the *transition moment integral*. To describe $\hat{M}$, one should recall that the electric dipole moment is defined as the distance between the centers of gravity of the positive and negative charges times the magnitude of these charges. The center of gravity of the positive charges in a molecule is fixed by the nuclei, but the center of gravity of the electrons is an average over the probability function. The vector for the average distance from the nuclei to the

electron is represented as $\bar{r}$. The electric dipole moment vector, $\bar{M}$, is given by $\bar{M} = \Sigma e\bar{r}$, with the summation carried out over all the electrons in the molecule. For the ground state, the electric dipole moment is given by the average over the probability function, or

$$\int \psi_g \sum e\bar{r} \psi_g d\tau$$

By comparison of this equation for the ground state dipole moment with the transition moment integral:

$$\int \psi_g \hat{M} \psi^{ex} d\tau$$

this integral can be seen roughly to represent charge migration or displacement during the transition.

When the integral in equation (3) is zero, the intensity will be zero and, to a first approximation, the transition will be forbidden. In general, we do not have good wave functions for the states to substitute into this equation to calculate the intensity, for reasons discussed in the previous section. However, symmetry can often tell us if integrals are zero, so it is important to examine the symmetry properties of the integrand of equation (3), for they enable us to make some important predictions. These symmetry considerations will also enable us to derive some selection rules for electronic transitions. The quantity $\bar{M}$ is a vector quantity and can be resolved into x, y, and z components. The integral in equation (3) then has the components:

$$\int \psi_{e1} \hat{M}_x \psi_{e1}^{ex} d\upsilon \qquad \ldots(4)$$

$$\int \psi_{e1} \hat{M}_y \psi_{e1}^{ex} d\upsilon \qquad \ldots(5)$$

$$\int \psi_{e1} \hat{M}_z \psi_{e1}^{ex} d\upsilon \qquad \ldots(6)$$

In order to have an allowed transition, at least one of the integrals in equations (4) to (6) must be non-zero. If all three of these integrals are zero, the transition is called forbidden and, according to approximate theory, should not occur at all. Forbidden transitions do occur; more refined theories give small values to the intensity integrals.

Symmetry considerations can tell us if the integral is zero in the following way. An integral can be non-zero only if the direct

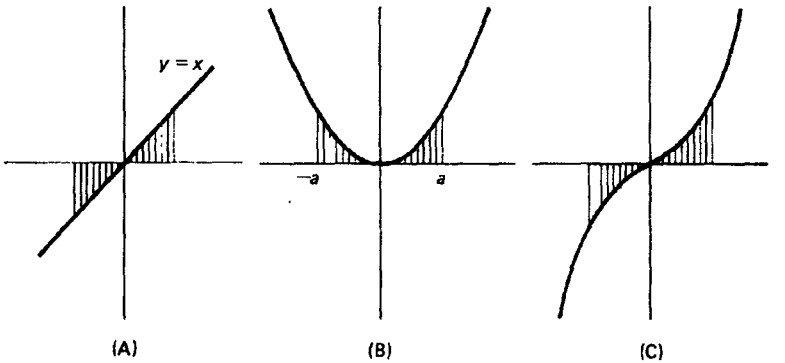


Figure 2.8 Plots of some simple functions, $y = f(x)$, and their symmetries.

product of the integrand belongs to symmetry species A_1. Another way of saying this is that such integrals are different from zero only when the integrand remains unchanged for any of the symmetry operations permitted by the symmetry of the molecule. The reason for the above statements can be seen by looking at some simple mathematical functions. Symmetry considerations tell us that this function does not have A_1 symmetry. The integral $\int y(x)\,dx$ represents the area under the curve. Since $y(x)$ is positive in one quadrant and negative in the other, the area from $+a$ to $-a$ or from $+\infty$ to $-\infty$ is zero. Now, plot a function $y(x)$ such that y is symmetric in x. This function is called "even" and does have A_1 symmetry, because it is unchanged by any of the operations of the symmetry group to which the function belongs. This can be shown by carrying out the symmetry operations of the C_{2v} point. The integral does not vanish. The shaded area gives the value for the integral $\int_{-a}^{a} y dx$ for $y = x^2$, and it is seen not to cancel to zero. Since y is a function of x^2, we can take the direct product of the irreducible representation of x with itself, $\Gamma_x \times \Gamma_x$, and reduce this to demonstrate that this is a totally symmetric irreducible representation. Next, the function $y = x^3$, will be examined. Symmetry considerations tell us that this is an odd function, and we see that $\int y\,dx$ (for $y = x^3$) $= 0$. Does $\Gamma_x \times \Gamma_x \times \Gamma_x = A_1$? The integral over all space of an odd function always vanishes. The integral over all space of an even function *generally* does not vanish.

Thus, to determine whether or not the integral is zero, we take the direct product of the irreducible representations of everything in the integrand. If the direct product is or contains A_1, the integral is non-zero and the transition is allowed.

The application of these ideas is best illustrated by treating an example. Consider the $\pi \rightarrow \pi^*$ transition in formaldehyde. The ground state, like all ground states containing no unpaired electrons, is A_1. The excited state is also A_1 ($b_1 \times b_1 = A_1$). The components $\hat{M}_x$, $\hat{M}_y$, and $\hat{M}_z$ transform as the x, y and z vectors of the point group. The table for the C_{2v} point group indicates that $\hat{M}_z$, the dipole moment vector lying along the z-axis, is A_1. Since $A_1 \times A_1 \times A_1 = A_1$, the integrand $\psi_{e1}\, \hat{M}_z \psi_{e1}^{ex}$ for the $\pi \rightarrow \pi^*$ is A_1 and the $\pi \rightarrow \pi^*$ transition is allowed.

For an $n \rightarrow \pi^*$ transition, the ground state is A_1 and the excited state is A_2. The character table indicates that no dipole moment component has symmetry A_2. Therefore, none of the three integrals [equations (4) to (6)] can be A_1 and the transition is forbidden ($A_2 \times A_2$ is the only product of A_2 that is A_1).

As was mentioned when we introduced this topic, the transition moment integral can be used to derive some important selection rules for electronic transitions.

2.6.3 Derivation of Some Selection Rules

1. For molecules with a center of symmetry, allowed transitions are $g \rightarrow u$ or $u \rightarrow g$. (The abbreviations g and u refer to *gerade* and *ungerade*, which are German for even and odd, respectively.) The d and s orbitals are g, and p orbitals are u. All wave functions in a molecule with a center of symmetry are g or u. All components of the vector $\hat{M}$ in a point group containing an inversion center are necessarily *ungerade*.

$$\Gamma\psi_g \times \Gamma_{op} \times \Gamma\psi_{ex} = \Gamma$$

$$u \times u \times u = u \quad \text{forbidden}$$

$$u \times u \times g = g \quad \text{allowed}$$

$$g \times u \times g = u \quad \text{forbidden}$$

$$g \times u \times u = g \quad \text{allowed}$$

This leads to the selection rule that $g \rightarrow u$ and $u \rightarrow g$ are allowed, but $g \rightarrow g$ and $u \rightarrow u$ are forbidden. Therefore, $d \rightarrow d$ transitions in transition metal complexes with a center of symmetry are forbidden. Values of e for the d-d transitions in $Ni(H_2O)_6^{2+}$ are ~20.

2. Transitions between states of different multiplicity are forbidden. Consider a singlet $\rightarrow$ triplet transition. Focusing on the electron being excited, we have in the singlet ground state $\psi\alpha\psi\beta$ and, in

the excited state, $\psi\alpha\psi\alpha$ or $\psi\beta\psi\beta$, where α and β are the spin coordinates. The dipole strength is given by

$$D = \left| \int \psi_i \alpha \hat{M} \psi_f \beta \, d\tau \, d\sigma \right|^2$$

(where $d\sigma$ is the volume element in the spin coordinates and the i and f subscripts refer to initial and final states). We can rewrite the integral corresponding to D as

$$\left| \int \psi_i \hat{M} \psi_f d\tau \int \alpha\beta \, d\sigma \right|^2$$

Since the second term is the product of $+1/2$ and $-1/2$ spins, it is always odd and zero, i.e., the spins are orthogonal. Since $\int\alpha\alpha \, d\sigma = 1$ and $\int\beta\beta \, d\sigma = 1$, in working out the intensity integral we only have to worry about the electron that is undergoing the transition, and we can ignore all the electrons in the molecule that do not change spin. The ε for absorption bands involving transitions between states of different multiplicity is generally less than one.

3. Transitions in molecules without a center of symmetry depend upon the symmetries of the initial and final states. If the direct product of these and anyone of $\hat{M}_x$, $\hat{M}_y$ or $\hat{M}_z$ is A_1 the transition is allowed. If all integrals are odd, the transition is forbidden.

2.6.4 Spectrum of Formaldehyde

We can summarize the above ideas and illustrate their utility by returning again to the ultraviolet spectrum of formaldehyde. The various possible excited states arising from electron excitations from the highest-energy filled orbitals (n_a, n_b, and π) are given by:

$$a_1^2 b_1^2 b_2^1 b_1^{1*} = {}^1A_2 \qquad (n_b \rightarrow \pi^*)$$
$$a_1^2 b_1^2 b_2^1 a_1^{1*} = {}^1B_2 \qquad (n_b \rightarrow \sigma^*)$$
$$a_1^2 b_1^2 b_2^2 b_1^{1*} = {}^1A_1 \qquad (\pi \rightarrow \pi^*)$$
$$a_1^2 b_1^1 b_2^2 a_1^{1*} = {}^1B_1 \qquad (\pi \rightarrow \sigma^*)$$
$$a_1^1 b_1^2 b_2^2 b_1^{1*} = {}^1B_1 \qquad (n_a \rightarrow \pi^*)$$
$$a_1^1 b_1^2 b_1^2 a_1^{1*} = {}^1A_1 \qquad (n_a \rightarrow \sigma^*)$$

Two bands are observed, one with $\varepsilon = 100$ at 2700 Å and an extremely intense one at 1850 Å. We see that the lowest-energy transitions are $n_b \rightarrow \pi^*$ and $\pi \rightarrow \pi^*$, better expressed as ${}^1A_1 \rightarrow {}^1A_2$ and ${}^1A_1 \rightarrow {}^1A_1$. The ${}^1A_1 \rightarrow {}^1A_2(n_b \rightarrow \pi^*)$ is forbidden, and

accordingly is assigned to the band at 2700 Å. Both $^1A_1 \rightarrow {}^1B_1(n_a \rightarrow \pi^*)$ and $^1A_1 \rightarrow {}^1A_1(\pi \rightarrow \pi^*)$ are allowed. The former may contribute to the observed band at 1850 Å or may be in the far uv.

The integrands for $^1A_1 \rightarrow {}^1B_1(\pi \rightarrow \sigma^*)$, $(^1A_1 \rightarrow {}^1B_2(n_b \rightarrow \sigma^*)$, and $^1A_1 \rightarrow {}^1A_1(n_a \rightarrow \sigma^*)$ are all A_1, leading to allowed transitions. These are expected to occur at very short wavelengths in the far ultraviolet region. This is the presently held view of the assignment of this spectrum, and it can be seen that the arguments are not rigorous. We shall subsequently show how polarization studies aid in making assignments more rigorous.

Next, it is informative to discuss the uv spectrum of acetaldehyde, which is quite similar to that of formaldehyde. The $n_b \rightarrow \pi^*$ transition has very low intensity. However, acetaldehyde has C_s symmetry; this point group has only two irreducible representations, A and B, with the x and y vectors transforming as A and the z vector as B. Accordingly, *all* transitions will have an integrand with A_1 symmetry and will be allowed. Though the $n_b \rightarrow \pi^*$ transition is allowed by symmetry, the value of the transition moment integral is very small and the intensity is low. The intensity of this band in acetaldehyde is greater than that in formaldehyde. We can well appreciate the fact that although monodeuteroformaldehyde [DC(O)H] does not have C_{2v} symmetry, it will have an electronic spectrum practically identical to that of formaldehyde. These are examples of a rather general type of result, which leads to the idea of *local symmetry*. According to this concept, even though a molecule does not have the symmetry of a particular point group, if the groups attached to the chromophore have similar bonding interactions, the molecule for many purposes can be treated as though it had this higher symmetry.

2.6.5 Spin-orbit and Vibronic Coupling Contributions to Intensity

The discrepancy between the theoretical prediction that a transition is forbidden and the experimental detection of a weak band assignable to this transition is attributable to the approximations of the theory. More refined calculations that include effects from *spin-orbit coupling* often predict low intensities for otherwise forbidden transitions. For example, a transition between a pure singlet state and a pure triplet state is forbidden. However, if spin-orbit coupling is present, the singlet could have the same total angular momentum as the triplet and the two states could interact. The interaction is indicated by equation (7):

$$\psi = a\,{}^1\psi + b\,{}^3\psi \qquad ...(7)$$

where ${}^1\psi$ and ${}^3\psi$ correspond to pure singlet and triplet states, respectively, ψ represents the actual ground state, and a and b are coefficients indicating the relative contributions of the pure states. If $a > b$, the ground state is essentially singlet with a slight amount of triplet character and the excited state will be essentially triplet. This slight amount of singlet character in the predominantly triplet excited state leads to an intensity integral for the singlet-triplet transition that is not zero; this explains why a weak peak corresponding to a multiplicity-forbidden transition can occur.

Another phenomenon that gives intensity to some forbidden transitions is *vibronic coupling*. We have assumed until now that the wave function for a molecule can be factored into an electronic part and a vibrational part, and we have ignored the vibrational part. When we applied symmetry considerations to our molecule, we assumed some symmetrical, equilibrium internuclear configuration. This is not correct, for the molecules in our system are undergoing vibrations and during certain vibrations the molecular symmetry changes considerably. For example, in an octahedral complex, the T_{1u} and T_{2u} vibrations remove the center of symmetry of the molecule. Since electronic transitions occur much more rapidly than molecular vibrations, we detect transitions occurring in our sample from many geometries that do not have high symmetry, e.g., the vibrationally distorted molecules of the octahedral complex. The local symmetry is still very close to octahedral, so the intensity gained this way is not very great; but it is large enough to allow a forbidden transition to occur with weak intensity.

The electronic transition can become allowed by certain vibrational modes but not by all. We can understand this by rewriting

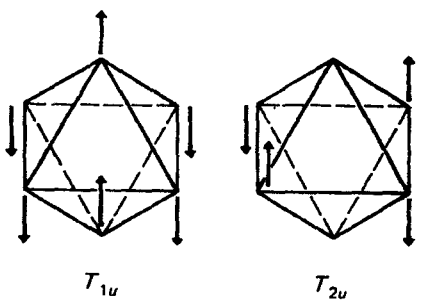


Figure 2.9 T_{1u} and T_{2u} vibrations of an octahedral complex.

the transition moment integral to include both the electronic and the vibrational components of the wave function as in equation (8):

$$f \propto D = \left| \int \psi_{el}^{*} \psi_{vib}^{*} \hat{M} \psi_{el}^{ex} \psi_{vib}^{ex} d\tau \right|^2 \qquad ...(8)$$

As we mentioned already, all ground vibrational wave functions are A_1 so the symmetry of $\psi_{el}\psi_{vib}$ becomes that of ψ_{el} which is also A_1 for molecules with no unpaired electrons. (In general discussion, we shall use the symbol A_1 to represent the totally symmetric irreducible representation, even though this is not the appropriate label in some point groups.) To use this equation to see whether a forbidden transition can gain intensity by vibronic coupling, we must take a product $\hat{M}_{(x,\ y,\ \text{or}\ z)}\psi_{el}^{ex}$ that is not A_1 and see whether there is a vibrational mode with symmetry that makes the product $\hat{M}_{(x,\ y,\ \text{or}\ z)}\psi_{el}^{ex}\psi_{vib}^{ex}$ equal to A_1. When ψ_{vib}^{ex} has the same symmetry as the product $\hat{M}_{(x,\ y,\ \text{or}\ z)}\psi_{el}^{ex}$ the product will be A_1.

This discussion can be made clearer by considering some examples. We shall consider vibrational spectroscopy in more detail in the next chapter. A non-linear molecule has $3N - 6$ internal vibrations; for formaldehyde these are $3a_1$ b_1, and $2b_2$. For the forbidden transition $^1A_1 \rightarrow {}^1A_2(n_b \rightarrow \pi^*)$, the vibrational wave function of a_1 symmetry does not change the direct product $\hat{M}_{(x,\ y,\ \text{or}\ z)}{}^1A_2$ so no intensity can be gained by this mode. Excitation of the b_1 vibrational mode leads to a direct product $\psi_{el}^{ex}\psi_{vib}^{ex}$ of $b_1 \times A_2 = B_2$. Since $\hat{M}_y$ has B_2 symmetry, the total integral $(\int \psi_{el}\psi_{vib}\hat{M}_y\psi_{el}^{ex}\psi_{vib}^{ex}\, d\tau)$ has A_1 symmetry, and the electronic transition becomes allowed by vibronic coupling to the b_1 mode.

It is informative to consider $Co(NH_3)_6^{3+}$ as an example, for it contains triply degenerate irreducible representations. The ground state is $^1A_{1g}$ (a strong field O_h d_6 complex). The excited states from d-d transitions are $^1T_{1g}$ and $^1T_{2g}$. $\hat{M}_x$, $\hat{M}_y$, and $\hat{M}_z$ transform as T_{1u}. For the $^1A_{1g} \rightarrow {}^1T_{1g}$ transition one obtains:

$$A_{1g} \times T_{1u} \times T_{1g}$$

The resulting direct product representation has a dimensionality of nine (the identity is $1 \times 3 \times 3 = 9$) and the total representation is reduced into a linear combination of $A_{1u} + E_u + T_{1u} + T_{2u}$ irreducible representations. With no A_{1g} component, the $^1A_{1g} \rightarrow {}^1T_{1g}$ transition is forbidden. However, the vibrations for an octahedral complex have the symmetries of a_{1g}, e_g, $2t_{1u}$, t_{2g}, t_{2u}. Since the direct products $t_{1u} \times T_{1u}$ and $t_{2u} \times T_{2u}$ have A_{1g} components, this transition becomes allowed by vibronic coupling.

2.6.6 Mixing of d and p Orbitals in Certain Symmetries

There is one further aspect of the intensity of electronic transitions that can be understood *via* the symmetry aspects of electronic transitions. The electronic spectra of tetrahedral complexes of cobalt(II) contain two bands assigned to *d-d* transitions at ~20,000 cm^{-1} and ~6000 cm^{-1}, assigned as $A_2 \rightarrow T_1$ and $A_2 \rightarrow T_2$ transitions respectively, with molar absorptivities of 600 and 50. Since the $\hat{M}$ components transform as T_2, we obtain for the $A_2 \rightarrow T_1$ transition

$$A_2 \times T_2 \times T_1 = A_1 + E + T_1 + T_2$$

so the transition is allowed. However, if only the *d*-orbitals were involved in this transition, the intensity would be zero for the integrals

$$\int \psi_{d_{xy}} \hat{M} \psi_{d_{xz}} d\tau = 0$$

However, in the T_d point group, the d_{xy}, d_{xz}, and d_{yz} orbitals and the *p*-orbitals transform as T_2 and therefore can mix. If the two states involved in the transition, A_2 and T_1, have differing amounts of *p*-character, intensity is gained by having some of the highly allowed $p \rightarrow d$ or $d \rightarrow p$ character associated with the transition.

Consider the consequences of this mixing on the $A_2 \rightarrow T_2$ transition. The transition moment integrand for this transition is

$$A_2 \times T_2 \times T_2$$

which, as the reader should verify, can be reduced to $A_2 + E + T_1 + T_2$. Since there is no A_1 component, the transition is forbidden. Mixing *p*-character into the wave functions will not help, for this type of transition is still forbidden. Accordingly, the ε for the $A_2 \rightarrow T_1$ transition is ten times greater than that of $A_2 \rightarrow T_2$. The latter transition gains most of its intensity by vibronic coupling.

2.6.7 Magnetic Dipole and Electric Quadrupole Contributions to the Intensity

So far, our discussion of the intensity of electronic transitions has centered on the electric dipole component of the radiation, for we concerned ourselves with the transition moment integral with an electric dipole operator, $e\vec{r}$. There is also a magnetic dipole component. The magnetic dipole operator transforms as a rotation $R_x R_y R_z$, and the intensity from this effect may be regarded as arising from the rotation of electron density. Transition moment integrals similar to those for electric dipole transitions can be written for the contribution from both magnetic dipole and electric quadrupole effects. In a molecule with a center of symmetry, both of these

operators are symmetric with respect to inversion, so $g \rightarrow g$ and $u \rightarrow u$ transitions are allowed. Approximate values of the transition moment integral for allowed transitions for these different operators are: 6×10^{-36} cgs units for an electric dipole transition, 9×10^{-41} cgs units for a magnetic dipole transition, and 7×10^{-43} cgs units for a quadrupole transition. Thus, we can see that these latter two effects will be important only when electric dipole transitions are forbidden. They do complicate the assignment of very weak bands in the spectrum.

2.6.8 Charge Transfer Transitions

A transition in which an electron is transferred from one atom or group in the molecule to another is called a *charge-transfer* transition. More accurately stated, the transition occurs between molecular orbitals that are essentially centered on different atoms. Very intense bands result, with molar absorptivities of 10^4 or greater. The frequency at maximum absorbancy, ν_{max}, often, but not always, occurs in the ultraviolet region. The anions ClO_4^- and SO_4^{2-} show very intense bands. Since MnO_4^- and CrO_4^{2-} have no *d* electrons, the intense colours of these ions cannot be explained on the basis of *d-d* transitions; they are attributed to charge transfer transitions. The transitions in MnO_4^- and CrO_4^{2-} are most simply visualized as an electron transfer from a non bonding orbital of an oxygen atom to the manganese or chromium ($n \rightarrow \pi^*$), in effect reducing these metals in the excited state. An alternate description for this transition involves excitation of an electron from a π bonding molecular orbital, consisting essentially of oxygen atomic orbitals, to a molecular orbital that is essentially the metal atomic orbital.

In the case of a pyridine complex of iridium(III), a charge transfer transition that involves oxidation of the metal has been reported. A metal electron is transferred from an orbital that is essentially an iridium atomic orbital to an empty π^* antibonding orbital in pyridine.

In gaseous sodium chloride, a charge transfer absorption occurs from the ion pair Na^+Cl^- to an excited state described as sodium and chlorine atoms having the same internuclear distance as the ion pair. A charge transfer absorption also occurs in the ion pair, *N*-methylpyridinium iodide in which an electron is transferred from I^- to a ring anti bonding orbital. A very intense charge transfer absorption is observed in addition compounds formed between iodine and several Lewis bases.

2.6.9 Polarized Absorption Spectra

If the incident radiation employed in an absorption experiment is polarized, only those transitions with similarly oriented dipole moment vectors will occur. In a powder, the molecules or complex ions are randomly oriented. All allowed transitions will be observed, for there will be a statistical distribution of crystals with dipole moment vectors aligned with the polarized radiation. However, suppose, for example, that a formaldehyde crystal, with all molecules arranged so that their z-axes are parallel, is examined. As indicated in the previous section, the integrand $\psi^* \hat{M}_z \psi$ has appropriate symmetry for the ${}^1A_{(\pi,\pi^*)} \leftarrow {}^1A$ transition, but $\psi^* \hat{M}_x \psi$ and $\psi^* \hat{M}_y \psi$ do not. When the z-axes of the molecules in the crystal are aligned parallel to light that has its electric vector polarized in the z-direction, light will be absorbed for the ${}^1A_{(\pi,\pi^*)} \leftarrow {}^1A$ transition. Light of this wavelength polarized in other planes will not be absorbed. If this crystal is rotated so that the z-axis is perpendicular to the plane of polarization of the light, no light is absorbed. This behaviour supports the assignment of this band to the transition ${}^1A_1 \leftarrow {}^1A_1$. To determine the expected polarization of any band, the symmetry species of the product $\psi_a \psi_b$ is compared with the components of $\hat{M}$, as was done above for formaldehyde.

Absorption of radiation will occur if $\hat{M}_z$ results in an A_1 transition moment integrand for equation (6). No absorption will occur if it is not A_1 regardless of the symmetries of the integrand for the $\hat{M}_x$ or $\hat{M}_y$ components [i.e., equations (4) and (5)]. The absorption will occur if the $\hat{M}_y$ component gives an A_1 transition moment integral.

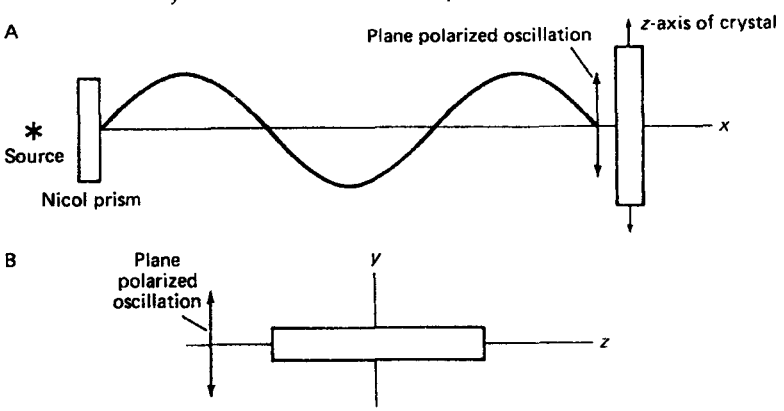


Figure 2.10 Schematic illustration of a polarized single crystal study.

Even if $\hat{M}_z$ has an integrand with A_1 symmetry, no absorption of the z-component will occur for this orientation and no absorption at all will occur if the $\hat{M}_y$ integrand is not A_1.

We can further illustrate these ideas by considering the electronic absorption spectrum of $PtCl_4^{2-}$. The transitions are charge transfers involving electron excitation from a mainly chlorine m.o. to the empty $d_{x^2-y^2}$ orbital on Pt(II). The symmetry is D_{4h}; using the same approach as on the NO_2^- ion, on a basis set of four p_z orbitals on chlorine, we obtain symmetry orbitals for chlorine of b_{2u}, e_u, and a_{2u} symmetry. This leads to the following possible charge transfer transitions:

$b_{2u}(\pi) \rightarrow b_{1g}(d_{x^2-y^2})$ with state labels $^1A_{1g} \rightarrow {}^1A_{2u}$ (here A_{2u} is the direct product of $b_{2u} \times b_{1g}$)

$e_u(\pi) \rightarrow b_{1g}(d_{x^2-y^2})$ with state labels $^1A_{1g} \rightarrow {}^1E_u$

$a_{2u}(\pi) \rightarrow b_{1g}(d_{x^2-y^2})$ with state labels $^1A_{1g} \rightarrow {}^1B_{2u}$

In the D_{4h} point group, $\hat{M}_x$ and $\hat{M}_y$ transform as E_u, and $\hat{M}_z$ as A_{2u}. If we first consider the $A_{1g} \rightarrow A_{2u}$ transition, we get for $\hat{M}_z$:

$$A_{1g}A_{2u}A_{2u} = A_{1g}$$

and for $\hat{M}_x$ and $\hat{M}_y$ we get:

$$A_{1g}E_uA_{2u} \neq A_{1g}$$

Accordingly, this transition is allowed and is polarized in the z-direction. If we use polarized light and a single crystal, light will be absorbed when the z-axis of the crystal is parallel to the z-direction of the light; but there will be no absorption when the z-axis is perpendicular to the light because the $\hat{M}_x$ and $\hat{M}_y$ integrands are not A_1.

For the band assigned to $^1A_{1g} \rightarrow {}^1E_u$, the $A_{1g}E_uE_u$ product has an A_{1g} component, so this transition is also allowed. Since $\hat{M}_z$ yields $A_{1g}A_{2u}E_u$, which does not have an A_{1g} component, there will be no absorption when the z-component is parallel to the plane of the polarized light but absorption will occur when the x- and y-axes of the crystal are parallel to the light.

The $A_{1g} \rightarrow B_{2u}$ transition turns out to be forbidden. Thus, we see that by employing polarized single crystal spectroscopy, we can rigorously assign the two intense charge transfer bands observed in the electronic spectrum of $PtCl_4^{2-}$. If the single crystal employed in these experiments did not have all of the molecular z-axes aligned, the polarization experiments would not work.

2.7 APPLICATIONS

Most applications of electronic spectroscopy have been made in the wavelength range from 2100 to 7500 Å, for this is the range accessible with most recording spectrophotometers. Relatively inexpensive commercial instruments can now be obtained to cover the range from 1900 to 8000 Å. The near infrared region, from 8000 to 25,000 Å, has also provided much useful information. Spectra can be examined through the 1900 to 25,000 Å region on samples of vapours, pure liquids, or solutions. Solids can be examined as single crystals or as discs formed by mixing the material with KCl or NaCl and pressing with a hydraulic press until a clear disk is formed. Spectra of powdered solids can also be examined over a more limited region (4000 to 25,000 Å) as reflectance spectra or on mulls of the solid compounds.

2.7.1 Fingerprinting

Since many different substances have very similar ultraviolet and visible spectra, his is a poor region for product identification by the "fingerprinting" technique. Information obtained from this region should be used in conjunction with other evidence to confirm the identity of the compound. Evidence for the presence of functional groups can be obtained by comparison of the spectra with reported data. For this purpose, ν_{max}, ε_{max}, and band shapes can be employed. It is also important that the spectra be examined in a variety of solvents to be sure that the band shifts are in accord with expectations.

Spectral data have been compiled by Sadtler, Lang, and Hershenson, and in "Organic Electronic Spectral Data" and the ASTM Coded IBM Cards. A review article by Mason and the text by Jaffe and Orchin are excellent for this type of application. It a functional group (chromophore) is involved in conjugation or steric interactions, or is attached to electron-releasing groups, its spectral properties are often different from those of an isolated functional group. These differences can often be predicted semiquantitatively for molecules in which such effects are expected to exist.

The spectra of some representative compounds and examples of the effect of substituents on the wavelength of a transition will be described briefly.

2.7.1.1 Saturated molecules

Saturated molecules without lone pair electrons undergo high-energy $\sigma \rightarrow \sigma^*$ transitions in the far ultraviolet. For example, methane

has a maximum at 1219 Å and ethane at 1350 Å corresponding to this transition. When lone pair electrons are available, a lower-energy $n \rightarrow \sigma^*$ transition is often detected in addition to the $\sigma \rightarrow \sigma^*$. For example, in triethylamine two transitions are observed, at 2273 and 1990 Å.

2.7.1.2 Carbonyl compounds

The carbonyl chromophore has been very extensively studied. Upon conjugation of the carbonyl group with a vinyl group, four π energy levels are formed. The highest occupied π level has a higher energy, and one of the lowest empty π^* levels has a lower energy, than the corresponding levels in a nonconjugated carbonyl group. The lone pair and σ electrons are relatively unaffected by conjugation. As a result, the $\pi \rightarrow \pi^*$ and $n \rightarrow \pi^*$ transition energies are lowered and the absorption maxima are shifted to longer wavelengths when the carbonyl is conjugated. The difference is greater for the $\pi \rightarrow \pi^*$ than for the $n \rightarrow \pi^*$ transition. The $n \rightarrow \sigma^*$ band is not affected appreciably and often lies beneath the shifted $\pi \rightarrow \pi^*$ absorption band. As stated earlier, electron-donating groups attached to the carbonyl cause a blue shift in the $n \rightarrow \pi^*$ transition and a red shift in $\pi \rightarrow \pi^*$.

It is of interest to compare the spectra of thiocarbonyl compounds with those of carbonyl compounds. In the sulfur compounds, the carbon-sulfur π interaction is weaker and, as a result, the energy difference between the π and π^* orbitals is smaller than in the oxygen compounds. In addition, the ionization potential of the sulfur electrons in the thiocarbonyl group is less than the ionization potential of oxygen electrons in a carbonyl. The n electrons are of higher energy in the thiocarbonyl and the $n \rightarrow \pi^*$ transition requires less energy in these compounds than in carbonyls. The absorption maximum in thiocarbonyls occurs at longer wavelengths and in some compounds is shifted into the visible region.

2.7.1.3 Inorganic systems

The SO_2 molecule has two absorption bands in the near ultraviolet at 3600 Å ($\varepsilon = 0.05$) and 2900 Å ($\varepsilon = 340$) corresponding to a triplet and singlet $n \rightarrow \pi^*$ transition. The gaseous spectrum shows considerable vibrational fine structure, and analysis has produced information concerning the structure of the excited state. In nitroso compounds, an $n \rightarrow \pi^*$ transition involving the lone pair electrons on the nitrogen occurs in the visible region. An

$n \rightarrow \pi^*$ transition involving an oxygen lone pair occurs in the ultraviolet.

The nitrite ion in water has two main absorption bands at 3546 Å ($\varepsilon = 23$) and 2100 Å ($\varepsilon = 5380$) and a weak band at 2870 Å ($\varepsilon = 9$). The band at 3546 Å is an $n \rightarrow \pi^*$ transition ($^1B_1 \leftarrow {}^1A_1$) involving the oxygen lone pair. The band at 2100 Å is assigned as $\pi \rightarrow \pi^*(^1B_2 \leftarrow {}^1A_1)$, and the band at 2870 Å is assigned to an $n \rightarrow \pi^*$ transition ($^1A_2 \leftarrow {}^1A_1$) involving the oxygen lone pair. For the simple ions (Br^-, Cl^-, OH^-) the absorption is attributed to charge transfer in which the electron is transferred to the solvent.

2.7.2 Molecular Addition Compounds of Iodine

The absorption band maximum for iodine (core plus $\sigma^2\pi^4n^4\pi^{*4}$) occurs at about 5200 Å in the solvent CCl_4 and is assigned to a $\pi^* \rightarrow \sigma^*$ transition. When a donor molecule is added to the above solution, two pronounced changes in the spectrum occur.

A blue shift is detected in the iodine peak, and a new peak arises in the ultraviolet region that is due to a charge transfer transition. The existence of an isosbestic point at 490 mμ indicates that there are only two absorbing species in the system; namely, free iodine and the complex B: $\underline{I} - \underline{I}|$. A 1:1 equilibrium constant can be calculated from absorbance measurements for this system. The constant value for K obtained over a wide range of donor concentrations is evidence for the existence of a 1:1 addition compound.

The bonding in iodine addition compounds can be described by the following equation:

$$\psi^0 = a\psi_{cov} + b\psi_{el}$$

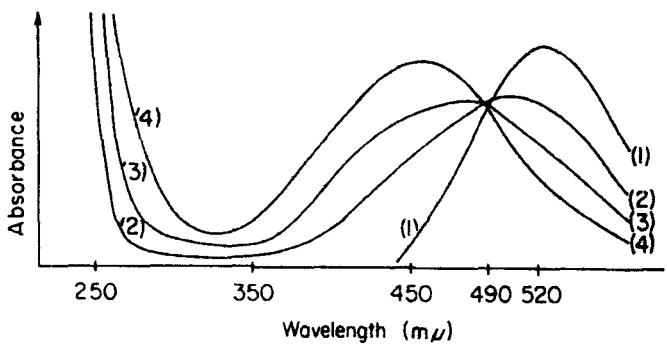


Figure 2.11 Spectra of iodine and base-iodine solutions.

where ψ_{el} includes contributions from purely electrostatic forces while ψ_{cov} includes contributions from covalent interactions (these are described as charge transfer interactions). In many of these complexes, $b > a$ in the ground state. In these cases, the band around 2500 Å arises from a charge transfer transition in which an electron from this ground state is promoted to an excited state in which $a > b$. In view of these coefficients, the charge transfer band assignment can be approximated by a transfer of a base electron, n_b, to the iodine σ^* orbital. These facts and the blue shift that occurs in the normal $\pi^* \rightarrow \sigma^*$ iodine transition upon complexation can be explained by consideration of the relative energies of the molecular orbitals of iodine and the complex. In Figure, n_b refers to the donor orbital on the base, and $\sigma_{I_2}^*$ and $\pi_{I_2}^*$ refer to the free iodine antibonding orbitals involved in the transition leading to iodine absorption. The σ_c, π_c^* and σ_c^* orbitals are molecular orbitals in the complex that are very much like the original base and iodine orbitals because of the weak Lewis acid-base interaction (2 to 10 kcal). The orbitals n_b and $\sigma_{I_2}^*$ combine to form molecular orbitals in the complex, σ_c and σ_c^*, in which σ_c, the bonding orbital, is essentially n_b and σ_c^* is essentially $\sigma_{I_2}^*$. Since σ_c^* is slightly higher in energy than the corresponding $\sigma_{I_2}^*$, the transition in complexed iodine requires slightly more energy than the corresponding transition in free I_2 and a blue shift is observed. The charge-transfer transition occurs at higher energy in the ultraviolet region and is designated by arrow. Some interesting correlations have been reported, which claim that the blue shift is related to the magnitude of the base-iodine interaction, i.e., the enthalpy of adduct formation. This would be expected qualitatively from the treatment as long as the energy of π_c^* differs very little from that of $\pi_{I_2}^*$ or else its energy changes in a linear manner with the enthalpy, ΔH. A rigorous evaluation of

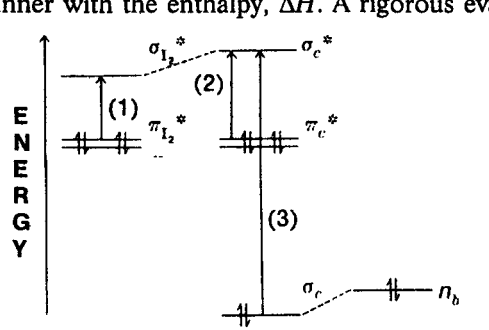


Figure 2.12 Some of the molecular orbitals in a base-iodine addition compound.

this correlation with accurate data on a wide range of different types of Lewis bases indicates that a rough general trend exists, but that a quantitative relation (as good as the accuracy of the data) does not exist. A relationship involving the charge transfer band, the ionization potential of the base, I_b, and the electron affinity of the acid, E_a, is also reported:

$$\nu = I_b - E_a - \Delta \qquad ...(9)$$

where Δ is an empirically determined constant for a related series of bases.

The enthalpies for the formation of these charge transfer complexes are of interest and significance to both inorganic and organic chemists. For many inorganic systems, especially in the areas of coordination chemistry and nonaqueous solvents, information about donor and acceptor interactions is essential to an understanding of many phenomena (for example, catalysis and metalloproteins). Since the above adducts are soluble in CCl_4 or hexane, the thermodynamic data can be interpreted more readily than results obtained in polar solvents, where large solvation enthalpies and entropies are encountered. As a result of these solvation effects, structural interpretations of the effect of substituents on pK_b values of bases and on stability constant data for various ligands and metal ions are often highly questionable. Some typical results from donor-I_2 systems in which such solvation effects are minimal illustrate the wide range of systems. The following few examples illustrate the information that can be obtained by studying enthalpies of association in non-polar, weakly basic solvents.

1. The donor properties of the p electron systems of alkyl substituted benzenes have been reported.
2. A correlation of the heat of formation of iodine adducts of a series of *para*-substituted benzamides with the Hammett substituent constants of the benzamides is reported.
3. The donor properties of a series of carbonyl compounds [$(CH_3)_2CO$, $CH_3C(O)N(CH_3)_2$, $(CH_3)_2NC(O)N(CH_3)_2$, $CH_3C(O)OCH_3$, $CH_3C(O)SCH_3$] have been evaluated and interpreted in terms of conjugative and inductive effects of the group attached to the carbonyl functional group.
4. The donor properties of sulfoxides, sulfones, and sulfites have been investigated. The results are interpreted to indicate that sulfur-oxygen π bonding is less effective in these systems than carbon-oxygen π bonding is in ketones and acetates.

5. The effect of ring size on the donor properties of cyclic ethers and sulfides has been investigated. It was found that for saturated cyclic sulfides, of general formula $(CH_2)_nS$, the donor properties of sulfur are in the order $n = 5 > 6 > 4 > 3$. The order for the analogous ether compounds is $4 > 5 > 6 > 3$. Explanations of these effects are offered.
6. The donor properties of a series of primary, secondary, and tertiary amines have been evaluated. The order of donor strength of amines varies with the acid studied. Explanations have been offered, which are based upon the relative importance of covalent and electrostatic contributions to the bonding in various adducts.

In addition to iodine, several other Lewis acids form charge transfer complexes that absorb in the ultraviolet or visible regions. For example, the relative acidities of I_2, ICl, Br_2, SO_2, and phenol toward the donor *N,N*-dimethylacetamide have been evaluated. Factors affecting the magnitude of the interaction and information regarding the bonding in the adducts are reported.

2.7.3 Effect of Solvent Polarity on Charge Transfer Spectra

The ion pair *N*-methylpyridinium iodide undergoes a charge-transfer transition. It has been found that the position of the charge-transfer band is a function of the solvating ability of the solvent. A shift to lower wavelengths is detected in the better solvating solvents. The positions of the bands are reported as transition energies, E_T. Transition energies (kcal mole^{-1}) are calculated from the frequency. The transition energy is referred to as the Z value. An explanation for the observed shift has been proposed. The dipole moment of the ion pair, $C_5H_5NCH_3{}^+I^-$, is reported to be perpendicular to the dipole moment of the excited state. Polar solvent molecules will align their dipole moments for maximum interaction with the ground state, lowering the energy of the ground state by solvation. The dipole moment of the solvent molecules will be perpendicular to the dipole moment of the excited state, producing a higher energy for the excited state than would be found in the gas phase. Since solvent molecules cannot rearrange in the time required for a transition, the 'relative lowering of the ground state and raising of the excited state increases the energy of the transition, E_T, over that in the gas phase, shifting the wavelength of absorption to higher frequencies. Hydrogen-bonding solvents are often found to increase E_T more than would be expected by comparing their dielectric constants with those of other solvents. This is due to the formation of hydrogen

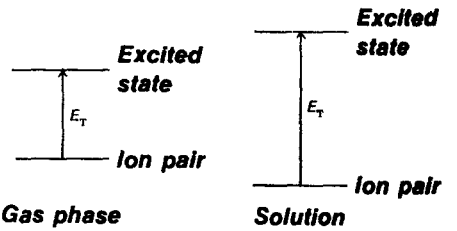


Figure 2.13 Effect of solvent on the transition energy, E_T.

bonds with the solute. The use of the dielectric constant to infer solvating ability can lead to difficulty because the local dielectric constant in the vicinity of the ion may be very different from the bulk dielectric constant.

The data obtained from these spectral shifts are employed as an empirical measure of the ionizing power of the solvent. The results can be correlated with a scale of "solvent polarities" determined from the effect of solvent on the rate of solvolysis of *t*-butyl chloride. Other applications of these data to kinetic and spectral studies are reported. Solvent effects are quite complicated, and these correlations at best provide a semiquantitative indication of the trends expected.

Significant differences exist between "solvating power" inferred from the dielectric constant and the results from spectral and kinetic parameters. Although methanol and formamide are found to have similar Z values, the dielectric constants are 32.6 and 109.5, respectively. Solvent effects cannot be understood solely on the basis of the dielectric constant. The extent of ion-pair association of the pyridinium salts in various solvents can be approximated from the apparent molar absorptivities of the charge-transfer absorption, because the dissociated ion pair is not expected to contribute to the charge transfer absorption. The tendencies of ion pairs to dissociate in solvents estimated in this way do not correlate with transition energies. The dissociating tendency is expected to be more closely related to the dielectric constant ($F = q_1q_2/Dr^2$, where F is the force between two ions with charges q_1 and q_2 separated by a distance r, and D is the dielectric constant). Specific Lewis acid-base interactions make the problem more complex than the simple dielectric model.

The band positions for the $n \rightarrow \pi^*$ transitions in certain ketones in various solvents are found to be linearly related to the Z values for the solvents. A constant slope is obtained for a plot of E_T versus

Z for many ketones. Deviations from linearity by certain ketones in this plot can be employed to provide interesting structural information about the molecular conformation. Cycloheptanone, for example, does not give a linear plot of E_T versus Z. The deviation is attributed to solvent effects on the relative proportion of the conformers present in solution.

2.7.4 Structures of Excited States

Considerable information is available about the structure of excited states of molecules from analysis of the rotational band contours in the electronic spectra. Both geometrical information and vibrational information about the excited states of large molecules can be obtained. By studying and analyzing the perturbation made on the vibrational fine structure of an electronic transition by an electric field, the dipole moment of the excited state can be obtained.

2.8 OPTICAL ROTATORY DISPERSION, CIRCULAR DICHROISM AND MAGNETOCIRCULAR DICHROISM

Plane polarized light consists of two circularly polarized components of equal intensity. The two types of circularly polarized light correspond to right-handed and left-handed springs. Circularly polarized light is defined as right-handed when its electric or magnetic vector rotates clockwise as viewed by an observer facing the direction of the light propagation (i.e., the source). The frequency of the rotation is related to the frequency of the light. Plane polarized light can be resolved into its two circular components, and the two components when added together produce plane polarized light in an optically isotropic medium. If plane polarized light is passed through a sample for which the refractive indices of the left and right polarized components differ, the components will, upon recombination, give plane polarized radiation in which the plane of the polarization has been rotated through an angle α, given by

$$\alpha = \frac{n_\ell - n_r}{\lambda} \qquad \ldots(10)$$

where the subscripts refer to left and right, n is the appropriate refractive index, and λ is the wavelength of light employed. The units are radians per unit length, with the length units given by those used for λ.

If the concentration of an optically active substance, c', is expressed in units of g cm^{-3} (corresponding to the density for a pure substance), the specific rotation $[\alpha]$ is defined as:

$$[\alpha] = \frac{\alpha}{c'd'} \qquad ...(11)$$

where d' is the thickness of the sample in decimeters. The molar rotation $[M]$ is defined as:

$$[M] = M[\alpha] \times 10^{-2} = M\alpha \times 10^{-2} / c'd' \qquad ...(12)$$

where M is the molecular weight of the optically active component. (The quantity 10^{-2} is subject to convention and not always included in $[M]$.)

The *optical rotatory dispersion* curve, ORD, is a plot of the molar rotation, $[\alpha]$ or $[M]$, against λ. When the plane of polarization rotates clockwise as viewed by an observer facing the direction of propagation of the radiation, $[\alpha]$ or $[M]$ is defined as positive; a counterclockwise rotation is defined as negative.

The technique whereby one determines that an optically active substance *absorbs* right and left circularly polarized light differently is called *circular dichroism, CD*. All optically active substances exhibit CD in the region of appropriate electronic absorption bands. The molar circular dichroism $\varepsilon_l - \varepsilon_r$ is defined as

$$\varepsilon_\ell - \varepsilon_r = \frac{k_\ell - k_r}{c} \qquad ...(13)$$

where k, the absorption coefficient, is defined by $I = I_0\, 10^{-kd}$ with I_0 and I being the intensity of the incident and resultant light and d being the cell thickness. By plotting $\varepsilon_l - \varepsilon_r$ versus λ, the CD curve results.

Wherever circular dichroism is observed in a sample, the resulting radiation is not plane polarized, but is elliptically polarized. The quantity a in the above equations is then the angle between the initial plane of polarization and the majór axis of the ellipse of the resultant light. One can define a quantity φ' (in radians), the tangent of which is the ratio of the major to minor axes of the ellipse. The quantity φ' is used to approximate the ellipticity; when it is expressed in degrees, it can be converted to a specific ellipticity $[\varphi]$ or molar ellipticity $[\theta]$ by

$$[\varphi] = \frac{\varphi'}{c'd'} \qquad ...(14)$$

and

$$[\theta] = M[\varphi]10^{-2} \qquad ...(15)$$

where the symbols are as defined in equations (11) and (12). The quantity $[\theta]$ is related to $\varepsilon_l - \varepsilon_r$ by the following equation:

$$\varepsilon_\ell - \varepsilon_r = 0.3032 \times 10^{-3}[\theta] \qquad \ldots(16)$$

Thus, one often sees the CD curve plotted as [θ] versus λ.

With CD one can measure only the optical activity if there is an accompanying electronic absorption band. On the other hand, ORD is measurable both inside and outside the absorption band.* The ORD and CD curves of D-(–)-$[Rh(en)_3]^{3+}$ are illustrated. Throughout most of the visible region, the ORD curve is negative. However, the CD curve associated with the visible *d-d* transitions at ~300 mμ is clearly positive. All the chromophores in a molecule contribute to the rotatory power at a given wavelength, but only the chromophore that absorbs at the given wavelength contributes to the CD. Thus, a transition in the far uv can make a significant contribution to the rotation in the region of *d-d* transitions in the ORD. The negative effect in the uv dominates the *d-d* contribution through most of the visible region, and the negative ORD curve results. For most of the applications to be discussed here, CD is the method of choice.

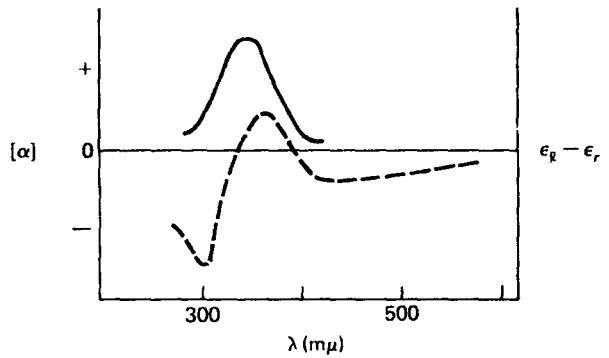


Figure 2.14 ORD (---) and CD (—) curves for D.(–)-$[Rh(en)_3]^{3+}$.

2.8.1 Selection Rules

We have previously given [equation (3)) the transition moment integral for an electric dipole transition, and we mentioned that a magnetic dipole transition integral has a similar form. In order for an electronic transition to give rise to optical activity, the transition must be both electric and magnetic dipole allowed, i.e.,

$$R \propto \left[\int \psi_{el} \hat{M} \psi_{el}^{ex} d\upsilon\right]\left[\int \psi_{el} \hat{M}D \psi_{el}^{ex} d\upsilon\right]$$

where R is the rotational strength, $\hat{M}$D is the magnetic dipole operator, and $\hat{M}$ is the electric dipole operator. If an electronic

absorption band is observed, there must be some mechanism for making this allowed; it then becomes important to be concerned with the magnetic dipole selection rules. The sign and magnitude of the activity can be calculated by evaluating both the electric and magnetic dipole integrals.

2.8.2 Applications

The use of CD in band assignments is an obvious application of our previous discussion of the selection rules. For example, octahedral nickel(II) complexes have three bands assigned as $^3A_{2g}$ to $^3T_{2g}$, $^3T_{1g}$(F), and $^3T_{1g}$(P) in order of increasing energy. Since the magnetic dipole operator (the magnetic dipole transforms like the rotations $R_xR_yR_z$) in O_h is T_{1g}, only the $^3A_{2g} \rightarrow {}^3T_{2g}$ transition is magnetic dipole allowed. Accordingly, it is found in the CD spectrum of $Ni(pn)_3^{2+}$ (pn = propylene diamine) that the low-energy band has a maximum ($\varepsilon_l - \varepsilon_r$) value of 0.8, while those for the other bands are less than 0.04. This confirms the original band assignments.

A second type of application involves using CD to show that certain absorption bands have contributions from more than one electronic transition. The CD bands are usually narrower and can be positive or negative. This idea is illustrated the absorption curve and CD curve for Δ(+)-$Co(en)_3^{3+}$ in aqueous solution.

Co(III) complexes with O_h symmetry commonly have two absorptions assigned to $^1T_{1g}$ and $^1T_{2g}$. The symmetry of $Co(en)_3^{3+}$ is

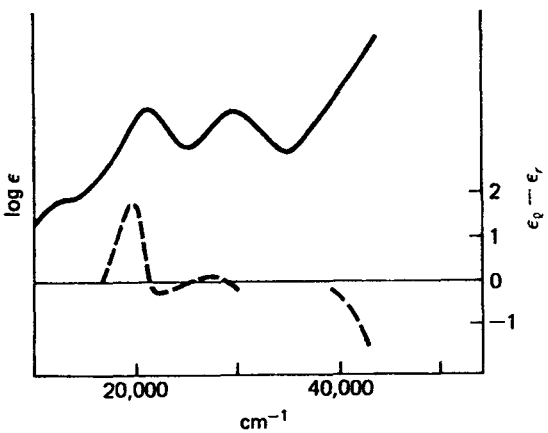


Figure 2.15 The absorption spectrum and CD curve of (+)-$Co(en)_3^{3+}$ in aqueous solution.

D_3, so the transitions to the T states are expected to show some splittings. This is not detected in the absorption spectrum. However, the effects of lower symmetry are observed in the CD spectrum. The magnetic dipole selection rules for D_3 predict that the low-energy band will have two magnetic dipole allowed components, $^1A_1 \rightarrow {}^1A_2$ and $^1A_1 \rightarrow {}^1E_a$. The $^1A_1 \rightarrow {}^1E_b$ transition of the high-energy band is magnetic dipole allowed, but the $^1A_1 \rightarrow {}^1A_1$ transition is not. In the CD, two components (+ and -) are seen in the low-energy band and one in the high-energy band, as predicted for a D_3 distortion. Applications of these ideas can be used to indicate the symmetry of molecules in solution and in uniaxial single crystals.

Another application involves using the sign of the CD to obtain the absolute configuration of a molecule. This application has been particularly successful for organic compounds. In inorganic systems, absolute configurations are often assigned by analogy to known systems. Particular care must be employed in determining what constitutes an analogous complex. However, by using complete operator-matrices for the electric and magnetic dipole components, the signs of the trigonal components in some Co(III) and Cr(III) complexes have been related to the absolute configurations of the complexes.

Finally, optically active transitions are polarized, and the polarization information can be used to support the assignment of the electronic spectrum.

2.8.3 Magnetocircular Dichroism

When plane-polarized light is passed through any substance in a magnetic field, H_0, whose component in the direction of the light propagation is non-zero, the substance appears to be optically active. Left and right circularly polarized light do not interact in equivalent ways. For atoms, for example, left circularly polarized (lcp) light induces a transition in which Δm_J is -1, while for right circularly polarized (rcp) light Δm_J is $+1$. Note that m_J has the same relationship to J as m_l has to l. If one observed a transition from an S state where $J = 0$ to a ^1P state where $J = 1$, the two transitions would occur as a consequence of this selection rule.

If the absorptions of left and right circularly polarized light corresponding to these transitions are measured separately, the curves are obtained. Here ν_0 is the band maximum for the absorption band. When the mcd curve is plotted, the result is obtained, provided that the band width is much greater than the Zeeman splitting of

the excited state. A curve of this sort is referred to as an A-term and can arise only for a transition in which $J > 0$ for one of the states involved. The sign of the A-term in molecules depends upon the sign of the Zeeman splitting and the molecular selection rules for circularly polarized light.

Next consider a transition from a ^1P to a ^1S state. The $\Delta m_J = +1$ transition for rcp light is now that of m_J from -1 to 0. Because the m_J states are not equally populated but Boltzmann populated, the two transitions will not have equal intensity. The relative intensities will be very much temperature dependent. The resultant mcd curve, is referred to as a C-term. The band shape and intensity are very temperature dependent. An A-term curve usually occurs superimposed upon a C-term curve.

A third type of curve (B-term) results when there is a field-induced mixing of the states involved (this phenomenon also creates temperature independent paramagnetism, TIP, and will be discussed in more detail in the chapter on magnetism). This is manifested in a curve that looks like a C-curve but that is temperature independent. Since this mixing is present to some extent in all molecules, *all substances have mcd activity*. The magnitude of the external magnetic field intensity will determine whether or not the signal is observed.

The following characteristics summarize the basis for detecting and qualitatively interpreting mcd curves:

1. An A-term curve changes sign at the absorption maximum, while B and C curves maximize or minimize at the maximum of the electronic absorption band.
2. A C-term curve's intensity is inversely proportional to the absolute temperature, while a B-term is independent of temperature.
3. An A-term spectrum is possible only if the ground or excited state involved in the electronic transition is degenerate and has angular momentum.
4. A C-term spectrum is possible only if the ground state is degenerate and has angular momentum.

As can be anticipated, mcd measurements are of considerable utility in assigning the electronic spectrum of a compound. Furthermore, the magnitude of the parameters is of considerable utility in providing information about many subtle electronic effects. The molecular orbital origin of an electron involved in a transition can be determined. The lowest energy band in RuO_4 is clearly an

oxygen to ruthenium charge-transfer band. One cannot determine from the electronic spectrum whether the oxygen electron involved in the transition came from a $t_1\pi$ or $t_2\pi$ type of oxygen molecular orbital. The sign of the mcd A-term established the transition as $t_1\pi$ (oxygen) $\rightarrow e_{x^2-y^2,z^2}$ (ruthenium). Another significant advantage of mcd is in the assignment of spin-forbidden electronic transitions that have very low intensity in the electronic absorption spectrum. The assignments of the components in a six-coordinate chromium(III) complex have been made with this technique. Other applications have been summarized in review articles. Mcd has been extensively applied to provide information about the properties of excited electronic states. Deductions regarding their symmetries, angular momenta, electronic splittings, and vibrational-electronic interactions are possible.

3

Nuclear Magnetic Resonance Spectroscopy

Protons and neutrons both have a spin quantum number of ½ and, depending on how these particles pair up in the nucleus, the resultant nucleus mayor may not have a net non-zero nuclear spin quantum number, I. If the spins of all the particles are paired, there will be no net spin and the nuclear spin quantum number I will be zero. This type of nucleus is said to have zero spin. When I is ½ , there is one net unpaired spin and this unpaired spin imparts a nuclear magnetic moment, μ, to the nucleus. The distribution of positive charge in a nucleus of this type is spherical. The properties for I = ½ are represented symbolically as a spinning sphere (*vide infra*). When $I \geq 1$, the nucleus has spin associated with it and the nuclear charge distribution is non-spherical. The nucleus is said to possess a quadrupole moment eQ, where e is the unit of electrostatic charge and Q is a measure of the deviation of the nuclear charge distribution from spherical symmetry. For a spherical nucleus, eQ is zero. A positive value of Q indicates that charge is oriented along the direction of the principal axis, while a negative value for Q indicates charge accumulation perpendicular to the principal axis.

Typical examples include ^{16}O, ^{12}C, and ^{32}S. The values for I, μ, and eQ for many other nuclei have been tabulated and are more easily looked up than figured out. NMR spectroscopy is most often concerned with nuclei with I = ½, examples of which include ^{1}H, ^{13}C, ^{31}P, and ^{19}F. Spectra often result from nuclei for which $I \geq 1$, but cannot be obtained on nuclei with I = 0.

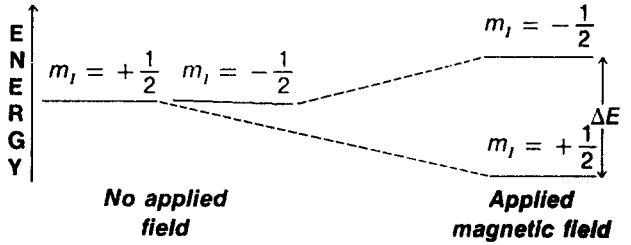


Figure 3.1 Splitting of the $m = \pm ½$ states in a magnetic field.

Unpaired nuclear spin leads to a nuclear magnetic moment. The allowed orientations of the nuclear magnetic moment vector in a magnetic field are indicated by the nuclear spin angular momentum quantum number, m_I. This quantum number takes on values I, $I - 1$, ..., $(-1 + 1)$, $-I$. When $I = ½$, $m_I = \pm ½$ corresponding to alignments of the magnetic moment with and opposed to the field. When $I = 1$, m_I has values of 1, 0, and -1, corresponding, respectively, to alignments with, perpendicular to, and opposed to the field. In the absence of a magnetic field, all orientations of the nuclear moment are degenerate. In the presence of an external field, however, this degeneracy will be removed. For a nucleus with $I = ½$, the $m_I = + ½$ state will be lower in energy and the $-½$ state higher. The energy difference between the two states, at magnetic field strengths commonly employed, corresponds to radio frequency radiation; it is this transition that occurs in the nmr experiment.

3.1 CLASSICAL DESCRIPTION OF THE NMR EXPERIMENT—THE BLOCH EQUATIONS

3.1.1 Some Definitions

We shall begin our classical description of nmr by reviewing some of the background physics of magnetism and defining a few terms necessary to understand nmr. It is very important to appreciate what is meant by angular momentum. Circulating charges have angular momentum associated with them. Planar angular momentum, $\bar{p}(\varphi)$ is given by $\bar{p}(\varphi) = \bar{r} \times m\bar{v}$ where $\bar{r}$, is the position vector of the particle e; v its linear momentum vector; m is the mass; and φ is the angular change that signifies an angular momentum. The angular momentum $\bar{p}$ is perpendicular to the plane of the circulating charge, and the direction of the angular momentum vector is given by the right hand rule.

The nucleus is a more complicated three-dimensional problem. The total angular momentum of a nucleus is given by $\bar{J}$, but it is

convenient to define a dimensionless angular momentum operator $\hat{I}$ by

$$\hat{J} = \hbar\hat{I} \qquad ...(1)$$

Associated with the angular momentum is a classical magnetic moment, $\vec{\mu}_N$ which can be taken as parallel to $\vec{J}$, so:

$$\vec{\mu}_N = \gamma\vec{J} = \gamma\hbar\vec{I} \qquad ...(2)$$

where γ, the magnetogyric (sometimes called gyromagnetic) ratio is a constant characteristic of the nucleus. From equation (2), we see that the magnetogyric ratio represents the ratio of the nuclear magnetic moment to the nuclear angular momentum.

3.1.2 Behaviour of a Bar Magnet in a Magnetic Field

A nucleus with a magnetic moment can be treated as though it were a bar magnet. If a bar magnet were placed in a magnetic field, the magnet would precess about the applied field, $\vec{H}_0$, as shown for a spinning nuclear moment. Here θ is the angle that the magnetic moment vector makes with the applied field; and ω, the Larmor frequency, is the frequency of the nuclear moment precession. The instantaneous motion of the nuclear moment (indicated by the arrowhead on the dashed circle) is tangential to the circle and perpendicular to $\vec{\mu}$ and $\vec{H}_0$. The magnetic field is exerting a force or torque on the nuclear moment, causing it to precess about the applied field. For use later, we would like to write an equation to describe the precession of a magnet in a magnetic field. The applied field $\vec{H}_0$ exerts a torque $\vec{\tau}$ on $\vec{\mu}$; which is given by the *cross product:*

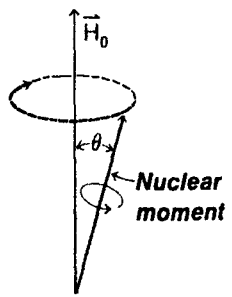


Figure 3.2 Procession of a nuclear moment in an applied field of strength H_0.

$$\vec{\tau} = \vec{\mu} \times \vec{H}_0 \qquad ...(3)$$

The right hand rule tells you that *the moment is precessing clockwise, with the torque and instantaneous motion perpendicular to* $\vec{\mu}$ and $\vec{H}_0$. According to Newton's Law, force is equal to the time derivative of the momentum, so the torque is the time derivative of the angular momentum.

$$\vec{\tau} = \frac{d}{dt}(\hbar\vec{I})$$

With the equation $\vec{\mu} = \gamma\hbar\vec{I}$, the equation for precession of the moment is given by:

$$\frac{d\vec{\mu}}{dt} = \frac{\gamma d(\hbar\vec{I})}{dt} = \gamma\vec{\tau}$$

or ...(4)

$$\frac{d\vec{\mu}}{dt} = \dot{\vec{\mu}} = -\gamma\vec{H}_0 \times \vec{\mu}$$

(The dot is an abbreviation for the time derivative of some property, in this case $\vec{\mu}$; and the minus sign arises because we have changed the order for taking the cross product.) Thus, ω, the precession frequency or the *Larmor frequency,* is given by:

$$\omega = |\gamma|H_0 \qquad ...(5)$$

where the sign of γ determines the sense of the precession. According to equation (5), the frequency of the precession depends upon the applied field strength and the magnetogyric ratio of the nucleus. The energy of this system is given by the dot product of $\vec{\mu}$ and $\vec{H}_0$.

$$E = -\vec{\mu} \bullet \vec{H} = -|\mu|\,|H_0|\cos\theta \qquad ...(6)$$

3.1.3 Rotating Axis Systems

There is one more mathematical construct that greatly aids the analysis of the nmr experiment, and this is the idea of a rotating coordinate system or *rotating frame.* In figure a set of *x, y,* and *z* coordinates is illustrated. The rotating frame is described by the rotating axes *u* and υ, which rotate at some frequency ω_1 in the *xy-plane.* The *z*-axis is common to both coordinate systems. If the rotating frame rotates at some frequency less than the Larmor frequency, it would appear to an observer on the frame that the

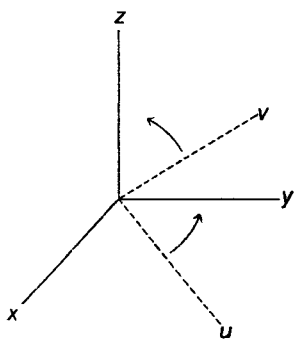


Figure 3.3 The rotating frame *u, v, z* in a Cartesian coordinate system *x, y, z.*

precessional frequency has slowed, which would correspond to a weakening of the applied field. Labeling this apparently weaker field as $\bar{H}_{eff}$, our precessing moment is described in the rotating frame by

$$\dot{\bar{\mu}} = -\gamma \bar{H}_{eff} \times \bar{\mu} \qquad ...(7)$$

where $|H_{eff}| < |H_0| \cdot \bar{H}_{eff}$ is seen to be a function of ω_1, and if ω_1 is faster than the precessional frequency of the moment, ω, it will appear as though the z-field has changed direction. The effective field, $\bar{H}_{eff}$' can be written as:

$$\bar{H}_{eff} = \left[H_0 - \frac{\omega_1}{\gamma} \right] \vec{e}_z \qquad ...(8)$$

where $\vec{e}_z$ is a unit vector along the z-axis. The time dependence of the moment, $\bar{\mu}$, can be rewritten as

$$\dot{\bar{\mu}} = -\gamma \left\{ \left(H_0 - \frac{\omega_1}{\gamma} \right) \vec{e}_z \right\} \times \bar{\mu} \qquad ...(9)$$

When $\omega_1 = \omega = \gamma H_0$, then $|H_{eff}| = 0$ and $\bar{\mu}$ appears to be stationary. Thus, in a frame rotating with the Larmor frequency, we have eliminated time dependency and, in so doing, greatly simplified the problem.

3.1.4 Magnetization Vectors and Relaxation

The concepts in this section are vital to an understanding of nmr. We shall develop the concept of relaxation in stages as the necessary background information is covered. For practical purposes, it is necessary to consider an ensemble or large number of moments because the nmr experiment is done with bulk samples. The individual moments in the sample add vectorially to give a *net magnetization*, $\bar{M}$.

$$\bar{M} = \sum_i \bar{\mu}_i \qquad ...(10)$$

In an ensemble of spins in a field, those orientations aligned with the field will be lower in energy and preferred. However, thermal energies oppose total alignment; experimentally, only a small net magnetization is observed. The equation for the motion of precession of $\bar{M}$ is similar to that for $\bar{\mu}$, i.e.,

$$\dot{\bar{M}} = -\gamma \bar{H}_0 \times \bar{M} \qquad ...(11)$$

If we place a sample in a magnetic field at constant temperature and allow the system to come to equilibrium, the resulting system is said to be at thermal equilibrium. At thermal equilibrium, the magnetization, M_0, is given by

$$|M_0| = \frac{N_0 \mu^2}{3kT} I(I+1)H_0 \qquad ...(12)$$

where k is the Boltzmann constant and N_0 is the number of magnetic moments per unit volume *(i.e.,* the number of spins per unit volume). This equation results from equation (6) and the assumption of a Boltzmann distribution. For H_2O at 27° and H_0 = 10,000 gauss, $M_0 \approx 3 \times 10^{-6}$ gauss.

The description of the magnetization is not yet complete for, at equilibrium, we have a dynamic situation in which anyone given nuclear moment is rapidly changing its orientation with respect to the field. The mechanism for reorientation involves time dependent fields that arise from the molecular motion of magnetic nuclei in the sample. Suppose that nucleus B, a nucleus that had a magnetic moment, passed by nucleus A, whose nuclear moment is opposed to the field. The spin of A could be oriented with the field via the fluctuating field from moving B. In the process, the translational or rotational energy of the molecule containing B would be increased, since A gives up energy on going to a more stable orientation aligned with the field. This process is referred to as *spin-lattice relaxation.* Nucleus B is the lattice; it can be a magnetic nucleus in the same or another sample molecule or in the solvent. These nuclei do not have to be hydrogens and may even be unpaired electrons in the same or other molecules. The fluctuating field from motion of the magnetic nucleus must have the same frequency as that of the nmr transition in order for this relaxation process to occur. However, moving magnetic nuclei have a wide distribution of frequency components, and the required one is usually included.

Another process can also occur when the two nuclei interact, whereby nucleus B, with $m_I = +½$ goes to the higher-energy $m_I = -½$ state while nucleus A changes from $-½$ to $+½$. There is no net change in spin from this process, and it is referred to as a *spin-spin relaxation* mechanism.

To gain more insight into the nature of relaxation processes, let us examine the decay in magnetization as a function of time. Consider an experiment in which we have our magnetization aligned along the u axis. We then switch on a field aligned with the z-axis;

M_u diminishes and M_z grows, generally at different rates. For example, the u-component has completely decayed, but the z-component has not yet reached equilibrium, for it gets larger. The growth and decay are first order processes, and Bloch proposed the following equation for the three-dimensional case (*i.e.*, u, υ, and z magnetizations are involved and the field is located along the z-axis):

$$\dot{M}_u = -\frac{M_u}{T_2} \quad \dot{M}_\upsilon = -\frac{M_\upsilon}{T_2} \qquad ...(13)$$

and

$$\dot{M}_z = -\frac{1}{T_1}(M_z - M_0) \qquad ...(14)$$

where $1/T_2$ and $1/T_1$ are first order rate constants, M_0 is the equilibrium value of the z magnetization, and u- and υ-components vanish at equilibrium. In general, it is round that $1/T_2 \geq 1/T_1$. Instead of rate constants, $1/T$, it is more usual to refer to their reciprocals (*i.e.*, the T's), which are lifetimes or relaxation times. Since T_1 refers to the z-component, it is called the *longitudinal relaxation time*, while T_2 is called the *transverse relaxation time*. The spin-lattice mechanism contributes to T_1, and the spin-spin mechanism is one of several contributions to T_2.

3.1.5 NMR Transition

Before proceeding with our classical description of the nmr experiment, it is advantageous to introduce a few concepts from the quantum mechanical description of the experiment. When the bare nucleus (no electrons around it) is placed in a magnetic field, H_0, the field and the nuclear moment interact as described by the Hamiltonian for the system

$$\hat{H} = -\vec{\mu}_N \bullet \vec{H}_0 \qquad ...(15)$$

where, as shown in equation (2), $\vec{\mu}_N = \gamma\hbar\vec{I}$ with the N subscript denoting a nuclear moment. When it is obvious that we are referring to a nuclear moment, the N subscript will be dropped. Combining (2) and (15) yields

$$\hat{H} = -\gamma_N \hbar H_0 \hat{I}_z = -g_N \beta_N H_0 \hat{I}_z \qquad ...(16)$$

where γ_N is constant for a given nucleus, g_N is the nuclear g factor and:

$$\beta_N = e\hbar/2Mc \qquad ...(17)$$

In equation (17), M is the mass of the proton, e is the charge of the proton, and c is the speed of light.

The expectation values of the operator $\hat{I}_z$, where z is selected as the applied field direction, are given by m_I where $m_I = I, I-1, \ldots, -I$. The degeneracy of the m_I states that existed in the absence of the field is removed by the interaction between the field H_0 and the nuclear magnetic moment μ_N. The quantized orientations of these nuclear moments relative to an applied field $\bar{H}_0$ for $I = ½$ and $I = 1$. Since the eignvalues of the operator $\hat{I}_z$ are m_I, the eigenvalues of $\hat{H}$ (i.e., the energy levels) are given by equation (18).

$$E = -\gamma\hbar m_I H_0 \qquad \ldots(18)$$

The energy as a function of the field is illustrated for $I = ½$. The quantity g_N is positive for a proton, as is β_N, so the positive m_I value is lowest in energy. The nuclear wave functions are abbreviated as $|\alpha\rangle$ and $|\beta\rangle$ for the $+½$ and $-½$ states, respectively. By inserting values of $m_I = +½$ and $m_I = -½$ into equation (18), we find that the energy difference, ΔE, between these states or the energy of the transition $h\nu$ for a nucleus of spin ½ is given by $g_N\beta_N H_0$ or $\gamma\hbar H_0$.

Since $m_I = +½$ is lower in energy than $m_I = -½$ there will be a slight excess population of the low energy state at room temperature, as described by the Boltzmann distribution expression in equation (19):

$$\frac{N(-½)}{N(+½)} = -\exp(-\Delta E/kt) \cong 1 - \frac{\Delta E}{kT} \text{ when } \Delta E < kT \qquad \ldots(19)$$

ΔE is $\sim 10^{-3}$ cm^{-1} in a 10,000 gauss field, and kT ~ 200 cm^{-1}. At room temperature, there is a ratio of 1.0000066 ($+½$) spins to one ($-½$). The probability of a nuclear moment being in the $+½$ state is (½)

$$\left(1 + \frac{\mu H_0}{kT}\right)$$

and that of it being in the $-½$ state is (½)

$$\left(1 - \frac{\mu H_0}{kT}\right)$$

[recall our discussion of equation (12)).

The energy separation corresponding to $h\nu$ occurs in the radio-frequency region of the spectrum at the magnetic field strengths usually employed in the experiment. One applies a circularly polarized radio frequency (r.f.) field at right angles to $\bar{H}_0$ (*vide infra*), and the magnetic component of this electromagnetic field, $\bar{H}_1$, provides

a torque to flip the moments from $m_I = +½$ to $-½$,causing transitions to occur.

3.1.6 Bloch Equations

In order to understand many of the applications of nmr, it is necessary to appreciate the change in magnetization of the system with time as the H_1 field is applied. This result is provided with the Bloch equation. Incorporating equations (13) and (14), describing relaxation processes, into (11), which describes the precession of the magnetization, and converting to the rotating frame gives the Bloch equation:

$$\dot{\bar{M}} = \underbrace{-\gamma \bar{H}_{eff} \times \bar{M}}_{\substack{\text{torque form} \\ \text{the magnetic} \\ \text{field}}} \underbrace{- \frac{1}{T_2}(M_u \bar{e}_u + M_\upsilon \bar{e}_\upsilon) - \frac{1}{T_1}(M_z - M_0)\bar{e}_z}_{\text{relaxation effects}} \qquad ...(20)$$

In the presence of H_1 and in the rotating frame, equation (8)—which described $\bar{H}_{\text{eff}}$ in the rotating frame-becomes:

$$\bar{H}_{eff} = \left(H_0 - \frac{\omega_1}{\gamma} \right) \bar{e}_z + \bar{H}_1 \bar{e}_u \qquad ...(21)$$

where the frame is rotating with the frequency ω_1 corresponding to the frequency of H_1, the oscillating field at right angles to $\bar{H}_0$. Equation (20) is a vector equation in the rotating frame that can best be written in terms of the components of $\bar{M}$, which are M_u, M_υ, and M_z.

The three components of the Bloch equation are:

$$\dot{M}_u = \frac{dM_u}{dt} = -(\omega_1 - \omega_0)M_\upsilon - \frac{M_u}{T_2} \qquad ...(22)$$

$$\dot{M}_\upsilon = \frac{dM_\upsilon}{dt} = (\omega_1 - \omega_0)M_u - \frac{M_\upsilon}{T_2} + \gamma H_1 M_z \qquad ...(23)$$

$$\dot{M}_z = \frac{dM_z}{dt} = -\gamma H_1 M_\upsilon + (M_0 - M_z)/T_1 \qquad ...(24)$$

In these equations, ω_0 is the Larmor frequency, which equals γH_0; and the u, υ reference frame is rotating at angular velocity ω_1.

Experimentally, we monitor the magnetization in the xy plane, referred to as the transverse component. Using a phase-sensitive detector, we monitor the component of magnetization induced along the u-axis. In the normal slow passage or steady state experiment, only a u-component of magnetization exists; but because of a 90°

phase lag associated with the electronic detection system, a component 90° out of phase with u is measured. Slow passage or steady state conditions require H_1 to be weak (of the order of milligauss) compared to H_0 (which is of the order of kilogauss). Then, according to equation (21), the z-component dominates unless ω_1 is close to ω_0 so that the first term ($H_0 = \omega_0/\gamma$) becomes small. (Note that ω_1 is the frequency of the r.f. field and *not* the Larmor frequency of precession about H_1; *i.e.*, $\omega_1 \neq \gamma H_1$.) When ω_1 is close to the Larmor frequency, ω_0, then $\vec{H}_{\text{eff}}$ is tipped toward the u-axis. Since H_0 is being changed slowly, the net effect is to change H_{eff} slowly. The individual moments continue to precess about $\vec{H}_{\text{eff}}$ as a consequence of the torque, which is perpendicular to $\vec{H}_{\text{eff}}$. As a result, $\vec{M}$ remains parallel to $\vec{H}_{\text{eff}}$ and a u-component results. Figure represents the tipping of $\vec{M}$ to remain aligned with $\vec{H}_{\text{eff}}$ as we pass through resonance. The $\vec{H}_1$ field is along the u-axis. It is also an alternating field with the frequency of the rotating frame. A static field will not tip the magnetization vector significantly because H_1 is so small.

Relaxation processes, T_1, are trying to preserve the orientation along the strong z-field, as shown by the arrow labeled T_1. Under steady state conditions, all time derivatives are zero, so equations (22), (23), and (24) are all equal to zero and can be solved to produce:

$$M_u = M_0 \frac{\gamma H_1 T_2^2 (\omega_0 - \omega_1)}{1 + T_2^2(\omega_0 - \omega_1)^2 + \gamma^2 H_1^2 T_1 T_2} \qquad \ldots(25)$$

$$M_v = M_0 \frac{\gamma H_1 T_2}{1 + T_2^2(\omega_0 - \omega_1)^2 + \gamma^2 H_1^2 T_1 T_2} \qquad \ldots(26)$$

$$M_z = M_0 \frac{1 + T_2^2(\omega_0 - \omega_1)^2}{1 + T_2^2(\omega_0 - \omega_1)^2 + \gamma^2 H_1^2 T_1 T_2} \qquad \ldots(27)$$

3.1.7 NMR Experiment

Equations (25) to (27) describe the magnetization of our sample in the so-called slow passage experiment. In this method, one applies a strong homogeneous magnetic field, causing the nuclei to precess. Radiation of energy comparable to ΔE is then imposed with a radio frequency transmitter, producing H_1. When the applied frequency from the radio transmitter is equal to the Larmor frequency, the two are said to be in resonance, and a u,v-component is induced

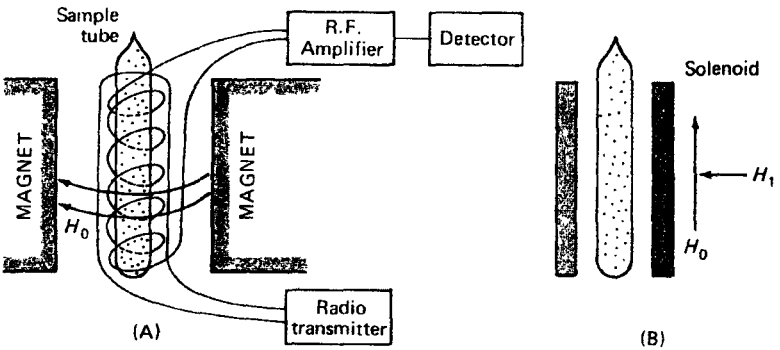


Figure 3.4 Schematic diagram of a simple nmr spectrometer.

which can be detected. This is the condition in (21) when $H_0 \approx \omega_1/\gamma$. Quantum mechanically, this is equivalent to some nuclei being excited from the low energy state (m_I =, $+\frac{1}{2}$) to the high energy state ($m_I = -\frac{1}{2}$) by absorption of energy from the r.f. source at a frequency equal to the Larmor frequency. Since $\Delta E = h\nu$ and $\omega = 2\pi\nu$, ΔE is proportional to the Larmor frequency, ω. Energy will be extracted from the r.f. source only when this resonance condition ($\omega = 2\pi\nu$) is fulfilled. With an electronic detector, one can observe the frequency at which a u,v-component is induced or at which the loss of energy from the transmitter occurs, allowing the resonance frequency to be measured.

It is possible to match the Larmor frequency and the applied radio frequency either by holding the field strength constant (and hence ω constant) and scanning a variable applied radio frequency until matching occurs or, as is done in the more common nmr apparatus, by varying the field strength until ω becomes equal to a fixed applied frequency. In the latter method, fixed frequency probes (source and detector coils) are employed and the field strength at which resonance occurs is measured. Two experimental configurations are used in this experiment. The H_0 field direction (z-direction) is perpendicular to the sample tube, while the H_0 field and the sample are coaxial. There are some applications in which this difference is important (*vide infra*).

We can now see why the relaxation processes discussed in equations (13) and 14) had to be added to our complete equation [(20), which led to (25) to 27) for the slow passage experiment] for the behaviour of the magnetization. If the populations of nuclei in the ground and excited states were equal, then the probability

that the nucleus would emit energy under the resonance condition would equal the probability that the nucleus would absorb energy [*i.e.*, transitions $m_I(+½) \rightarrow m_I(-½)$ would be as probable as $m_I(-½) \rightarrow m_I(+½)$]. No net change would then be detected by the radio frequency probe. As mentioned earlier, in a strong magnetic field there will be a slight excess of nuclei aligned with the field (lower energy state) and consequently a net absorption of energy results. As energy is absorbed from the r.f. signal, enough nuclei could be excited after a finite period of time so that the population in the lower state would be equal to that in the higher sate. Initially, absorption might be detected but this absorption would gradually disappear as the populations of ground and excited states became equal. When this occurs, the sample is said to be saturated. If the nmr instrument is operated properly, saturation usually does not occur, because the relaxation mechanisms allow nuclei to return to the lower energy state without emitting radiation. As a result there is always an excess of nuclei in the lower energy state, and a continuous absorption of energy from the r.f. source by the sample occurs.

Remembering that the u-component of the magnetization is induced, equation (25) predicts that an absorption band in the nmr spectrum will have the general form of a Lorentzian function,

$$A\left(\frac{1}{a+bx^2}\right),$$

for when H_1 is small the last term in the denominator is negligible. Thus, as we sweep through resonance, the magnitude of M_u gives a Lorentzian plot which is the nmr spectrum. When a large H_1 is employed, conditions leading-to saturation are observed and the shape of the spectral line is grossly distorted from that of a Lorentzian by contributions to M_u from the $\gamma^2 H_1^2 T_1 T_2$ term of the denominator. In the extreme of high source power, no signal is observed, for we in effect destroy the population difference between $m_I = +½$ and $-½$ and the resulting net z-component of the magnetization.

The nmr experiment has significance to the chemist because the energy of the resonance (*i.e.*, the field strength required to attain a Larmor frequency equal to the fixed frequency) is dependent upon the electronic environment about the nucleus. The electrons shield the nucleus, so that the magnitude of the field seen at the nucleus, H_N, is different from the applied field, H_0:

$$H_N = H_0(1-\sigma) \qquad ...(28)$$

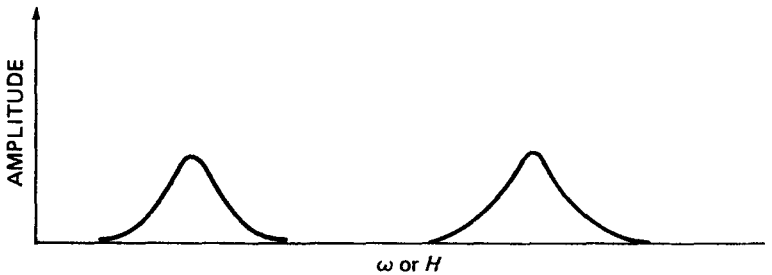


Figure 3.5 A low resolution nmr spectrum of a sample containing two different kinds of protons.

where σ, the shielding constant, is a dimensionless quantity that represents the shielding of the nucleus by the electrons. The value of the shielding constant depends on several factors, which will be discussed in detail later. Equation (28) states that the value of H_0 in equation (21) is different from the H applied for different nuclei in the molecule, and H_0 should then be replaced by H_N. Consequently, in a slow sweep of H_0, the various magnetizations of the different nuclei are sampled individually, because when H_1 dominates for one kind of nucleus the H_N term will still be dominant for others. For a molecule with two different kinds of hydrogen atoms this will lead to a spectrum.

One other point is worth making here. The differences in the magnetogyric ratios of different kinds of nuclei are much larger than the effects from σ, so there is no trouble distinguishing signals from the different kinds of nuclei in a sample; *e.g.*, ^{19}F and ^{1}H are never confused. The ranges are so vastly separated that different instrumentation is required to study different kinds of nuclei. The resonance conditions ($h\nu = g_N\beta_N H_0$) for various nuclei are given for an applied field of 10,000 gauss.

The frequencies are in megaHertz (10^6 Hertz or 10^6 c/s), and the variation in the proton resonances of typical organic compounds >from different shielding constants is only 600 Hz. In some paramagnetic complexes, shifts of the order of magnitude of 840,000 Hz have been observed; but even these could not be confused with a fluorine resonance.

Another limiting experiment is the pulse experiment. We shall discuss this experiment in more detail later, but now wish to briefly consider it in the context of equation (21). Suppose that a single strong pulse is imposed (*e.g.*, $H_1 \approx 100$ gauss $> H_0 - \omega_1/\gamma$ for all of the protons). Referring to the equation for $\bar{H}_{\text{eff}}$,

$$\vec{H}_{\text{eff}} = H_2\vec{e}_u + \left(H_0 - \frac{\omega_1}{\gamma}\right)\vec{e}_z$$

the local H_0 will vary slightly from nucleus to nucleus, but $\vec{H}_1$ is so large that the term $H_1\vec{e}_u$ is completely dominant, and $\vec{H}_{\text{eff}}$ is almost the same for all nuclei present. All nuclei are sampled simultaneously. Since $\vec{H}_1$ is directed along $\vec{u}$ and the net magnetization along $\vec{z}$, the cross product requires that $\vec{M}$ begin to precess about $\vec{u}$. The pulse time is so short, however, that the magnetization does not have time to precess around $\vec{u}$, but merely tips toward $\vec{v}$. The pulse duration is shorter than the time corresponding to one full precession about $\vec{u}$. Accordingly, the v-component is measured in this experiment as $\vec{M}$ is tipped toward $\vec{v}$. This is how ^{13}C experiments, for example, are now being done by Fourier transform procedures. If a strong pulse were employed for a very long time, the nuclei would precess around $\vec{H}_{\text{eff}}$ as in the slow passage experiment; a u-component would arise and v would disappear. Pulse experiments cannot be done this way. When a strong pulse is employed for a short enough duration that there is no time for any relaxation to occur during the pulse ($t_p \approx 10$ microseconds), the Bloch equations predict the following expression for the v-component, as a function of time, t:

$$M_v = \sin(\gamma H_1 t_p) M_0 \exp(-t/T_2) \qquad \ldots(29)$$

This expression results from the Fourier transform of equation (23), which allows conversion from the frequency domain (23) to the time domain (29). The quantity dM_v/dt cannot be set equal to zero in this experiment. The v-component of the magnetization decays with time according to equation (29). This is called a free induction decay curve and it can be analyzed by computer (via Fourier transformation, *vide infra*) to give a frequency spectrum identical to the Lorentzian slow passage result.

3.2 QUANTUM MECHANICAL DESCRIPTION OF THE NMR EXPERIMENT

3.2.1 Properties of $\hat{I}$

Now, to gain valuable insight, let us reexamine this whole problem using a quantum mechanical instead of a classical mechanical approach. Quantum mechanics shows us that for $I = ½$, there are two allowed orientations of the spin angular momentum vector in a magnetic field and will also indicate the necessary requirements to induce transitions between these energy states by an appropriate

perturbation, *i.e.*, the application of an oscillating magnetic field with energies corresponding to r.f. radiation. The necessary direction for this field can be determined from a consideration of the spin angular momentum operators.

The $\hat{I}^2$ operator has eigenvalues $I(I + 1)$ (as in an atom, where the orbital angular momentum operator $\hat{L}^2\psi = \ell(\ell + 1)\hbar^2\psi$). Anyone of the components of $\hat{I}$ (*e.g.*, $\hat{I}_z$) commutes with $\hat{I}^2$, so we can specify eigenvalues of both $\hat{I}^2$ and $\hat{I}_z$. The eigenvalues of $\hat{I}_z$ are given by I, $I - 1$, ... $-I$ (like m_l values in an atom, where the z-component of angular momentum, $\hat{L}_z$, is given by $\hat{L}_z\psi = m_\ell\hbar\psi$ with $m_\ell = \pm\ell,\ldots,0$). In general, if two operators commute, then there exist simultaneous eigenfunctions of both operators for which eigenvalues can be specified.

It is a simple matter to determine whether two operators commute. If they do, then by our earlier definition of commutation

$$\hat{I}^2\hat{I}_z - \hat{I}_z\hat{I}^2 = 0$$

The above difference is abbreviated by the symbol $[\hat{I}^2, \hat{I}_z]$. Similar equations can be written for $\hat{I}_x$ and $\hat{I}_y$, *i.e.*,

$$[\hat{I}^2, \hat{I}_x] = [\hat{I}^2, \hat{I}_y] = 0$$

However, $\hat{I}_z$ does not commute with $\hat{I}_x$ or $\hat{I}_y$, *e.g.*,

$$\hat{I}_z\hat{I}_y - \hat{I}_y\hat{I}_z \neq 0$$

Eigenvalues for $\hat{I}^2$ exist, and if we decide to specify eigenvalues for $\hat{I}_z$, then eigenvalues for $\hat{I}_x$ and $\hat{I}_y$ do not exist.

Average values could be calculated for the $\hat{I}_x$ and $\hat{I}_y$ operators. We shall make these concepts more specific by applying these operators to the spin wave functions α and β for $I = ½$ nuclei and showing the results:

$$\begin{aligned} \hat{I}_z\alpha &= (+½)\alpha \\ \hat{I}_z\beta &= (-½)\beta \end{aligned} \qquad \ldots(30)$$

$$\begin{aligned} \hat{I}_x\alpha &= (½)\beta \quad \hat{I}_x\beta = (-½)\alpha \\ \hat{I}_y\alpha &= (½)i\beta \quad \hat{I}_y\beta = (-½)i\alpha \end{aligned} \qquad \ldots(31)$$

Thus, the $\hat{I}_z$ operator yields eigenvalues, since operation on α gives back α and operation on β gives β. The $\hat{I}_x$ and $\hat{I}_y$ operators do not yield eigenvalues, since operation on α produces β and operation on β yields α. The average value for the property $\hat{I}_x$ or $\hat{I}_y$ is given by an equation of the sort $\int\psi^* \text{op}\psi d\tau / \int\psi^2 d\tau$. The following

relations hold for α and β (as they do for orthonormal electronic wave functions):

$$\int \alpha^2 d\tau = \int \beta^2 d\tau = 1 \quad \text{and} \quad \int \alpha\beta d\tau = 0$$

As mentioned, the integrals encountered in quantum mechanical descriptions of systems are written employing the *bra* and *ket notation* for convenience. Recall that the symbol | ⟩ is referred to as a ket and ⟨ | as a bra. An integral of the form $\int(\psi^*$ Operator $\psi)d\tau$ is written as $\langle\psi|\text{Op}|\psi\rangle$, while an integral of the form $\int\psi_1^*\psi_2 d\tau$ is written as $\langle\psi_1|\psi_2\rangle$.

3.2.2 Transition Probabilities

Consider the effect of the H_1 field in the quantum mechanical description. If the alternating field is written in terms of an amplitude H_x°, we get a perturbing term in the Hamiltonian of the form of equation (32):

$$\hat{H}_{pert} = -\gamma\hbar H_x^\circ \hat{I}_x \cos\omega_1 t \qquad ...(32)$$

Recall from equation (16) that $\hat{H}$ for a nucleus in a z-field was $\hat{H} = -\gamma\hbar H_0 \hat{I}_z = -g_N\beta_N H_0 I_z$, so now the perturbation is of a similar form but for an x-field that is alternating.

The equation describing the probability of a transition in the nmr, P, is similar to that in u.v. and i.r., and is given by

$$P = 2\pi\gamma_N^2 H_1^2 \left|\left\langle \varphi^{ex}\middle|\hat{I}_x\middle|\varphi\right\rangle\right|^2 g(\omega) \qquad ...(33)$$

where $g(\omega)$ is a general line shape function, which is an empirical function that describes how the absorption varies near resonance. To apply equation (33), we need to evaluate matrix elements of the form $\langle\varphi^{ex}|\hat{I}_x|\varphi\rangle$ and determine whether they are zero or non-zero.

The solution is best accomplished by constructing a matrix that summarizes all of the integrals that must be evaluated to describe the system in a magnetic field with and without H_1. Rows and columns are constructed, which are headed by the basis set. In this case, we have one nucleus with α and β nuclear spin wave functions leading to:

	α	β
α	$\left\langle\alpha\middle\|\hat{H}+\hat{H}_{\text{pert}}\middle\|\alpha\right\rangle$	$\left\langle\alpha\middle\|\hat{H}+\hat{H}_{\text{pert}}\middle\|\beta\right\rangle$
β	$\left\langle\beta\middle\|\hat{H}+\hat{H}_{\text{pert}}\middle\|\alpha\right\rangle$	$\left\langle\beta\middle\|\hat{H}+\hat{H}_{\text{pert}}\middle\|\beta\right\rangle$

By this procedure, we have considered all possible matrix elements and also have them in such a form that if E were subtracted from the diagonal elements we would have the secular determinant, which can be solved to give the energies of these states in an applied field $(H_0 + H_1)$. The resulting energies could then be used in the secular equations (produced by matrix multiplication of the secular determinant with a matrix of the basis set) to give the wave functions in the field.

We begin by evaluating the elements in the secular determinant when the applied field is H_0 (with a z-component only), *i.e.*, the Zeeman experiment. Since there is no x or y field component (only z), there are no $\hat{I}_x$ or $\hat{I}_y$ operators and all matrix elements of the form $\langle|\hat{I}_x|\rangle$ or $\langle|\hat{I}_y|\rangle$ are zero. The off-diagonal elements, $\langle\alpha|\hat{I}_z|\beta\rangle$ and $\langle\beta|\hat{I}_z|\alpha\rangle$, are also zero because $\hat{I}_z\,|\beta\rangle = -\frac{1}{2}\beta$, and $\langle\alpha|\beta\rangle$ and $\langle\beta|\alpha\rangle$ are zero. The only nonzero matrix elements are $\langle\alpha|\hat{I}_z|\alpha\rangle$ and $\langle\beta|\hat{I}_z|\beta\rangle$, with the former corresponding to stabilization, $+\frac{1}{2}$ (*i.e.*, $\frac{1}{2}\langle\alpha|\alpha\rangle$), and the latter to destabilization, $-\frac{1}{2}$. With no off-diagonal elements, the eigenvalues are directly obtained and the basis set is not mixed, so the two wave functions are α and β. When these are substituted into equation (33) for φ and φ^{ex}, the matrix element is zero and the transition is not allowed, $\langle\alpha|\beta\rangle = 0$.

Next, we shall consider what happens when an H_1 field along the x-axis is added to the Zeeman experiment described above. We must now worry about matrix elements involving $\hat{I}_x$. The diagonal elements $\langle\alpha|\hat{I}_x|\alpha\rangle$ are zero $[\langle\alpha|\hat{I}_x|\alpha\rangle = \frac{1}{2}\langle\alpha|\beta\rangle = 0]$, but the off-diagonal elements $\langle\alpha|\hat{I}_x|\beta\rangle$ are non-zero. Since the H_1 field is small compared to H_0 (z-component), these off-diagonal matrix elements are so small as to have a negligible effect on the energies (the effect of $\hat{I}_z$ on the diagonal is the same as in the Zeeman experiment). However, the small off-diagonal matrix elements are important because they provide a mechanism for inducing transitions from α to β because the new wave functions for the system with H_1 present (*i.e.*, obtained after diagonalizing our matrix) mix a little β character into the α-Zeeman Slate and a little α into the β-Zeeman state. The new wave function for the α-Zeeman state now is $\varphi = \sqrt{1-a^2}\,|\alpha\rangle + a|\beta\rangle$, where $a < 1$, and a similar change occurs in the β state. When these new wave functions φ are substituted into equation (33), P is non-zero, making the transition allowed. This corresponds to the classical picture of H_1 exerting a torque to give a transverse component to the magnetization. Thus, we see that the

probability, P, of a transition occurring depends upon the off-diagonal matrix elements in the α,β basis being non-zero, so that equation (33) is non-zero.

Next, consider what would happen if the r.f. perturbing field H_1 was placed along the z-axis. The off-diagonal matrix elements would again be zero, so equation (33) becomes

$$P = \langle \alpha | \hat{I}_z / \beta \rangle = 0$$

This would correspond to a slight reinforcement of H_0, but would not allow transitions. This exercise shows that H_1 cannot be collinear with H_0 to get an nmr transition. (In the classical model, there has to be a perpendicular component for H_1 to exert a torque.)

Finally, consider the case in which $I=1$. Matrix elements $\langle m' | \hat{I}_x | m \rangle$ vanish unless $m' = m \pm 1$. Consequently, for $I=1$, the allowed transitions are between adjacent levels with $\Delta m = \pm 1$, giving $\hbar\omega = \Delta E = \gamma\hbar H_0$.

In summary, simply placing a sample in a magnetic field, H_0, removes the degeneracy of the m_I states. Now, a radio frequency source is needed to provide $h\nu$ to induce the transition. Absorption of energy occurs provided that the magnetic vector of the oscillating electromagnetic field, H_1, has a component perpendicular to the steady field, H_0, of the magnet. Otherwise (*i.e.*, if H_1 is parallel to H_0), the oscillating field simply modulates the applied field, slightly changing the energy levels of the spin system, but no energy absorption occurs.

3.3 RELAXATION EFFECTS AND MECHANISMS

The influence of relaxation effects on nmr line shapes leads to some very important applications of nmr spectroscopy. Accordingly, it is worthwhile to summarize and extend our understanding of these phenomena. We begin with a discussion of relaxation processes and their effect on the shapes of our resonance line. The lifetime of a given spin state influences the spectral line width via the Uncertainty Principle, which is given in equation (34):

$$\Delta E \Delta t \approx \hbar \qquad \ldots(34)$$

Since $\Delta E = h\Delta\nu$ and $\Delta t = T_2$, the lifetime of the excited state, the range of frequencies is given by $\Delta\nu \approx 1/T_2$. The quantity $1/T_2$ as employed here lumps together all of the factors influencing the line width (i.e., all the relaxation processes) and is simply one-half the width of the spectral line at half-height. When the only contribution to T_2 is from spin-lattice effects, then $T_2 = T_1$. Most molecules

contain magnetic nuclei; in the spin-lattice mechanism a local fluctuating field arising from the motion of magnetic nuclei in the lattice (where the lattice refers to other atoms in the molecule or other molecules including the solvent) couples the energy of the nuclear spin to other degrees of freedom in the sample, *e.g.*, translational or rotational energy. For liquids, T_1 values are usually between 10^{-2} and 10^{2} seconds but approach values of 10^{-4} seconds if certain paramagnetic ions are present. The extent of spin-lattice relaxation depends upon (1) the magnitude of the local field and (2) the rate of fluctuation. Paramagnetic ions have much more intense magnetic fields associated with them and are very efficient at causing relaxation. The water proton signal in a 0.1 M solution of $Mn(H_2O)_6^{2+}$ is so extensively broadened by efficient relaxation from Mn^{2+} that a proton signal cannot be detected in the nmr spectrum. As mentioned before, the spin-lattice process can be described by a first order rate constant $(1/T_1)$ for the decay of the z-component of the magnetization, say, after the field is turned off, and is referred to as *longitudinal relaxation.*

Next, we shall discuss some processes that affect the xy-components of the magnetization and are referred to as *transverse relaxation processes.* The spin-lattice effect discussed above always contributes to randomization of the xy-component; therefore,

$$\frac{1}{T_2} = \frac{1}{T_1} + \frac{1}{T_2'}$$

where $1/T_2'$ includes all effects other than the spin-lattice mechanism. When the dominant relaxation mechanism is spin-lattice for both longitudinal and transverse processes, *i.e.*, $1/T_1 = 1/T_2$, then $1/T_2'$ is ignored. The quantity we used in the Bloch equations is $1/T_2$ and not $1/T_2'$. In solution, these other effects, $1/T_2'$, are small for a proton compared to $1/T_1$ so $(1/T_2) = (1/T_1)$. The other effects include field inhomogeneity, spin-spin exchange, and the interaction between nuclear moments. We shall discuss these in detail, beginning with field inhomogeneity. When the field is not homogeneous, protons of the same type in different parts of the sample experience different fields and give rise to a distribution of frequencies. This causes a broad band. The effect of inhomogeneity can be minimized by spinning the sample.

Interaction between nuclear moments is also included in T_2'. When a neighbouring magnetic nucleus stays in a given relative position for a long time, as in solids or viscous liquids, the local

field felt by the proton has a *zero frequency* contribution; *i.e.*, it is not a fluctuating field as in the T_1 process from the field of the neighbouring magnetic dipoles. A given type of proton could have neighbours with, for instance, a + + − + − combination of nuclear moments in one molecule, + − + − − in another, and so forth. Variability in the static field experienced by different protons of the same type causes broadening just like field inhomogeneity did. To give you some appreciation for the magnitude of this effect, a proton, for example, creates a field of about 10 gauss when it is 1 Å away. This could cause a broadening of 10^5 Hz. As a result of this effect, extremely broad lines are observed when the nmr spectra of solids are taken, but this effect is averaged to zero in non-viscous solutions.

The process of spin-spin exchange contributes to T_2, but not to T_1, for it does not influence the z-component of the magnetization. In this process, a nucleus in an excited state transfers its energy to a nucleus in the ground state. The excited nucleus returns to the ground state in the process and simultaneously converts the ground state nucleus to the excited state. No net change in the z-component results, but the u and v-components are randomized .

In common practice, $1/T_2'$ is never really used. Either $1/T_2 = 1/T_1$ (with $1/T_2'$ negligible) or we just discuss $1/T_2$. In a typical nmr spectrum, $1/T_2$ is obtained from the line shape (as will be shown) and it either equals $1/T_1$ or it does not.

As mentioned earlier, the true form of a broadened line is described empirically by a shape function $g(\omega)$, which describes how the absorption of energy varies near resonance according to equation (33). Since magnetic resonance lines in solution have a Lorentzian line shape:

$$g(\omega) = \frac{T_2}{\pi} \frac{1}{1 + T_2^2(\omega - \omega_0)^2} \qquad ...(35)$$

The width of the band between the points where absorption is half its maximum height is $2/T_2$ in units of radians sec^{-1}. In units of Hz, the full band width at half height is given by $1/\pi T_2$.

3.3.1 Measuring the Chemical Shift

The discussion of equation (28) indicated how the shielding constant and contributes to making nmr of interest to a chemist. We now must consider ways of making the abscissa quantitative. In the typical nmr spectrum, the magnetic field is varied until all of the protons in the sample have undergone resonance. This is

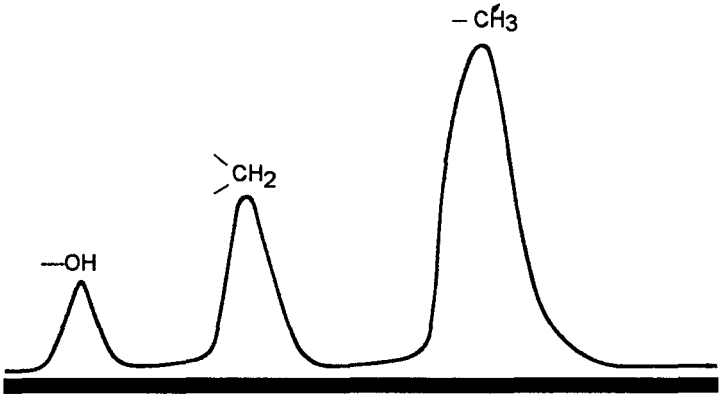


Figure 3.6 Low resolution proton nmr spectrum of C_2H_5OH.

illustrated by the low resolution spectrum of ethanol. The least shielded proton (smallest σ) on the electronegative oxygen atom interacts with the field at lowest applied field strength. The areas under the peaks are in direct proportion to the numbers of equivalent hydrogens, 1:2:3, on the hydroxyl, methylene, and methyl groups. Note that the separation of absorption peaks from $-CH_3$ and $-CH_2-$ hydrogens in this spectrum is much greater than that in the infrared spectrum.

We wish to calibrate the horizontal axis so that the field strength (or some function of it) at which the protons absorb energy from the radio frequency probe can be recorded. Equation (28) could be employed, but accurate measurement of H_N and H_0 is difficult. Instead, a reference material is employed, and the difference between the field strength at which the sample nucleus absorbs and that at which the nucleus in the reference compound absorbs is measured.

The reference compound is added to the sample (*vide infra*), so it must be unreactive. Furthermore, it is convenient if its resonance is in a region that does not overlap other resonances in the molecules typically studied. Tetramethylsilane, TMS, has both properties and is a very common reference material for non-aqueous solvents. In view of the limited solubility of TMS in water, the salt $(CH_3)_3SiCD_2CD_2CO_2^-$ is commonly used in this solvent. The position of its resonance is set at zero on the chart paper. The field strength is swept linearly, and the sweep is geared to a recorder producing a spectrum that at low resolution would resemble. The magnetic field differences indicated by the peaks are very small, and it is difficult

to construct a magnet that does not drift on this scale. Accordingly, most instruments pick a resonance and electronically adjust the field circuit to maintain or lock this peak at a constant position. This can be accomplished by having some material in a sealed capillary in the sample tube or in the instrument to lock on, or else by picking a resonance in the spectrum of the solution being studied for this purpose. The former procedure is described as an *external lock* and the latter as an *internal lock.* The internal lock produces more accurate results, for the field is being locked on a resonance that has all of the advantages of an internal standard *(vide infra).* In a typical experiment employing an internal lock, TMS is added to the sample for this purpose.

With TMS at zero, it is possible to measure the differences in peak maxima, Δ. Although the field is being varied in this experiment, the abscissa and hence the differences in peak positions are calibrated in a frequency unit of cycles per second, referred to as Hertz. This should cause no confusion, for frequency and field strength are related by equation (5):

$$\omega = \gamma H$$

According to equation (28), the shielding of the various nuclei, σH_0, depends upon the field strength, H_0. If a fixed frequency probe of 60 MHz is employed, the field utilized will be different than if a 100 MHz probe is employed. The peak separations are 10/6 as large at 100 MHz as at 60 MHz. To overcome this problem and to obtain values for the peak positions, which are independent of field strength, the chemical shift, δ, is defined as

$$\delta = \frac{\delta \times 10^6}{\text{fixed frequency of the probe, } \nu_0} \qquad \ldots(36)$$

Since the probe frequency of ν_0 is in units of MHz, and Δ has units of Hertz, the fraction is multiplied by the factor 10^6 to give convenient

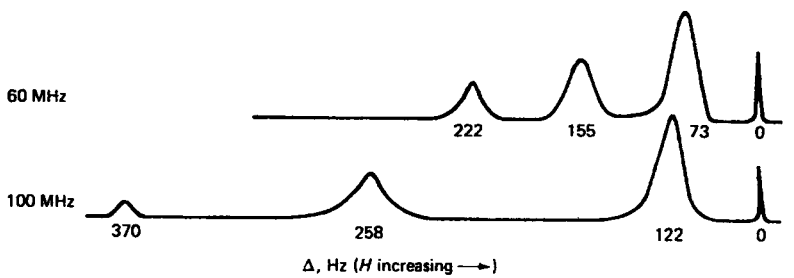


Figure 3.7 The nmr spectrum of CH_3CH_2OH at 60 MHz and at 100 MHz.

numbers for δ in units of parts per million, ppm. The chemical shift, δ, is independent of the probe frequency employed.

If the sample resonance peak occurs at a lower field strength than the reference peak, Δ is positive by convention. When $Si(CH_3)_4$ is employed as a standard, almost all δ values for organic compounds are positive and the larger positive numbers refer :0 lesser shielding. The τ scale has been commonly used in the past in organic chemistry for chemical shifts relative to TMS measured in CCl_4 as solvent. It is defined by the equation

$$\tau = 10 - \delta \qquad ...(37)$$

In recent work, its use has been de-emphasized.

For very accurate chemical shift determination, the side band technique can be employed to insure that the field is being swept linearly. In this experiment, the field axis is calibrated by displacing part of this spectrum electronically, by imposing on :' e r.f. field a fixed audio frequency field. Part of the intensity of all the lines in the spectrum will be displaced a certain number of cycles per second, equal to the audio frequency. If the audio frequency were 300 Hz, the distance on the chart paper between the original and the displaced peaks would be 300 Hz. The displaced peaks are referred to as side bands. For example, the distance between the $Si(CH_3)_4$ and CH_3I peaks is divided by the distance between one of the main peaks and its side band. Multiplication of this ratio by 300 Hz (or by whatever imposed frequency is selected) gives the value, Δ, in Hertz for the difference between the sample signal (CH_3I) and the reference signal ($Si(CH_3)_4$). The ordinate is calibrated very precisely by this procedure. By working with several peaks, the linearity of the sweep can be verified. The spectrum has been drawn to scale (a 60 MHz probe was employed), so that with a ruler and the

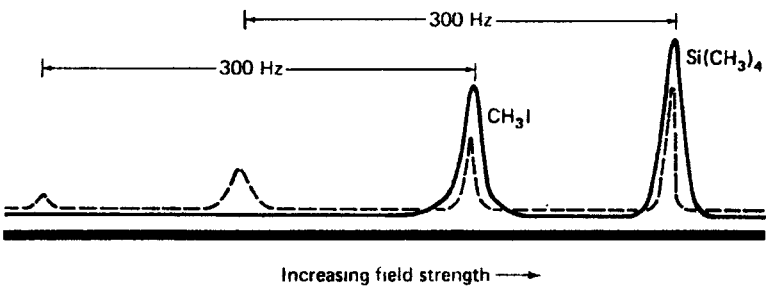


Figure 3.8 Proton nmr spectrum of CH_3I with $Si(CH_3)_4$ as an internal standard (solid line). The dashed line is the same spectrum with a 300 Hz side band.

information given, the reader should be able to calculate the value: $\delta = 2.2$.

For several reasons, it is advantageous to add the TMS directly to the solution being studied as opposed to having it in a separate sealed capillary tube, *i.e.*, an *internal* vs. an *external standard*. Magnetic field-induced circulations of the paired electron density in a molecule give rise to a magnetic moment that is opposed to the applied. In diamagnetic substances, this effect accounts for the repulsion, or diamagnetism, experienced by these materials when placed in a magnetic field. The magnitude of this effect varies in different substances, giving rise to varying diamagnetic susceptibilities. This diamagnetic susceptibility in turn gives rise to magnetic shielding of a molecule in a solvent, and is variable for different solvents. The effect is referred to as the volume diamagnetic susceptibility of the solvent. Thus, the chemical shift of a solute molecule in a solvent will be influenced not only by shielding of electrons but also by the volume diamagnetic susceptibility of the solvent. If the solute were liquid, the δ value obtained for a *solution* of the liquid relative to another liquid external standard would be different from that obtained for the *pure* liquid solute (this is often referred to as the neat liquid) to the extent that the volume diamagnetic susceptibilities of the solvent and neat liquid were different.

Because of the variation in the contribution from the volume diamagnetic susceptibility, the chemical shifts of neat liquids relative to an external standard are difficult to interpret. Problems are also encountered when substances are examined in solution. The diamagnetic contributions to the shielding of the solute depend upon the average number of solute and solvent molecules, *i.e.*, the number of solvent and solute neighbours. Consequently, the chemical shift will be concentration dependent. To get a meaningful value for δ, it is therefore necessary to eliminate or keep constant the contribution to δ from the diamagnetic susceptibility of the solvent. This can be accomplished by measuring δ at different concentrations and extrapolating to infinite dilution, in effect producing a value for δ under volume susceptibility conditions of the pure solvent. If all values for different solutes are compared in this solvent, the effect is constant.

Contributions to the measured δ from the volume diamagnetic susceptibility of the solvent can be more easily minimized by using an *internal standard*. The internal standard must, of course, be

unreactive with the solvent and sample. Under these conditions the standard is subjected to the same volume susceptibility (from solvent molecules) as is the solute, and the effects will tend to cancel when the difference, Δ, is calculated. (Due to variation in the arrangement of solvent around different solutes, an exact cancellation is often not obtained.) Cyclohexane and $Si(CH_3)_4$ are commonly employed as internal standards for protons. In order for the results from an internal and an external standard to be strictly comparable, δ for the standard as a pure liquid must be identical with δ for the standard in the solvent. It is now common practice to employ an internal standard. For accurate work, spectra at two different concentrations should be run as checks and the results relative to two internal standards compared to insure that the internal standard approximations are working.

The solvent employed for an nmr experiment should be chemically inert, and should have a symmetrical electron distribution. Carbon tetrachloride is ideal for proton resonance. Chemical shift values close to those obtained in carbon tetrachloride are obtained in chloroform, deuterochloroform, and carbon disulfide. Acetonitrile, dimethylformamide, and acetone are frequently used but can be employed satisfactorily only if there are no specific interactions (H bonding or other Lewis acid-base interactions) and if an internal standard is employed. It is also possible to measure δ in such a solvent for compounds whose chemical shifts in CCl_4 are known to determine a correction factor for the volume susceptibility of the solvent. This is then applied to the chemical shifts of other solutes in that solvent.

As seen in the above discussion, the solvent in which the chemical shift is examined has a pronounced effect on the value obtained. It is found that even the relative values of δ for the different protons in a given molecule may vary in different solvents. A quantity σ_{solv} has been proposed to encompass all types of shielding from solvent effects. This quantity consists of the following:

$$\sigma_{solv} = \sigma_B + \sigma_A + \sigma_W + \sigma_E$$

where σ_B is the shielding contribution from the bulk magnetic susceptibility of the solvent, σ_A arises from anisotropy of the solvent, σ_W arises from van der Waals interactions between the solvent and solute, and σ_E is the shielding contribution from a polar effect caused by charge distribution induced in neighbouring solvent molecule: by polar solutes (*i.e.*, an induced solvent dipole-solute interaction). As

mentioned above, σ_B is eliminated by the use of internal standards. Some indication of the importance of the other effects for a given solute can be obtained by determining δ in different solvents.

3.3.2 Simple Applications of the Chemical Shift

The main concern in this section will be with demonstrating the wide range of systems that can be studied by nmr. Only the simplest kinds of applications will be discussed. The nuclei that have been most frequently studied are ^{1}H, ^{19}F, ^{13}C, and ^{31}P. There are also extensive reports of ^{11}B, ^{17}O, ^{15}N, and ^{59}Co nmr spectra.

The range of proton chemical shifts measured on pure liquids for a series of organic compounds is illustrated. Proton shifts outside this total range have been reported, and sometimes shifts outside the range indicated for a given functional group occur. In general, the data serve to give a fairly reliable means of distinguishing protons on quite similar functional groups. Note the difference in CH_3–C, CH_3C=, CH_3–O, HC=, HCO, etc. Very extensive compilations of proton chemical shifts have been reported, which can be employed for the fingerprint type of application. Care must be exercised with –OH groups, for the shift is very concentration- and temperature-dependent. For example, when the spectrum of ethanol is examined as a function of concentration in an inert solvent (*e.g.*, CCl_4), the total change in chemical shift in going from concentrated to dilute solution amounts to about 5 ppm. When the data are extrapolated to infinite dilution, the hydroxyl proton appears at a higher field than the methyl protons, in contrast to the spectrum of pure ethanol. These changes are due to differing degrees of hydrogen bonding. There is more hydrogen bonding in concentrated than in dilute solutions. The effect of this interaction is to reduce the screening of the proton, causing a shift to lower field. This behaviour on dilution can be employed to verify the assignment of a peak to a hydroxyl group. This dilution technique has been used to investigate the existence of steric effects in hydrogen bonding and should aid in distinguishing between intermolecular and intramolecular hydrogen bonding. Solvent effects are quite large whenever hydrogen bonding or other specific interactions cur and, as will be shown later, nmr has been very valuable in establishing the existence or absence of these interactions.

For a limited number of compounds, a correlation has been reported between the electronegativity of X and the proton chemical shift of the CH_3 group in CH_3X compounds, and also between the

electronegativity of X and δCH_3–δCH_2 for a series of C_2H_5X compounds. The more electronegative X is, the less shielded are the portions. The severe limitations of such a correlation are discussed in the section on chemical shift interpretation.

Protons attached to metal ions are in general very highly shielded, the resonance often occurring 5 to 15 ppm to the high field side of TMS and, in some cases, occurring over 60 ppm upfield from TMS. The shifts are attributed to paramagnetic shielding (*vide infra*) of the proton by the filled *d* orbital electrons of the metal. The proton nmr in $HRh(CN)_5^{3-}$ is doublet, the center of which occurs at 10.6 ppm on the high field side of TMS. This splitting (*vide infra*) and shift establish :he existence of a bond between rhodium and hydrogen.

The range of fluorine chemical shifts is an order of magnitude larger than that normally encountered for protons. The difference between the fluorine resonances in F_2 and in F^-(aq) is 542 ppm, compared to the range of about 12 ppm for proton shifts. A wide range (~500 ppm) of phosphorus chemical shifts has also been reported. The fingerprint application is immediately obvious for compounds containing these elements. Before additional applications are discussed, it is necessary to consider in more detail some other effects influencing the nmr spectrum.

3.4 SPIN-SPIN SPLITTING

3.4.1 Effect of Spin-spin Splitting on the Spectrum

When an nmr spectrum is examined under high resolution, considerable fine structure is often observed. The difference in the high and low resolution spectra of ethanol can be seen by comparing.

The chemical shift of the CH_2 group relative to the CH_3 group is indicated by Δ measured from the band centers. The fine structure in the $-CH_3$ and $-CH_2-$ peaks arises from the phenomenon known as *spin-spin splitting,* and the separation, *J,* between the peaks comprising the fine structure is referred to as the *spin-spin coupling constant.* This parameter is usually expressed in Hertz. As mentioned earlier, the magnitude of Δ depends upon the applied field strength. However, the magnitude of the spin-spin coupling constant in Hz is field independent.

The cause of the fine structure and the reason for the field-independent character of *J* can be understood by considering an H-D molecule. If y some mechanism the magnetic moment of the

proton can be transmitted to the deuteron, the field strength at which the deuteron precesses at the probe frequency will depend upon the magnetic quantum number of the neighbouring hydrogen nucleus. If the proton nucleus has a spin of $+½$, its magnetic moment is aligned with the field so that the field experienced by the deuterium is the sum of the proton and applied fields. A lower applied field strength, H_0, will be required to attain the precession frequency of the deuterium nucleus in this molecule than in the one in which the hydrogen has a magnetic quantum number of $-½$. In the latter case the field from the proton opposes the applied field and must be overcome by the applied field to attain a precessional frequency equal to the probe frequency. The m_I values for the hydrogen nuclei in the different molecules that give rise to different peaks are indicated above the respective peaks. The two peaks are of equal intensity because there is practically equal probability that the hydrogen will have $+½$ or $-½$ magnetic quantum numbers. These proton resonance peaks correspond to magnetic quantum numbers of $+1,0$, and -1 for the attached deuterium nuclei in different molecules. The spin-spin coupling constants J_{DH} and J_{HD} respectively, have the same value. Subsequently, we shall discuss mechanisms for transmitting the magnetic moment of a neighbouring atom to the nucleus undergoing resonance.

Returning to ethyl alcohol, we shall examine the splitting of the methyl protons by the methylene protons. The two equivalent protons on the $-CH_2-$ group can have the various possible combinations of nuclear orientations. In case 1, both nuclei have m_I values of $+½$, giving a sum of $+1$ and accounting for the low field peak of the $-CH_3$ resonance. Case 2 is the combination of $-CH_2-$ nuclear spins that gives rise to the middle peak, and case 3 causes the high field peak. The probability that the spins of both nuclei will cancel is twice as great as that of either of the combinations represented by case 1 and case 3. (There are equal numbers of $+½$ and $-½$ spins.) As a result, the area of the central peak will be twice that of the others.

The nuclear configurations of the $-CH_3$ group that cause splitting of the $-CH_2-$ group are indicated. The four different total net spins give rise to the four peaks with the largest separation in this multiplet. The relative areas are in the ratio 1: 3: 3: 1. The separation between these peaks in units of Hz is referred to as $J_{H-C-C-H}$ or as ${}^3J_{H-H}$. The latter symbol indicates H–H coupling between three (the

superscript) bonds. The peak separation in the methylene resonance from this coupling is equal to the peak separation in the methyl group. The spectrum of the CH_2 resonance is further complicated by the fact that each of the peaks in the quartet from the methyl splitting is further split into a doublet by the hydroxyl proton. In the actual spectrum, some of the eight peaks expected from this splitting overlap, so they are not all clearly seen.

The OH peak is split into a triplet by the CH_2 protons with the same separation as J_{H-C-OH} in the methylene resonance. Usually, though not always, the effects of spin-spin coupling are not seen over more than three bonds. Accordingly, the interaction of the –OH proton with the methyl group is not seen. This "stick-type spectrum" is constructed by drawing a line for each chemically shifted different nucleus. On the next row, the effect of the largest J is shown. Additional lines are added for each J until the final spectrum is obtained.

Because of the selection rules for this process (*vide infra*), equivalent nuclei do not split each other; *e.g.*, one of the protons in the $-CH_3$ group cannot be split by the other two protons. A general rule for splitting can be formulated that eliminates the necessity for going through a procedure. For the general case of the peak from nucleus A being split by a non-equivalent nucleus B, the number of peaks, n, in the spectrum due to A is given by the formula

$$n_A = 2\Sigma S_B + 1 \qquad ...(38)$$

where ΣS_B equals the sum of the spins of equivalent B nuclei. The relative intensities of the peaks can be obtained from the coefficients of the terms that result from the binomial expansion of $(r + 1)^m$, where $m = n - 1$ and r is an undefined variable; *e.g.*, when there are four peaks, $n = 4$ and $m = 3$, leading to $r^3 + 3r^2 + 3r + 1$. The coefficients 1:3:3:1 produce the relative intensities. The Pascal triangle is a convenient device for remembering the coefficients of the binomial expansion for nuclei with $I = ½$.

```
                1
              1   2
            1   2   1
          1   3   3   1
        1   4   6   4   1
      1   5  10  10   5   1
    1   6  15  20  15   6   1
```

The triangle is readily constructed, for the sum of any two elements in a row equal the element between them in the row below.

When two groups of non-equivalent nuclei Band C split a third nucleus A, the number of peaks in the A resonance is given by

$$n_A = (2\Sigma S_B + 1)\ (2\Sigma S_C + 1) \qquad ...(39)$$

This is equivalent to what was done in the discussion of the methylene resonance of ethanol, when each of the four peaks from spin coupling by $-CH_3$ was further split into a doublet from coupling to $-OH$, leading to a total of eight peaks. When a nucleus with $I \geq 1$ is coupled to the observed nucleus, the number of lines is still given by equation (38); however, the intensities are no longer given by the binomial expansion. For instance, the hydrogen signal is split into three lines by the deuterium, for which $I = 1$. The intensities of the three lines are all equal and not in the ratio 1:2:1, since there is near equal probability that the deuterium will have +1, 0, and −1 magnetic quantum numbers. The same situation occurs with splitting from ^{14}N.

4

Autoradiography

Autoradiography is one of the most important techniques for the study of synthesis, turn-over, transport and localization of macromolecular constituents in the cell. As the structure of the tissue remains intact, it is indeed the only method which permits a direct study of the inter-relationships between cell structure and function. Autoradiography has already led to important results on nucleocytoplasmic interactions in RNA and protein synthesis and to new insights into the nature of chromosomal replication both in higher organisms and bacteria.

Autoradiography (also called radioautography) is based on the use of substances labelled with a suitable radioisotope. During the synthesis of macromolecular constituents, labelled soluble precursor is incorporated particles. A layer of photographic emulsion on the top of the tissue is employed as a radiation detector. After exposure to the β-radiation, the microscope slide and the attached emulsion are passed together through photographic developer and fixer like a photographic plate. Exposure may have to be extended to several weeks in order to obtain enough silver grains for evaluation. The location of silver grains with respect to the underlying tissue shows the location of labelled molecules in the cell at the time of fixation. The preparation of autoradiographs includes thus the following steps:

1. Incubation: Inject the radioisotope or transfer small organisms into the radioactive solution. Root tips can be grown in water containing labelled precursor. The duration of incubation depends on the purpose of the study.
2. Fixation and preparation of microtome section or tissue squashes.

3. Cover with photographic emulsion.
4. Exposure in a light-tight container.
5. Photographic processing (development, rinsing, fixing, washing, drying).
6. Microscopic study of the resulting autoradiograph.

As section and emulsion remain always in intimate contact and are studied together in the microscope, the silver grains observed in the emulsion can be related to the biological structures underneath. The labelled precursor itself and metabolic intermediate of low molecular weight are water-soluble and therefore lost during the usual procedures employed in preparing sections or tissue squashes. The radioactivity which remains in the tissue is therefore largely or exclusively due to the previous incorporation of soluble precursor into insoluble macromolecules.

The precursor which is most frequently used for autoradiographic studies of DNA synthesis is thymidine labelled with tritium (^{3}H). If this labelled nucleoside is available at the time of cellular DNA synthesis, it is incorporated together with unlabelled thymidine from the cell's own precursor pool into newly formed DNA. Its location can be demonstrated within the interphase nucleus or in the chromatids of the next or any later metaphase stage. From the distribution of chromatid labelling in the two metaphases following incubation of ^{3}H-thymidine, Taylor, Woods and Hughes (1957) could conclude that chromosomal replication follows a semi-conservative pattern. Both biochemical and autoradiographic studies played a decisive role in demonstrating that the nucleus is the main if not the sole site of DNA-dependent RNA synthesis, messenger RNA being synthesized on the chromosomes and ribosomal RNA in the nucleolus, to be subsequently transported into the cytoplasm. Most autoradiographic studies of RNA synthesis were carried out with ^{3}H-uridine, a nucleoside which is found only in RNA. Similarly, the RNA synthesis at specific loci in giant polytenic chromosomes of diptera has been studied both under normal and experimentally modified conditions by autoradiographic methods. A full complement of labelled amino acids is now also available from various manufacturers for autoradiographic studies of protein synthesis and turnover. Enzyme histochemistry is another field of application: Enzymes can be localized with the aid of a radioactively labelled substrate if the latter will not diffuse away from the reaction site. Barnard and Ostrowski (1962, 1964) as well as Ostrowski et al.

(1964) succeeded on localizing esterases by their reaction with tritiated di-isopropyl-fluorophosphate (^{3}H-DFP) which is irreversibly bound to the enzyme as a substrate inhibitor. In recent years, autoradiography has also been applied at the electron microscope level.

4.1 LABELLING, APPLICATION AND SPECIFICITY OF PRECURSORS

Radioisotopes are formed in the nuclear reactor. Tritium ($^3_1$H), a hydrogen isotope of primary importance for autoradiography, is formed by the action of slow neutrons on an isotope of lithium. The nuclear reaction is accompanied by the emission of an α-particle:

$$^6_3Li + {}^1_0n \rightarrow {}^3_1H + {}^4_2He$$

Tritium has a half life of 12.3 years. The nuclear reaction involves the emission of a β-particle and the formation of a non-radioactive isotope of helium:

$$^3_1H \rightarrow \beta^- + {}^3_2He$$

The most important radioisotopes for autoradiographic work — all β-emitters. The emission of electrons is due to neutron decay in the nucleus whose atomic number is thereby raised by one unit. Although the energy difference between the two nuclides involved in the transformation is a constant, the electrons which are emitted show a continuous energy spectrum ranging from zero to a certain maximal energy (E_{max}) which is a characteristic of the particular isotope used. Energy balance is maintained by the simultaneous ejection of an antineutrino from the nucleus. Because of their different maximal kinetic energies measured in million electron volts (MEV), the β-particles of different radioisotopes can penetrate the tissue and the photographic emulsion to different depths. This is proportional to the square of the initial energy and, roughly, inversely proportional to the density of the material through which the rays pass. The values for fixed tissues should be very similar to this. The mean ranks are, of course, considerably smaller than the maximal ranges. Thus, even the high energy β-particles from a tritium source at the lower surface of a 6μ tissue section cannot reach the emulsion above as they are deviated from a straight path by elastic and inelastic scattering. Simultaneous emission of gamma rays (^{131}I) is of no consequence for autoradiography as the majority of the energy rich-γ-quanta pass through the photographic layer without energy.

Organic molecules are labelled with radioisotopes by an exchange process. Thymidine, for example, can be labelled by catalytic exchange

with tritium oxide on reduced palladium black in acid solution. The hydrogen bound to the carbon at the number 4 position of the pyrimidine ring is preferentially exchanged for tritium in a certain proportion of the thymidine molecules. Depending upon experimental conditions, the tritiated thymidine obtained may have a specific activity as high as some curies per millimole (1c = $3.7 \cdot 10^{10}$ decays per second, 1 millicurie = 1 mc = 10^{-3}c, 1 microcurie = 1 μc = 10^{-6}C).

Some precautions are necessary when working with radioisotopes. Although weak sources such as tritium present no danger from external radiation, it must be remembered that compounds like ^{3}H-thymidine, which is incorporated into the genetic material itself, may not be so innocuous as an internal source of radiation. This is especially true for all isotopes with a long half-life. Direct or indirect contact when pipetting solutions or through smoking during work should be avoided. Work with carbon-14-requires special precautions as the experimental organisms may release $^{14}CO_2$ into the atmosphere. Particular attention should therefore be paid to local regulations on the use of radioisotopes in biological and medical research and the disposal of radioactive wastes.

Most radioactive compounds for autoradiography are sold by the manufacturer in sterile aqueous solutions with the concentrations (μc/ml) and specific activity (c/mole) indicated on the label. The bottles are capped with a rubber seal and secured against accidental opening. Desired amounts are withdrawn with a calibrated syringe. The needle of the syringe is pushed through the rubber stopper which acts as a seal when it is again withdrawn. The syringe should be handled with surgical gloves. All pipettes and glassware which were in contact with radioactive solutions are first carefully washed in running water and then sterilized.

The isotope is administered either directly or after appropriate dilution with physiological saline to experimental animals by intravenous or intraperitoneal injection. In the case of small aquatic organisms, protozoa or tissue cultures, the radioisotope is added to the culture fluid in the desired concentration. Labelled molecules can penetrate into Hydra only after the animal has been wounded by a cut. Higher plants may simply be put with their roots into the radioactive solution. In their experimental on chromosomal replication, Taylor, Woods and Hughes (1957) put growing roots for 8 hours into 1 μc/ml ^{3}H-thymidine. Sites of cellular synthesis may be localized by pulse labelling lasting only a few minutes. If the intracellular

transport of macromolecules is to be studied, a different technique must be employed. In such cases a short labelling pulse should be followed by washing in an excess of unlabelled precursor to dilute the intracellular pool of labelled material. The centre of synthesis will now lose radioactivity which accumulates at the site of transport. The connection of silver grains will usually increase with time and finally decrease again as the labelled constituent is metabolized. The optimal conditions of incubation (time concentration and specific activity of the radioisotope) have to be determined empirically by pilot experiments. Extreme dosages should be avoided to minimize the danger of radiation damage to the object during and after incubation. Not all radioactively labelled molecules can either penetrate into the cells or be used as precursors for the synthesis of macromolecular constituents. *Neurospora crassa*, for the synthesis of macromolecular constituents. *Neurospora crassa*, for example, seems to be unable to utilize thymidine, probably because the necessary enzyme for the phosphorylation of thymidine is lacking. As *Neurospora* DNA doubtlessly contains thymidine, it seems likely that a related precursor is first incorporated into DNA and then enzymatically changed to thymidine. Pilot experiments are also necessary in order to determine whether the isotope concentration and the incubation time are sufficient to obtain autoradiographs for qualitative or quantitative study after a reasonable exposure time. Aside from the rate of synthesis of the macromolecular constituent, this will depend upon the concentration and specific activity of the radioisotope and its dilution by intracellular pools of pre-existing, unlabelled precursor.

The biochemical specificity of incorporation is another question of importance. Inorganic, ^{32}P-labelled phosphate, for example, is incorporated into a variety of different compounds such as phosphoproteins as well as nucleic acids. The latter can be specifically digested with nucleases indicating that all remaining radioactivity is due to other compounds. C^{14}-labelled precursor is not only incorporated as such but also enzymatically degraded to labelled carbon dioxide and possibly other compounds which may be re-utilized for the synthesis of entirely different macromolecules. The danger of biochemical exchange of label is therefore present and should not be neglected. In studies with C^{14}-labelled nucleosides the specificity should therefore be controlled with nucleases. Thymidine is present exclusively in DNA. High labelling specificity is therefore

to be expected in this case, although exceptions do occur. Very specific labelling of RNA is obtained with ^{3}H-uridine as a precursor since this nucleoside is found only in RNA but not in DNA. Nevertheless, Bielavsky and Tencer (1960) found in amphibian eggs the label of ^{3}H-uridine in DNA instead of RNA as expected. These examples emphasize the necessity of having proper controls of labelling specificity. Most of the water-soluble precursor is already lost during fixation. Its extraction can be completed by washing the sections for 5 minutes at 4°C in 5% trichloracetic acid. The acid must be washed out carefully for several hours in three changes of 70% alcohol.

4.2 AUTORADIOGRAPHIC RESOLUTION

The smallest distance between two point sources of radiation within the tissue which yield separate autoradiographs in the emulsion layer is defined as autoradiographic resolution. That this depends on three factors, namely the distance between source and emulsion, the path length of the radiation from the particular isotope used and the properties of the emulsion. Cross-fire from the two sources leads to a loss of resolving power. As the β-rays are emitted in all directions, this is more pronounced if the sources are deep within a thick section, if the emulsion is thick and if the maximal range of the radiation is high. An emulsion consisting of a monolayer of very small and very densely packed silver bromide grains would obviously be ideal from the standpoint of resolution. With some of the special techniques used in electron microscope autoradiography, this ideal is already very closely approximated. The resulting silver grains are, however, too fine for detection with the light microscope. For conventional autoradiography, emulsion of about 5 μ thickness are used. Thick microtome sections can be the cause of poor resolution especially if the sources emit β-particles of high energy and therefore high penetration. These particles can pass for long distances obliquely through the section and cause the formation of silver grains far away from the source. To obtain optimal resolution, sections should therefore be made as thin as possible, if necessary by cutting methacrylate blocks with a glass knife, and isotopes which emit soft, short-range β-rays should be used.

The best resolution is obtained with tritium, the β-emitter of the shortest range. It is largely for this reason that tritium-labelled precursors achieved such predominance in autoradiography. On the other hand, the limited range of β-particles from tritium may lead

to misinterpretations. Even with sections of only a few micra thickness, the radiation from sources at the lower surface of the section can no longer reach the emulsion and structures located as this level may be erroneously regarded as nonradioactive. Errors due to self-absorption can only be eliminated by reducing the specimen thickness to the mean range of the β-rays.

Resolution is also limited by the grain size in the emulsion. According to Caro (1964). Iiford L4 has the lowest grain size at maximal sensitivity (~ 0.12μ). Ilford K 5 has a grain size of about 0.18μ and a somewhat lower sensitivity. It is better suited for light microscope work than L 4 whose grains are too fine for easy identification. It should be noted that grain size increases with developing time. The values above correspond to the minimal diameters after chemical development.

4.2.1 Techniques of Tissue Preparation

4.2.1.1 Fixation

In principle, any fixative can be used which will not extract the compounds to be localized by autoradiography. However, reducing mixtures should be avoided as these may lead to chemical blackening of the emulsion if washed out insufficiently. Fixation in alcohol acetic acid (3 : 1) or Carnoy's fluid has been most generally useful. Ficq (1962) states that soluble transfer RNA is retained after fixation by freeze-substitution.

4.2.1.2 Preparation of glass slides for autoradiography

To ensure good adhesion of the photographic emulsion, slides should be carefully cleaned for 1 day in a freshly prepared solution of chrome-sulphuric acid. Wash for several hours in running water and then in distilled water. Cleaned slides are dipped into chromo alum gelatine as an adhesive (dissolve 5 g white gelatin in 1 litre of warm distilled water and add 0.5 g chrome alum after cooling). The slides are put upright into a desiccator to dry.

4.2.1.3 Squash preparations and sections

Squash preparations for tritium autoradiography must be very flat to avoid excessive self-absorption of radiation within the tissue. This is of special importance for studies of nuclei and chromosomes after incubation with ^{3}H-thymidine where the β-radiation has to penetrate a layer of cytoplasm. With tissue culture material, the method of Prescott and Bender (1964) for the preparation of

metaphase chromosomes free of cytoplasm can be employed. Thin methacrylate sections are prepared, stretched and mounted following the method described. After drying, mark the positions of individual sections with a diamond pencil on the back of the slide. Methacrylate can be completely removed within 10 seconds by putting the slides into amyl acetate. With paraffin sections, special care must be taken to remove all paraffin in two changes of xylene.

4.2.1.4 Staining before applying the emulsion

Only a few staining methods can be used before the emulsion is applied. Of these, the Feulgen reaction for DNA and the aceto-orcein technique for squash preparation deserve special mention. Staining with basic dyestuffs has to be performed after photographic processing.

4.2.2 Applying the Emulsion

4.2.2.1 Stripping film method

Stripping film emulsions are about 5μ thick and are attached with a gelatine layer of 10μ thickness to glass plates (9 × 12 cm) from which they have to stripped before use. Kodak AR 10 with a resolution of about 2 μ is recommended for cytological studies. Packages containing 12 plates are stored in the refrigerator until used to keep the background of silver grains to a minimum. Plates should come to room temperature before attempts are made to strip the thin films. For stripping, use a deep red safe light in the dark room and cut rectangular pieces of film with a scalpel against the glass surface. At the extreme margins of the plate the layers are too thin to strip off easily. Do not pull the film off with forceps as mechanical tension may lead to silver grains. Depending upon humidity, the films roll up more or less when stripped off the plate. Breathe gently on the films to flatten and transfer immediately with the emulsion side down on a clean water surface. Surface tension is sufficient to smooth the films within about 1 minute. Put the slide with the specimen on its upper side into the water and lift it up. Do not pull the film off with forceps as mechanical tension may lead to sliver grains. Depending upon humidity, the films roll up more or less when stripped off the plate. Breathe gently on the films to flatten and transfer immediately with the emulsion side down on a clean water surface. Surface tension is sufficient to smooth the films within about 1 minute. Put the slide with the specimen on its upper side into the water and lift it up until the emulsion attaches to the

slide. Lift the slide completely and slowly out of the water until the film is completely attached without air bubbles. The piece of film, which should be broader than the slide, becomes smoothly attached and the free edges fold around the slide. The slide is now set upright and dried. Drying may be accelerated with a small fan. The emulsion are exposed to the radiation from the specimen in a light-tight box containing some drying agent (silica gel or grains of $CaCl_2$) in the refrigerator.

4.2.2.2 Liquid emulsions

The relative sensitivity and the grain size of different emulsion have been studied by Caro (1964). A highly sensitive emulsion for light microscope autoradiography is Ilford K5. The Kodak emulsions NTB, NTB-2 and NTB-3 which show, in this order, increasing sensitivity and decreasing grain size are altogether somewhat less sensitive than Ilford K 5 but yield better contrast for photomicrography. NTB-3 has about the same sensitivity as the Kodak stripping film AR-10 mentioned above. With increasing sensitivity the tendency to form a background due to spontaneous disintegrations of silver halide increases also.

It is very easy to cover slides with liquid emulsion. The liquid emulsion gels, which are solid at room temperature, are melted in a water bath at 42–45°C. As soon as the emulsion has reached the temperature of the water bath, two slides put back-to-back are dipped into cover the material with the photographic layer. Slides are then put upright to dry and kept for exposure in a light-tight box in the refrigerator. The thickness of the emulsion is more uniform if dry slides are dipped. If necessary, extract free nucleosides with 5% trichloracetic acid, wash in 70% alcohol, pass through 95% and 100% alcohol to dehydrate, dry in air and dip into the emulsion. The backs of the slides are wiped before the emulsion dries. The emulsion layer becomes thinner if the liquid emulsion is diluted after melting with an equal volume of distilled water. Squash preparations whose thickness is variable are best covered with undiluted emulsion.

Old emulsions contain considerably more background. Prescott (1964)) recommends dipping a slide without specimen into the emulsion and developing it right away in order to test for background. Later silver grains already present at the beginning of exposure can be erased by Caro's method (1964): Slides covered with emulsion are put into a Coplin jar containing some filter paper soaked with 3% hydrogen peroxide at the bottom. Glass rods are

put in the bottom to prevent the emulsions from actually touching the wet filter paper. The jar is closed to obtain an atmosphere saturated with H_2O_2. After 4 to 6 hours, the slides are dried in air and exposed in a light-tight container. To save emulsion, Perry (1964) recommends flat vessels for dipping. The specimen can also be mounted at the end of the slide to save emulsion. Used emulsions should be discarded.

4.2.3 Exposure Time

The exposure time depends upon the grain yield per decay event, the rate of decay, and the concentration of the radioisotope within the tissue. With isotopes of a short half life like I^{131} and P^{32}, it is useless to prolong the exposure to several times the half life as the gain in grain density becomes less and less with time. Aside from this, the correct exposure time must be determined either by a pilot experiment or by taking a slide out from time to time and developing it to check if longer exposure is needed. If possible, the isotope concentration should be adjusted to get an autoradiograph within three weeks' exposure.

4.2.4 Photographic Processing

The X-ray fine grain developer Kodak D 19b can be used for developing autoradiographs. Develop under a deep red safe light for 2–5 minutes at 18°C. Longer development leads to coarse silver grains. The slides should reach room temperature before they are put into the developer.

One litre D 19b developer contains:

2.2 g 4-methylamino-phenosulphate (commercial names: 'Elon', 'Metol', 'Photorex' (Merck)).

72.0 g anhydrous sodium sulphite (or twice this amount of the product with crystal water).

8.8 g crystalline hydroquinone

48.0 g anhydrous sodium carbonate

4.0 g potassium bromide.

Dissolve the chemicals in the order given above in about 750 ml distilled water at 35–40°C and adjust the total volume to 1,000 ml. Take care that each substance is completely dissolved before the next is added. After developing, rinse for about 10 seconds in distilled water and transfer to photographic fixer for about 10 minutes. Wash for 10 minutes in running tap water, rinse in distilled water and dry, using a small fan if desired

4.2.5 Microscopy of Autoradiographs

Autoradiographs may be simply be studied by putting a drop of immersion oil directly on the dry emulsion and using the oil immersion objective without a cover glass. This method is especially suited for pre-stained slides. For phase contrast microscopy of unstained preparations, use a drop of glycerol as a mounting medium and add a coverslip. The refractive index of hardening resins is generally too high to obtain adequate phase contrast. Glycerol can be diluted with water if the phase contrast of the specimen should still be insufficient. After removing the cover glass, the glycerol is washed out with water and the slide is dried again for storage. The grains themselves are most easily recognized in dark ground illumination, but the structure of the biological specimen underneath is better visible in phase contrast.

Autoradiographs can also be stained after photographic processing. Basic dyestuffs such as toludine blue, azure B at pH 4 or methyl green/pyronin are most suitable for this. The gelatin layer on top of the emulsion of stripping film stains very heavily but it will usually lose the stain just as readily by washing in several changes of distilled water. Clearing can be accelerated in 40-50% ethyl alcohol. More concentrated alcohol must be avoided to prevent cloudy precipitation of the gelatin. Dry the slides in air after the gelatine has been cleared of the stain and transfer directly into xylene. Permanent preparations are made with resin usual.

The depth of focus of high power oil immersion objectives is to restricted that the silver grains of the autoradiographs and the tissue underneath are not simultaneously in focus. Oil immersion objectives of somewhat lower numerical aperture and 40-54 times magnification are preferable for photomicrography. Their resolving power is somewhat less but their depth of focus is sufficient to obtain a sharp image of both the sliver grains and the cell structures. Alternatively, a 100 × phase contrast oil immersion objective may be employed when working with thin sections in the 1-μ range. As the microscope is focused on the upper level of a section mounted in Harleco synthetic resin, the silver grains are slightly out of focus and appear as bright, sharp dots in phase contrast.

4.2.6 Quantitative Evaluation

Absolute quantitation with the aim of determining the actual amount of isotope incorporated from grain counts and the known

rate of radioactive decay is fraught with so many difficulties of standardization and calibration that this approach has not been widely applied at the present time. It is simpler and often just a valuable to determine the relative amounts of radioisotope incorporated in different regions of the specimen. Both track and grain counting have been used for this purpose.

4.2.6.1 Track counting

The β-particles from all isotopes except tritium contain enough energy to activate a number of silver bromide grains in succession and thus produce regular tracks in thick emulsions which are easily distinguished from the single grains of the background. Even a few tracks which emanate from a particular source are easily recognized as such. Track counting leads therefore to very accurate results even in moderately labelled specimens whose grains would otherwise be difficult to separate from the ever present background.

4.2.6.2 Grain counting

Under certain conditions which are discussed in detail by Perry (1964) the number of silver grains per unit area is proportional to the amount of isotope incorporated. Sufficient autoradiographic resolution and negligible self-absorption within the tissue are basic prerequisites for the quantitative evaluation of autoradiographs by the grain counting method. The latter is especially important with tritium whose β-radiation has an extremely restricted range of penetration. In the case of squash preparations whose thickness varies, quantitative evaluation must be based on a knowledge of the thickness profile of the cell. However, self-absorption depends not only upon thickness, but also upon density which may be determined by interferometry or immersion refractometry. A correction can be achieved by multiplying the number of silver grains per unit area with the reciprocal of the density per unit area. The influence of thickness and density has been studied quantitatively by Maurer and Primbsch (1964). With transparent methacrylate sections, the thickness can be measured with the interfence microscope. Maurer and Primbsch found already at a section thickness of only 0.25 μ 'saturation' of grain density over the nucleolus, i.e. thicker slices of nucleolar material led to no increase in grain density because of self-absorption of the β-radiation from the lower levels. In nucleoplasm and cytoplasm, which are less dense, saturation was reached at about 1 μ. It is evident from this that most published

comparisons of grain densities after tritium labelling are based on saturation densities from layers of 'infinite' thickness as far as self-absorption within the selection is concerned. Such comparisons are valid if the thickness is 'infinite' for all structures of different densities.

The concentration of the radioactive precursor, the specific activity of the isotope and the exposure time must be properly chosen to obtain an optimal grain density for counting. This is necessary not only because dense clumps of grain can no longer be counted, but also because, during long exposures, two β-particle may strike the same AgBr-particle twice leading to only one silver grains in the developer (coincidence error). As the size of the silver grains themselves is already close to the limit of microscopic resolution, both types of error are excluded to a large extent if the maximal countable density of the most densely labelled cell regions is first determined by a pilot experiment.

Counts may be made directly over a given structure whose area is then determined by planimetry on a photomicrograph or a camera lucida drawing, or with the aid of an eyepiece reticle. Photometric methods have been used in addition to visual grain counting to measure the density of labelling. Designs which have been used with success incorporate a vertical illuminator and utilize the light reflected by the silver grains for photometric measurement. The photometric response should increase linearly with increasing numbers of grains in the measured area.

4.2.7 Autoradiography with two Emulsions

The incorporation of ^{3}H- and ^{14}C-labelled precursors can be studied simultaneoulsy if the slide is covered with two emulsion layers, making use of the difference in penetrating power of β-radiation from the two radioisotopes. The two precursors may be administered either simultaneously or successively. For example, H^3-uridine may be used to label RNA and C^{14}-thymidine to label DNA, or C^{14}-adenine and H^3-phenylalanine canbe used to label nucleic acids and proteins. The preparation is first covered with one layer of liquid emulsion and developed after exposure. The autoradiograph of this layer is due to both H^3 and C^{14}. The first layer is now covered with one layer of liquid emulsion and developed after exposure. The autoradiograph of this layer is due to both H^3 and C^{14}. The first layer is now covered with a cellodin film (dip into 1% cellodin in 1:1 Mixture of ether and alcohol) and covered with the second

layer of liquid emulsion. This is again exposed, developed and fixed. The second layer can be reached only by the energy rich-β-paricle of C^{14} and not by the H^3. The proportion of silver grains is the first layer which is due to either of the two isotopes is determined by a comparison with the grains in the second layer which are due only to C^{14}. The distinction is made easier if two emulsions of different grain sizes are used, e.g. Ilford L4 and Kodak G5. Attempts at multiple autoradiography have been made with colour film which consists of three layers. The grains appear in different colours depending upon the penetration of the β-rays.

A new method of double isotope autoradiography with two emulsions at high resolution has been published by Trelstad (1965). The two eulsions differ in their sensitivity to β-radiation. They are applied to the top and the bottom of the section respectively. A thin section of about 1 μ cut from araldite or methacrylate-embedded tissue with a glass knife on an ultramicrotome is attached to a slide covered with a layer of celloidin and dipped into liquid emulsion (Ilford K-1 or Kodak NTA) which is sensitive practically only to the low energy radiation from tritium. After drying, it is covered with double-faced Scotch tape which has a hole punched into it to fit over the section. Emulsion and celloidin are cut around the mask of Scotch tape and all three components are slowly and carefully peeled off the slide. They are then turned on a slide with a hole to be covered with a second, highly sensitive emulsion (Ilford K-5 or L-4) which can only be reached by the high energy radiation of ^{14}C because the intervening layer of celloidin which absorbs the radiation from tritium. The lower surface of the hole in the slide has to be temporarily covered with Scotch tape to prevent access of the liquified second emulsion to the tritium-sensitive layer. After processing, the 'sandwich' of the two emulsions the section and the intervening celloidin layer is transferred to a new slide by sticking it to the wet surface of the ^{14}C-sensitive emulsion.

4.2.8 Autoradiography of Water-Soluble Substance

Water-soluble materials are usually lost in the microtechnical procedure employed for specimen preparation. The usual type of autoradiograph is therefore due solely to insoluble macromolecules. A simple method to retain water-soluble substances has been developed by Miller, Stone and Prescott (1964). Tissues are fixed by freeze-drying, microorganisms are simply dried to the slide after incubation. A wire loop whose diameter must be larger than the

width of the slide is dipped into liquid emulsion to obtain a thin film. After this has gelled, the slide is slowly pushed through the loop carrying the film. Breathe gently onto the film to attach it to the slide by the moisture of the breath. Expose and process as usual.

4.2.8.1 Artifacts

Artifacts are all changes of the density of silver grains in the processed film which are not due to the β-irradiation from radioisotope incorporated into molecular constituents of the specimen.

4.2.8.2 Background

A certain amount of background is present in all emulsions. The number of background grains increases with the age of the emulsion due to spontaneous and chemically induced degradation of silver halide, both of which are temperature-dependent. Both the radioactivity of the atmosphere and even safe light exposure during photographic processing contribute to some degree to the total amount of background. Sensitive emulsions always show more background. To reduce the background to a minimum, use only fresh emulsions and store liquid emulsions and striping film plates in the refrigerator. The background can be erased before processing with hydrogen peroxide.

4.2.8.3 Chemical artifacts

A latent image can also be formed by the action of reducing chemicals on the emulsion. If this action is due to certain localized compounds within the tissue, it may lead to misinterpretations. Such effects were noted by Kulangara (1961) in sections of the rabbit uterus and in ciliates by this author. Insoluble mercury salts in the tissue due to animal experiments or sublimate fixation can also lead to chemical artifacts. These are easily recognized by comparison with a control slide of non-radioactive tissue. It is possible to avoid these artifacts altogether by covering the sections with a protective layer of celloidin or chrome alum gelatine. Both methods degrade autoradiographic resolution slightly. With isotopes of short ranges, the grain density also may be affected.

A chemical artifact of an entirely different nature can arise due to the binding of labelled amino acid to tissue protein during fixation. Peters and Ashley found that appreciable amounts of ^{3}H-leucine can be bound to protein, especially during fixation of the tissue with glutaraldehyde for electron microscope autoradiography. It remains

to be seen whether other labelled amino acids can also give rise to this type of artifact and what role this might play in light microscope autoradiography. With ^{3}H-leucine and aldehyde fixation, the results of quantitative pulse-labelling experiments can be seriously affected.

4.2.8.4 Latent image fading

If exposed emulsions are stored, the latent image will gradually fade away. In autoradiography, this would lead to a decrease of grain density, n contrast to all other artifacts. Latent image fading is therefore not as easily recognized as an increase in background. In quantitative studies it may, however, play a certain role, as the formed latent images may fade during prolonged exposure. Experimental conditions should therefore be chosen to avoid the necessity of excessively long exposure times. Vapours of oxidizing substances may also lead to fading and should be avoided during exposure.

5

Separation Through Machines

Centrifuges are designed to accelerate sedimentation by utilizing centrifugal force. Various types of centrifuges are used in the clinical laboratory for separating suspended particles from a liquid in which the particles are not soluble. Liquids of differing specific gravities (density) may also be separated. There are three general types of centrifuges: the horizontal head, the angle head, and the ultracentrifuge. Many variations of the horizontal and angle head units are found in the clinical laboratory. These include bench top and floor standing units, refrigerated units, and such special-purpose instruments as the microhematocrit, microsample, cytospin, and continuous-flow systems.

5.1 BASICS

Centrifugation is one of the most useful techniques available to the bio-chemist. Not only can it be applied to the separation of materials but it can also be used to study the physical properties of macromolecules.

Basically, centrifugation involves placing particles and their suspending medium in an applied field of *centrifugal force.* The centrifugal field causes the particles to migrate more rapidly in a direction outward from the axis of rotation.

The operation is usually accomplished by placing the particles with their suspending medium in a container near the edge of a rotating device called a *rotor.* The movement of particles or solute in a centrifugal field is called *sedimentation,* and the rate of movement is referred to as the *sedimentation rate* or *sedimentation velocity.* The material that has been sedimented to the bottom of

the container is called the *pellet* or *residue,* whereas the solution above the pellet is called the *supernatant.* The sedimentation rate is directly proportional to the cetnrifugal force, *G,* which is defined by equation

$$G = \omega^2 r = \frac{4\pi^2 (rpm)^2}{3600} r \qquad ...(1)$$

where ω is the angular velocity in radians per second, r is the distance of the particles from the axis of rotation, and rpm is the speed of the centrifuge rotor in revolutions per minute. The intensity of the centrifugal force is usually expressed in terms of the *relative centrifugal force, RCF,* where

$$RCF = \omega^2 r/g = 1.119 \times 10^{-5}(r)\ (rpm)^2 \qquad ...(2)$$

and g is the gravitational constant. A monogram, which can be used to determine the relative centrifugal force when the speed and radius of the rotor in centimeters are known, is illustrated.

There are factors in addition to the centrifugal force that influence the sedimentation rate. Among the more significant are the shape of the particles being sedimented, their radius, the difference between their density and that of the medium in which they are suspended, and the viscosity of the suspending medium. These factors have the following relationship with the sedimentation velocity:

$$s = \frac{(2r_p)^2(\rho' - \rho)}{18\eta(f/f_0)} \qquad ...(3)$$

where s is the sedimentation velocity, r_p is the average radius of the sedimenting particles, ρ' is the density of the particles, ρ is the density of the suspending medium, η is the viscosity of the medium, and f/f_0 is the friction factor which represents the ratio of the sedimentation velocity of a hypothetical spherical particle compared to the velocity of the particle being sedimented. Any external changes which influence the values of these factors must be accounted for in considering the sedimentation velocity. Thus, a change in temperature might be expected to alter significantly the viscosity of the solution, or a change in the concentration of the buffer in the medium might change the density.

One other internal factor which affects the behaviour of particles in a centrifugal field is their electrical change. The presence of charges on particles of biological interest is fairly common. These

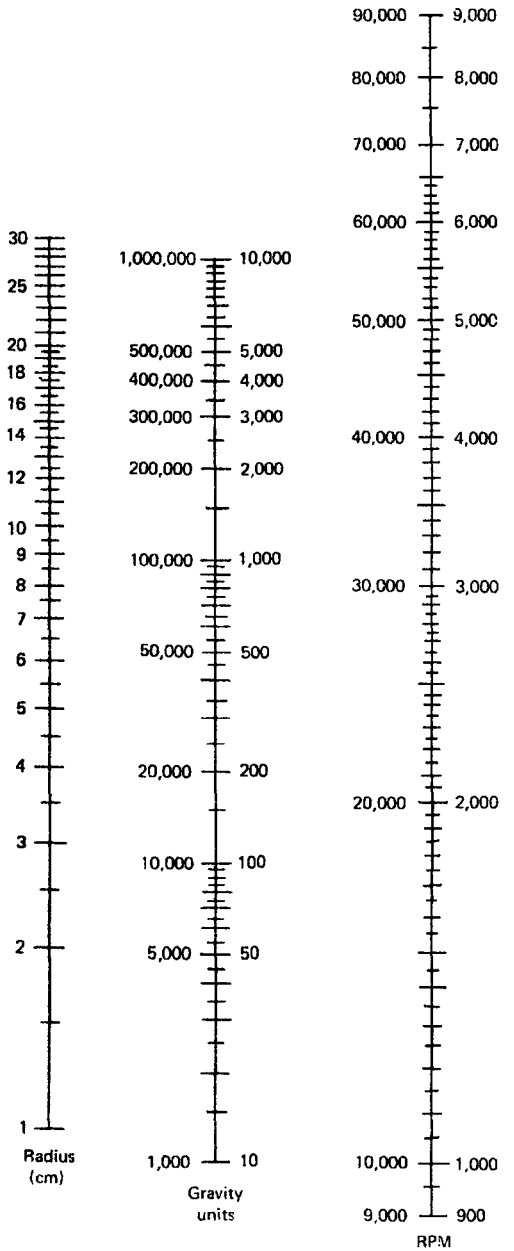


Figure 5.1 Nanogram for the determination of the gravitational force of a centrifuge rotor.

charges decrease the rate of sedimentation and increase the rate of the particles diffusion; however, the opposing forces do not cancel each other. Therefore, in situations where the charge effects alter sedimentation properties, as for example in determinations of the molecular weight of macromolecules, an electrolyte is introduced into the supporting medium at a concentration sufficient to reduce the electrical field strength.

In expressing sedimentation rates the term *sedimentation coefficient* is employed and is usually expressed at standard conditions. By convention, the coefficient is converted to standard conditions of water at 20°C and is denoted by the term $S_{20,w}$. A further description of this term may be found in the references dealing with ultracetnrifugation. The *molecular weight* of a solute can be determined by sedimentation velocity methods when the value of the sedimentation coefficient is known. The valve for the molecular weight, *M*, can be calculated from the well-known *Svedberg equation*

$$M = \frac{RTs}{D(1-\nu\rho)} \qquad ...(4)$$

where R is the gas constant; T is the absolute temperature; s is the sedimentation coefficient; D is the diffusion coefficient of the solute; $1-\bar{\nu}\rho$ is the buoyancy term where $\bar{\nu}$ is the partial specific volume of the solute and ρ is the density of the solvent. Sedimentation coefficients of spherical particles vary as the two-thirds power of their molecular weights (that is, $M^{2/3}$).

The molecular weight of macromolecules also can be determined by using a method called *sedimentation equilibrium.* This method uses a relatively low centrifugal force and takes advantage of the competition for the macromolecules between the forces of sedimentation and the forces of diffusion. In practice, it is usually necessary to centrifuge the sample until a steady state of concentration is reached by the macromolecules in the centrifuge sample holder. The molecular weight M may then be calculated from equation (5).

$$M = \frac{2RT\ln(c_2/c_1)}{\omega^2(1-\bar{\nu}\rho)(d_2^2-d_1^2)} \qquad ...(5)$$

Here c_1 and c_2 are the concentrations of the macromolecules at distances d_1 and d_2 from the axis of rotation, ω is the angular velocity, $\bar{\upsilon}$ is the partial specific volume, and ρ is the density of the solution.

Because the time required to reach the steady state in the sedimentation equilibrium method is often very long, a modification, in which the equilibrium condition is approached, often is used. This modification is called *approach-to-equilibrium*. Additional information on these techniques may be found in any of the references on ultracentrifugation.

A knowledge of the time required to sediment particles often saves valuable time and energy. This *sedimentation* or *precipitation time* can be calculated provided that one knows the sedimentation coefficient or the size of the particles ad the speed and dimensions of the rotor. Equation (6) is useful in the calculation of the precipitation time.

$$t_s = \left(\frac{\overline{ST_1} - \overline{ST_2}}{s} \right) \left(\frac{(rpm)^2_{max}}{(rpm)^2} \right) \qquad ...(6)$$

Here t_s is time in hours; $\overline{ST_1}$ and $\overline{ST_2}$ are centrifugation functions for a given rotor for the volumes of solution containing the solute at the beginning and at the end of the centrifugation, respectively; s is the sedimentation coefficient; $(rpm)_{max}$ is the maximum possible speed of the rotor; (rpm) is the speed at which the rotor is operated.

$\overline{ST}$ can be calculated for any rotor or centrifuge-tube combination as a function of the volume of solution in the tube according to equation (7),

$$\overline{ST} = \frac{10^{13}}{4\pi^2} \left(\frac{\ln(r_b / r)}{(rpm)^2_{max}} \right) \qquad ...(7)$$

where 10^{13} is the conversion value of the Svedberg unit; r_b is the distance from the axis of rotation to the bottom of the centrifuge tube; r is the distance from the axis of rotation to a given level in the tube. On the basis of equation (7) a graph such as may be constructed to relate the volume traversed by the solute to the $\overline{ST}$ in any given rotor.

Using equation (6) and the graph, the sedimentation time can be calculated for the clarification of a hypothetical 12-ml suspensions containing particles with a sedimentation coefficient of 200S in a Spinco No. 40 angle rotor at 20,000 rpm. From the graph it is apparent that a 12-ml traverse in a No. 40 rotor corresponds to a $\overline{ST}$ value of 102. By inserting the appropriate values into equation (6) one obtains

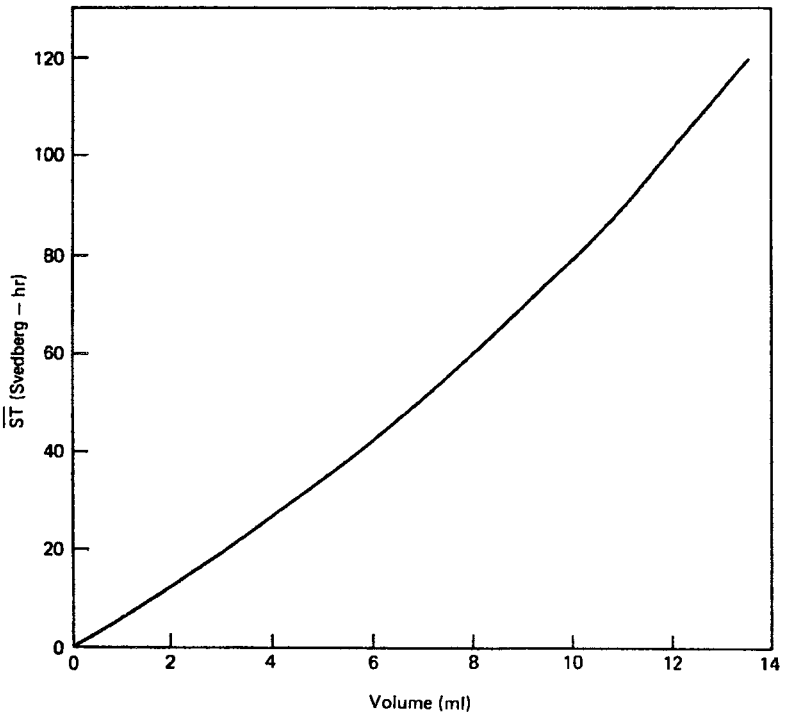


Figure 5.2 Rotor characteristics for the rapid computation of $\overline{ST}$ values for given volumes traversed by particles in Spinco No. 40 rotor having a maximum speed of 40,000 rpm.

$$t_s = \left(\frac{102}{200}\right)\left(\frac{40{,}000}{20{,}000}\right)^2 = 2.04 \text{ hours} \qquad \ldots(8)$$

If a speed of 40,000 rpm had been used for the sedimentation, the clarification could have been accomplished in 0.51 hours or about 31 min.

Occasionally a term called the *performance index,* P_i is used to express the relative performance of rotors in the complete sedimentation of a given material under idealized conditions. The value for this term can be calculated from equation (9).

$$P_i = \frac{rpm^2}{\ln r_{max} - \ln r_{min}} \qquad \ldots(9)$$

Here r_{max} and r_{min} are the distances in centimeters from the axis of rotation to the bottom of the tube and to the meniscus, respectively. The performance index can be utilized to calculate the precipitation

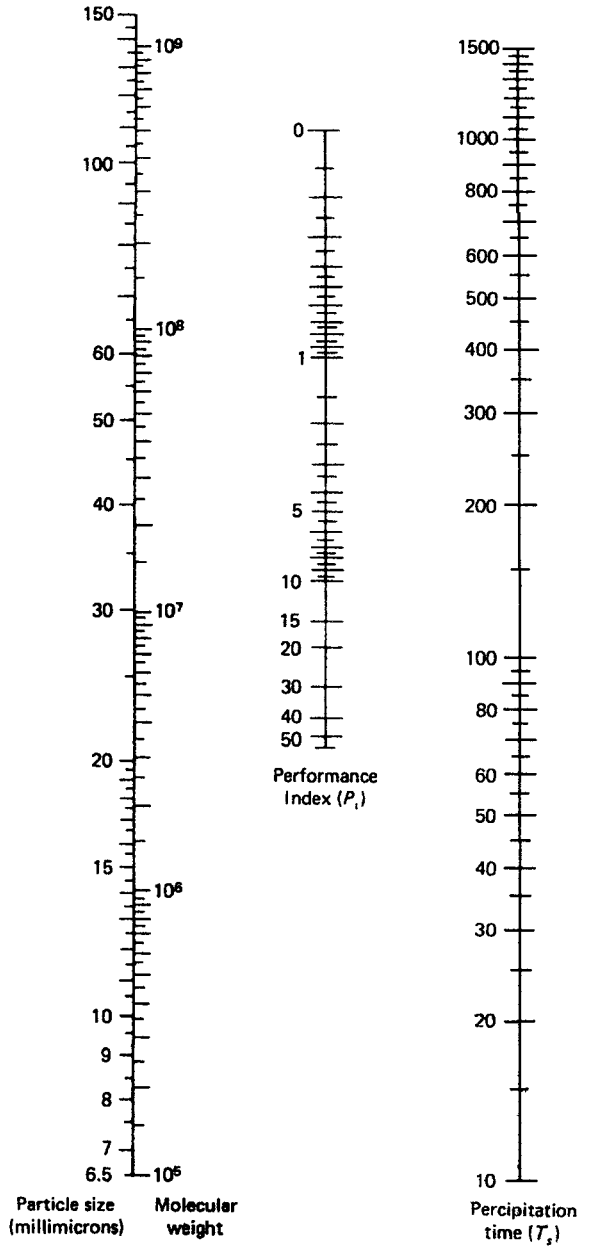


Figure 5.3 Nanogram for the determination of precipitation time when the performance index of the rotor and the size of the sedimenting particles are known.

time provided that the size or the molecular weight of the particles is known.

A knowledge of precipitation times is of great importance in the separation of subcellular particles which differ significantly in size. Several centrifugation runs using various compilations of speeds and times yield relatively homogeneous fractions of particles. This procedure is known as *differential centrifugation.*

5.2 ROTORS

Centrifuge rotors are usually designed either for preparative or analytical purposes. Most analytical rotors are constructed so that light, which is used to determine the position of the solute boundary, can pass through the sample and its container while it is spinning in the rotor. Details of analytical rotors are beyond the scope of the present text but may be found in most of the references to ultracentrifugation.

Preparative rotors are of three principal types: (a) angle, (b) swinging-bucket, and (c) zonal. Angle and swinging-bucket rotors are much more common than the zonal type. In the angle rotor, containers, which may be either tubes or bottles of glass, metal, or plastic, are inserted into cavities in the rotor that are constructed at an angle to the axis of rotation. In the case of the swinging-bucket

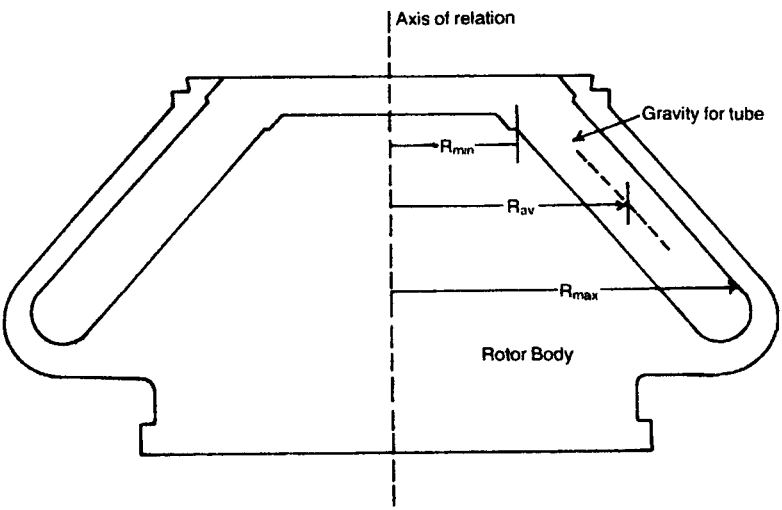


Figure 5.4 Cutaway diagram of an angle centrifuge rotor. R_{min}, R_{av}, and R_{max} represent the minimum, average, and maximum radii, respectively, from the axis of rotation of the rotor.

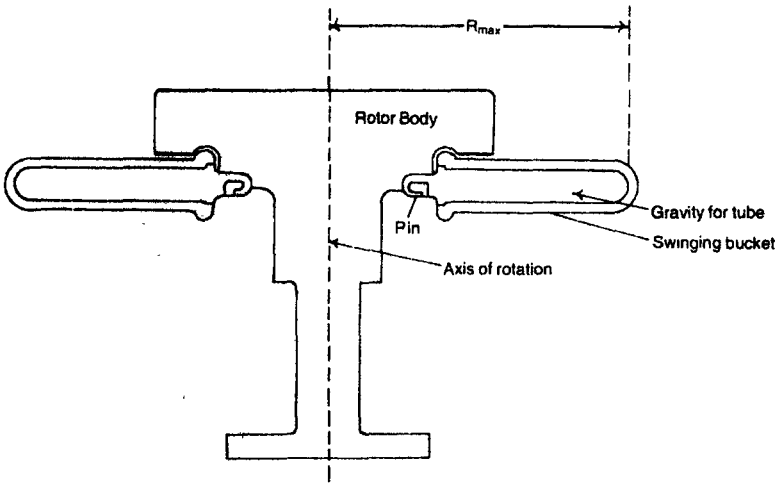


Figure 5.5 Cutaway diagram of a swinging-bucket rotor. R_{max} represents the maximum radius from the axis of rotation of the rotor.

rotor, the container is placed in a holder which is itself held to the rotor body by means of pins that allow it to swing. The vertical axis of the container is parallel to the axis of rotation when the rotor is at rest, but as it begin to spin the container and its holder swing out because of the centrifugal force.

At operational speeds the container is usually perpendicular to the axis of rotation. Zonal rotors which effectively separate particles of different sizes or densities have been developed recently. In these rotors the particles are sedimented through a gradient liquid in sector-

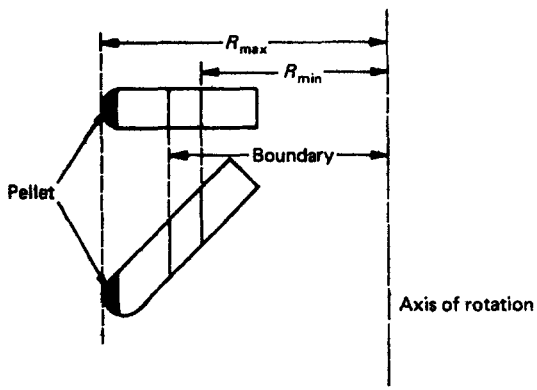


Figure 5.6 Relationship of R_{max} and R_{min} to the axis of rotation in swinging-bucket and angle centrifuge rotors.

shaped compartments. A clear description of this process may be found in the paper by Anderson [1]. In addition to the three types of rotors mentioned above, a special type of rotor is used in preparative centrifuges manufactured by the Sharpless Equipment Co. This type of rotor, which is a hollow stainless-steel cylinder, has a bottom fitted with crossed baffles.

As the rotor spins in the centrifuge, the solution containing the particles flows from the bottom across the baffles into the interior of the rotor. The vortex of liquid created by the spinning baffles causes the heaviest particles to sediment onto the lower wall of the rotor whereas the lighter particles sediment higher on the wall. If the speed and flow rate are controlled properly, the solvent and the lightest particles flow out through holes in the top of the rotor. This type of rotor is very effective for *continuous-flow centrifugation* and can be used to harvest large batches of microorganisms from their liquid cultures and subcellular particles from large batches of solution.

Angle centrifuge rotors have also been adapted for continuous-flow centrifugation, and several models are available commercially. All rotors must be designed so that they do not come apart under the stress of centrifugal fields. They should be constructed of a strong metal and should be machined so that their opposite sides strike a balance in weight. Those rotors which are used for very high speeds must be stronger and better-balanced than those used at low speeds. In addition, the metal used for the rotor should be protected with a resistant coating.

5.3 TYPES OF CENTRIFUGES

Most students of biochemistry are already familiar with standard bench-top laboratory centrifuges. These instruments are used in most laboratories for routine sedimentation work when refrigeration or high speeds are not required. The standard laboratory or clinical centrifuges are capable of reaching speeds up to about 5000 rpm with forces of about 3500 times the accelerating of gravity ($\times$ g) depending upon the particular rotor and centrifuge. They employ either angle or swinging-bucket rotors, most commonly for 50-ml or 12-ml and 15-ml tubes. Certain bench-top centrifuges can attain speeds up to about 16,000 rpm. Other laboratory centrifuges are capable of attaining higher speeds than the bench-top models. Many of these are equipped with refrigerated rotor chambers which are essential for most work involving enzymes.

These centrifuges are usually available for use with many different rotors that offer a wide choice of speeds and sample volumes. The most highly refined centrifuges are called ultracentrifuges. These instruments are available in both preparative and analytical models although, with the proper techniques, preparative models can be used for analytical purposes.

Ultracentrifuges are capable of attaining very high speeds, up to about 75,000 rpm, with centrifugal forces of up to about 500,000 × g. Since such speeds result in considerable heating of the rotor, because of the friction with air, ultracentrifuges are usually constructed with a vacuum rotor chamber and refrigeration. Rotors for use at the highest speeds are constructed of titanium. Analytical models of the ultracentrifuge are equipped with optical systems to measure the position of the solute boundary during centrifugation. These systems utilize refractive or absorptive properties of the sample to make measurements.

Many analytical ultracentrifuges are also equipped with controls which enable them to be operated at carefully regulated slow speeds so that solutes may be studied by the approach-to-equilibrium technique. These systems and methods of analytical ultracentrifugation are discussed extensively in the references to ultracentrifugation found at the end of this chapter. Several different centrifuge models are designed for or are modified for continuous-flow sedimentation. These either utilize angle rotors or the Sharpless cylindrical type of rotor discussed earlier. Most centrifuges are equipped with electrical motors to propel the rotors. A few centrifuges utilize compressed air or steam for this purpose.

5.4 CARE OF ROTORS AND CENTRIFUGES

Before using any centrifuge read the operational directions or obtain instructions from an experienced operator. Observe the following rules:

1. Do not overfill centrifuge tubes or bottles.
2. Balance the containers with their solutions before placing them in opposite cavities of a centrifuge.
3. Check the rotor to see that it is empty and undamaged before beginning a centrifuge run.
4. Do not exceed allowable speeds for a given rotor.
5. Turn off the centrifuge power immediately if any irregular noise or vibration occurs.

6. Keep rotors and rotor chambers clean. Chambers can be cleaned with a sponge, warm water, and a mild detergent; rotors can be cleaned with a bristle brush, warm water, and a mild detergent followed by rinsing with distilled water and drying by resting them upside-down.
7. See that the centrifuge is maintained according to instructions.

Some rotors for ultracentrifuges suffer from metal fatigue with prolonged use. Therefore, after a certain amount of use their maximum allowable speed must be decreased, a procedure called derating. For this reason, a record of their use must be maintained. Instructions on derating are provided with these rotors.

5.5 DENSITY GRADIENT TECHNIQUES

During centrifugation convective disturbances often affect the migration of particles in a sedimenting system. These disturbances arise from differences in the density of the solution caused by migration of particles, by temperature gradients across the tube, and by a torque placed on the solution during deceleration. Convective disturbances can be counteracted by placing the particles in a solution that increases in density with increasing distance from the axis of rotation. This technique is called *density-gradient centrifugation.* At present, there are three main types of density-gradient techniques: (a) stabilized moving-boundary centrifugation, (b) zone centrifugation, and (c) isopycnic gradient centrifugation.

The stabilized moving-bounary method in the analytical ultracentrifuge is similar to classical techniques for determining sedimentation coefficients. The coefficient is determined by observing the speed of migration of the solute boundary using the optical system of the centrifuge. In the case of the preparative ultracentrifuge a swinging-bucket rotor is used, and the sample is separated into fractions in which are then assayed to determine the position of the boundary. In this case the stabilizing density gradient is a shallow gradient that simply stabilizes the system against convection; the sample is distributed throughout the gradient at the start of the centrifugation.

In the case of zone density-gradient centrifugation, particles with a similar sedimentation rate are sedimented into a layer or zone. For this type, the solution containing the sample is layered on top of a steep-gradient solution. A steep gradient is necessary because the layer of sample material at the top creates an unstable negative

gradient just below the zone of particles which causes convective disturbances. The steep gradient reverses the negative gradient in a very short distance. A swinging-bucket rotor is used for this type of centrifugation. After centrifugation, each substance, now separated into its own zone or layer, can be removed separately for further analysis and for determination of its sedimentation rate. This has been a very popular type of density-gradient centrifugation.

In practice the gradient has most often consisted of buffered sucrose solution varying in concentration from about 5 to 20%. Other materials such as CsCl have also been used to form the gradient. At the end of the centrifugation run, the contents are divided in fractions by a suitable technique such as puncturing the bottom of the centrifuge tube with a needle and catching the drops in appropriate containers for further assay. With isopycnic gradient centrifugation, the separation of particles is based solely on their density. For this reason the gradient solution must cover the density range of all the particles in the sample, and the centrifugation must continue until all the particles have reached a density equilibrium with the surrounding solution. This type of gradient centrifugation results in particles of lower density migrating to near the top and particles of higher density migrating to near the bottom of the gradient solution.

In practice, the sample solution is usually distributed throughout the gradient solution at the beginning of the run although it is possible to start with the sample as a layer either on the top or on the bottom of the gradient solution. Either discontinuous or continuous gradients can be constructed. A discontinuous gradient is created by pipetting solutions of decreasing density to form layers on top of each other in a suitable centrifuge tube. Continuous gradients, which usually yield better results, can be prepared by using suitable mixing devices.

The sedimentation coefficient of a macromolecule can be determined using density-gradient centrifugation by comparison with another similar macromolecule of known sedimentation coefficient. Several proteins and nucleic acids are available commercially for this purpose. Both the unknown and the standard are sedimented in the same centrifuge tube, and the coefficient is calculated from relationship (10).

$$s_{20,w}(\text{unknown}) = s_{20,w}(\text{standard})\frac{d\ (\text{unknown})}{d\ (\text{standard})} \quad \ldots(10)$$

where d (unknown) and d (standard) are the distances travelled from the meniscus by the unknown and by the standard, respectively.

5.6 FUNCTION AND CONTROL

Daily inspection along with periodic function verification and proper preventive maintenance are vital to the efficiency and longevity of the centrifuge. A regular schedule for checks must be established and followed to insure proper operation of the instrument.

5.6.1 Daily Operation

Daily operation should include observation and inspection of the following:

1. The centrifuge should not be on the same circuit as sensitive electronic measuring devices such as spectrophotometers, since it generates electrical noise and has a high current drain at start-up.
2. Check the cleanliness of the chamber, and immediately clean up all spills. Be aware of biohazardous materials (microbiologic, radioactive, chemically) which require specific decontamination procedures.
3. Always balance the load of the centrifuge before operating: use the correct tube sizes and types for the particular centrifuge.
4. Always insure that the cover is closed and latched while the unit is operating. This will prevent the dangerous scattering of both biohazardous and physically dangerous material out into the environment where the operator and others may be exposed.
5. Observe for unusual noises or vibrations during operation.

5.6.2 Function Verification

Frequency of function verification procedures should be appropriate for the application of the centrifuge. It is generally recommended that procedures be performed at 3-month intervals with more frequent checks on those units with critical application. All checks performed and data gathered should be recorded in such a fashion that the information is readily available to the operator. Correction factors for speed, timing, or temperature should be posted on the unit. Tolerance limits for function checks should be established. The limits set will again depend on the application of the centrifuge. Include information for corrective action should tolerance limits be exceeded. Function verification procedures should include the following:

5.6.2.1 rpm calibration

One of the best ways to check the function of a centrifuge is to check its speed. Depending on how the unit is equipped, both the speed control and built-in tachometer should be checked with an external device. This may be accomplished with either a strobe light or mechanical or electronic tachometer of good accuracy. Several speeds used regularly on the unit should be checked. Values obtained with the external measuring device should agree within 5% of those with a built-in tachometer.

5.6.2.2 Timer

The timer should be set for common timining travels and these checked against an accurate stopwatch or electronic timer. General laboratory centrifuges should be accurate to 10% of the total timed interval.

5.6.2.3 Temperature

The thermometer on refrigerated units should be checked against a certified thermometer, and a correction factor should be derived if necessary.

5.6.3 Objectives

Preventive maintenance should be performed on the same time schedule as are function verification procedures. The preventive maintenance schedule should include the following:

5.6.3.1 Lubrication

Depending on the type of centrifuge, bearings on the upper and lower end of the motor shaft may be permanently lubricated or sealed. If this is not the case, then the manufacturer's instructions must be followed for lubrication. Bearing wear may be checked at this time by determining the amount of side play in the shaft.

5.6.3.2 Motor components

Brushes should be removed and checked, for wear. Replacement is recommended if they are worn to more than one-half their original length. When reinserting used brushes, replace them in the same orientation, and be certain that spring tension is adequate to maintain good contact with the commutator. The condition of the commutator should be examined. In order to avoid electrical arcing, the commutator and brush holders must be free of dirt, oil, and dust. If the commutator is scratched or scored, it will have to be removed and machined smooth with a lathe. On refrigerated units the

manufacturer's instruction manual should be consulted for maintenance procedures on the refrigeration system.

5.6.3.3 Electrical integrity

Both grounding resistance and current leakage should be checked periodically on each centrifuge. The fuse should be checked for proper rating; if the unit has a circuit breaker, it should be checked for proper operation. In addition, the line cord, plug, lamps, and wiring should be examined for defects.

5.6.3.4 Mechanical integrity

If the unit is equipped with a safety interlock, this device should be checked for proper working order. Gaskets, lathes, hinges, and control knobs should be examined to determine that all are functioning and in good condition. The head and shields or carriers should be examined for signs of mechanical stress (cracks) and for cleanliness and balance.

6

Photoseparation

Because of the similarity of function and use, both spectrophotometers and photoelectric colorimeters will be presented here. These instruments that are similar in design with respect to having many identical components can be markedly different in application and versatility. The colorimeter, which utilizes a filter as a monochromator, has proven to be a most rugged and stable instrument requiring very little maintenance. However, these systems are of limited or specialized usage because of the fixed wavelength of the monochromatic system. This wavelength can be changed only by changing the filter.

The spectrophotometers, on the other hand, are more versatile but require much more maintenance. This is because the condition of the light source and the dispersing element is directly related to the quality of the monochromatic light; therefore this system requires closer monitoring and more maintenance. However, the dispersion element enables the selection of a continuously variable source of monochromatic light with either a simple manual or automatic adjustment. It seems that as the efficiency of the monochromatic system to produce a narrow half-band pass with a relative high percentage of light transmitted increases, the complexity of the associated optics and light source also increases. As the sensitivity of the detectors increases, the requirements for a more stable power supply and meter system also increase. Because of the interrelation of the aforementioned, it must be understood that a good knowledge of the component parts, and their function as a part of that system, is a prerequisite to understanding the system. With a knowledge of

these systems, principles of operations, and of the laws that govern the absorption of light, the student is more able to cope with the problems of photometry.

6.1 BASICS

In relation to analytical chemistry, photometry refers to the measurement of the light-transmitting power of a solution in order to determine the concentration of light-absorbing substances present within. Photometry can, of course, be applied to measure the transmission of energy in the ultraviolet, infrared, and visible regions of the radiant energy spectrum. Instruments that are used to measure transmittance at various wavelengths are called spectrophotometers or photoelectric colorimeters, depending upon certain essential differences in construction, particularly the method of producing monochromatic light.

In all such instruments, monochromatic light is passed through an absorbing column of an often coloured solution of a fixed depth and directed upon a photosensitive device which converts the radiant energy into electrical energy. The current produced under these conditions is measured by means of a galvanometer or a sensitive voltmeter. Absorbance as measured in photometers involves not only the absorbance of the solute in the solution being evaluated, but also all of the molecules of the liquid through which the light passes.

It is necessary, therefore, to adjust the instrument by means of a "blank". This blank is prepared by placing in the absorption cell (cuvette) all of the constituents of the unknown solution (solvents, reagents, and so forth), but under conditions that will not permit the colour reaction to take place. The blank solution is used to set the meter of the instrument at a fixed point (100% T or zero absorbance). After adjusting the meter, the cuvette containing the unknown is placed in the instrument and read.

By this means, the absorbance of the reagents used can be cancelled. It follows that the reading of the meter with the unknown cuvette in place is a measure of the amount of absorbance of monochromatic light by the unknown. The greater the number of molecules or ions of absorbing substance present, the greater is the absorption of light. In other words, the deeper the colour, the greater is the deflection of the galvanometer from its original setting. Thus, the concentration of absorbing component present in a solution may be accurately measured by a photometer, provided that the necessary monochromatic light is used.

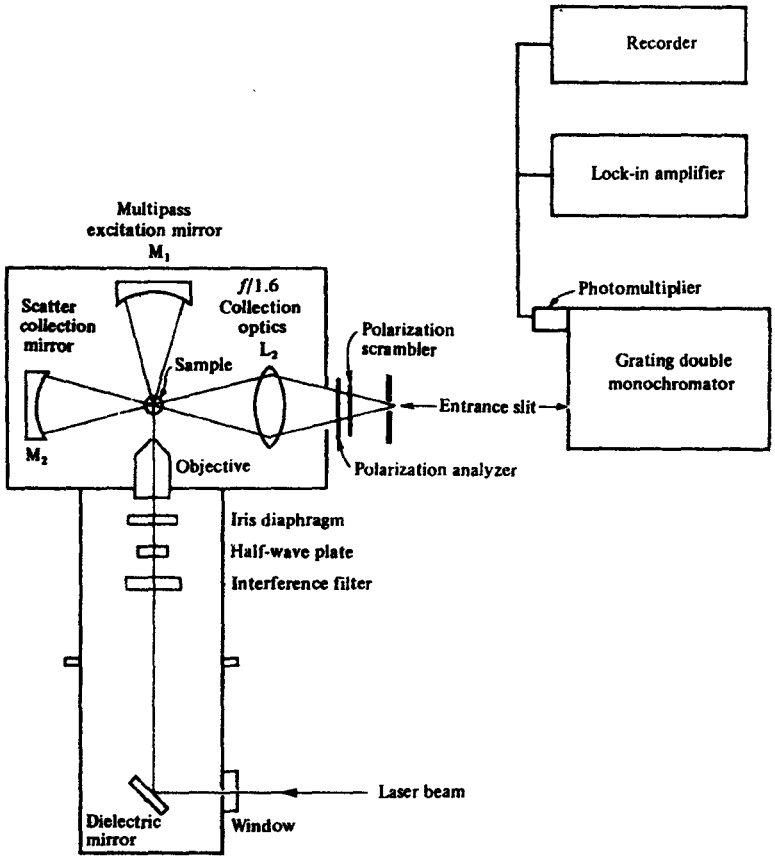


Figure 6.1 Schematic diagram of a Raman spectrometer.

6.1.1 Beer's Law

Although the laws of photometry have been referred to as "Beer's Law" (for convenience sake), they are actually laws derived by five different individuals. Bouguer showed (1729) that when light passes through an absorbing body, there is always the same difference between the logarithms of the amount of light entering and light leaving any section of the same thickness; or, as Lambert expressed it (1760), if one-half the amount of light entering leaves the first section, then ½ × ½ of the original light will leave the second section.

Beer (1852) found that one solution of copper sulphate will absorb the same amount as another if the concentration of the first

is twice that of the second and the length of the light path of the first is one-half that of the second. These laws then state that light, in passing through a coloured medium, is absorbed in direct proportion to the amount of the coloured substance in the light path. Thus, the strength of the observed "colour" is directly proportional to the concentration of the absorbing chromagen in solution.

If we let P equal the transmitted energy and P_0 the incident energy, then the ratio P/P_0 represents the transmittance (T) of the absorbing material. If the material does not absorb at all, then P and P_0 are the same, and $P/P_0 = T$, and $\%T = 100$. By using the negative logarithm of T, measurements are transformed to the energy absorbed $1/T$. The equation now becomes $A = -\log T$; since 2 is the log of 100, the $\%T$ formula becomes $A = 2 - \log \%T$. In practice, we adjust the photometer to read 100% T with a cuvette containing a blank solution. We then substitute a standard or unknown solution and read the meter. Thus, we may use $\%T$ and semilong paper or A with regular (Cartesian) graph paper to get a straight line when plotting against concentration as the abscissa. When a straight line is obtained on one of the above graphs, we may also calculate the unknown from the formula: Concentration of the unknown equals A of unknown divided by A of standard times the concentration of the standard.

6.2 PARTS

6.2.1 Circuit

The commercial power supply is adequate for most general uses and the usual fluctuations of voltage and current do not significantly upset the equilibrium of lamps, heaters, or motors. However, on stable operation of light-measuring instruments is possible without better control of the power source. Therefore, a power supply capable of furnishing adequately regulated electrical energy for the particular need of the instrument must be utilized. These power supplies fall into three general categories: batteries, voltage-regulating transformers (Sola), and electronic power supplies.

Batteries, either wet cell or dry cell, produce quite stable voltages and are relatively inexpensive. Dry cell batteries are mainly used for the operation of some small photometers, Wheatstone bridges, and radiation detectors, or wherever probability is necessary and a low direct current supply can be used. The wet cell (lead-acid) attery,

as used in automobiles, also produces a fairly stable current supply. However, it is also limited to being a source of low voltage and direct current. In addition, it requires regular maintenance consisting of frequent recharging and addition of water. After charging, the battery must go through a short period of discharging before stable current is obtained.

The constant-voltage transformers (such as Sola) will regulate power output very closely as long as the input remains between 95 and 125 V and at 60 Hz (cps). These devices are essentially trouble-free and have no moving parts. The electronic power supplies will give more precise regulation than can be easily obtained from the "Sola-type" transformer. Situations may be encountered where frequency variations must be taken into account. This necessitates the use of an electronic device to regulate and supply the desired voltage. These devices are very complex and quite expensive. They utilize either vacuum tube or solid-state regulators. The latter are becoming much more common and it is claimed that they are more dependable. However, it should be remembered that just about anyone can change a vacuum tube, but no inexperienced person should even try to tinker with solid-state components.

6.2.2 Sources of Energy

The function of the light source is to provide incident light of sufficient intensity for measurement. For work in the visible, near infrared, and near ultraviolet regions, the most common source is the glass-enclosed, tungsten filament, incandescent lamp. These lamps with prefocused bases are most useful with respect to easy replacement in an optical system. However, when used with a grating or prism monochromator, wavelength calibration must be checked when lamps are changed. Although these lamps produce a continuous spectrum over a wide range, they do have some shortcomings. Most of their energy is emitted in the near-infrared region of the spectrum, about 15% in the visible, when operated normally.

Operating temperatures can be increased to produce a greater percentage of short-wavelength energy, but this drastically shortens lamp life. For work in the ultraviolet region, the high-pressure hydrogen (or deuterium) discharge lamp is adequate from 200 nm to an upper limit extending to about 375 nm. At longer wavelengths the emission is no longer continuous. When deuterium is used instead of hydrogen, the light intensity is tripled. For very high levels of

ultraviolet illumination, the xenon arc or high-pressure mercury vapor lamp provide a large amount of continuous radiation plus high energies at the spectral lines of these elements. These lamps become very hot in operation and may even require thermal insulation, with or without auxiliary cooling, to protect the surrounding components. Although these special lamps are available and can be used, they usually require special power supplies and mountings. When planning to use these types of energy sources, it is better to purchase the equipment as a unit so that the light source will be properly set up and the necessary insulation will be in place.

6.2.3 Monochromator

When measuring the absorption or emission of radiant energy by an absorbing solution, it is necessary to be able to isolate the desired wavelength of that energy and exclude the rest. In other words, by restricting the band of wavelengths passing through the sample to those absorbed by the substance of interest, the sensitivity of the instrumental measurements to concentration changes is greatly enhanced. Thus the important characteristics of a monochromator (dispersing device) are its bandpass width, the nominal wavelength and peak transmittance. There are numerous ways of isolating the desired spectrum.

The simplest device is a *filter* (glass, Wratten, or interference) placed just in front of the sample holder. The usual glass and Wratten filters are of relative wide bandwidths and low peak emission. Since the function of the filter is to absorb unwanted energy, the dissipation of the heat produced in the filter must be considered. The filter must resist change in the spectral characteristics over long periods of usage. For these reasons, the gelatin (Wratten) filter is now seldom used because it tends to deteriorate quickly with time. Narrower bandwidths are obtained with *interference filters*. This type consists of an evaporated coating of a transparent dielectric spacer of low refractive index sandwiched between semitransparent silver films.

Sharp cutoff filters are placed on each side of the films to eliminate other than first-order effects. These filters have a bandwidth of 10-17 mm and a peak transmittance of 40-60%. *Multilayer interference filters* consist of successive layers of high and low refractive index dielectrics on alternating layers. These filters are characterized by a bandpass width of 8 mm or less and a peak transmittance of 60-90%. These multilayer interference filters will transmit only light with a wavelength two times the distance between

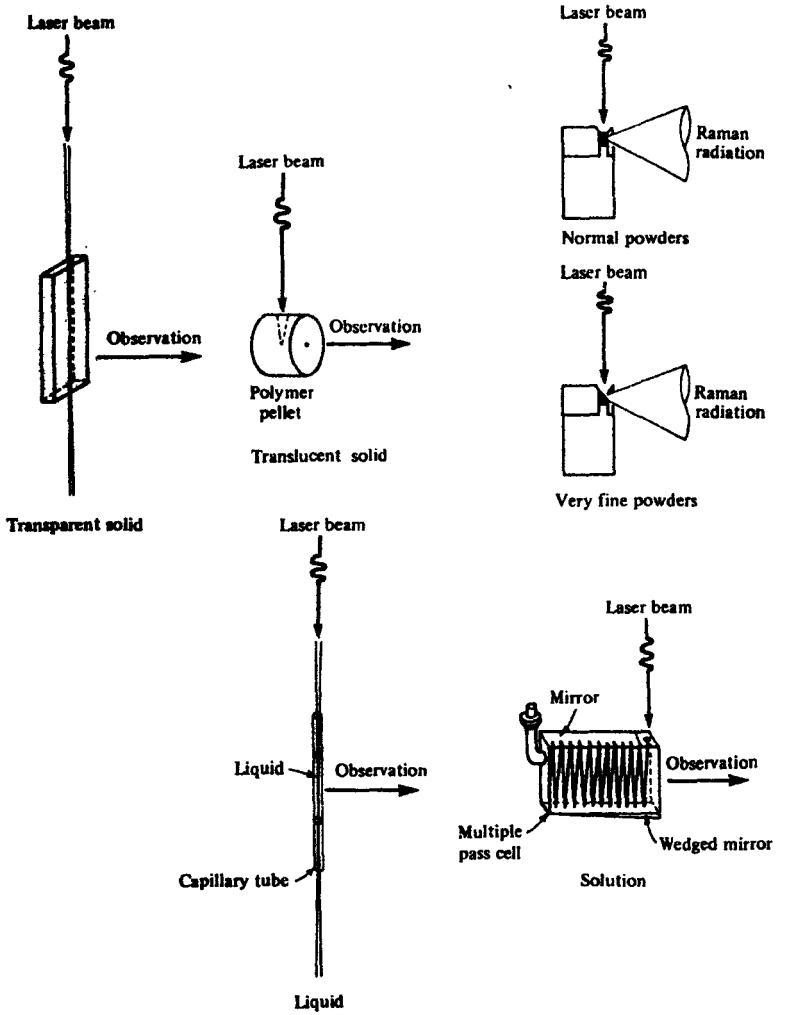


Figure 6.2 Experimental arrangements for laser-excitation of specimens in various physical forms.

the silver films (in phase), that is, with complete constructive interference. All other wavelengths will show partial or complete destructive interference as the reflected light rays are retransmitted out of phase with those of the same wavelength which are transmitted directly. These interference filters can be used with high-intensity light sources since they remove un-wanted radiation by transmission and reflection and not by absorption.

Many instruments use *prisms* or *diffraction gratings* as monochromators. These devices separate the various wavelengths of radiant energy as produced by a tungsten lamp by refraction or diffraction, and present them as a spectrum from which the desired wavelengths may be selected. The action of a prism in dispersing a polychromatic beam of radiation into a spectrum depends on the variation of the index of refraction with wavelength. As used in spectrometers, the light from the source is directed through a convergent lens into an entrance slit at the focal point of the lens, then through the prism and a second convergent lens. At an air-prism interface an entering light ray at an angle of incidence will be bent toward the vertical to the surface and, at the prism-air interface, it is bent away from the vertical. The image of the entrance slit is projected onto the exit slit as a series of coloured images arranged next to each other. Violet light is refracted (bent) farther than red.

The actual separation between two wavelengths depends on the dispersive power of the prism and the apical angle of the prism. A nonlinear wavelength scale results. The longer wavelengths are not refracted as much and thus are crowded. Dispersion by glass is about three times that of quartz. Quartz or fused silica is mandatory for inclusion of the spectra below 350 nm. *Diffraction gratings*, on the other hand, produce a linear noncurved spectrum. A diffraction grating is based on the principle that light rays of radiant energy will bend around a sharp corner, the degree of bending depending on the wavelength. Thus, with a beam of energy containing many wavelengths striking a grating, many tiny spectra are produced, one for each line of the grating. As the wavelengths move past the corners, wave fronts are formed. Where these cross, those that are in phase reinforce one another, and those that are to cancel out and disappear, thus leaving a complete spectrum to display upon the exit slit, much the same as with the prisms.

Some grating monochromators have a half-bandwidth as high as 35 mm, while other more expensive units have half-bandwidths of less than 0.5 nm. Despite the good dispersive qualities and the linear spectrum presentation, diffraction gratings have problems with stray light. This stray light may be defined as radiant energy at unwanted wavelengths reaching the detector. One problem is slight imperfections in the ruling. Another is caused by second-order effects of the grating. Part of this trouble can be eliminated by using a double monochromator or by a special mounting such as the Ebert system.

6.2.4 Cuvette Cell

Absorption spectrophotometry usually evaluates the absorption of a solute in a liquid solvent. The square or rectangular absorption cells have plane parallel faces and a light path of constant length. They are free of optical aberrations. Sets of cuvettes may be matched to very close tolerances. This type of cuvette is rather expensive and is usually utilized in the more expensive instruments. For ultraviolet work (below 340 nm) it is mandatory that this type of cell, constructed from silica or quartz, be used. For most work in the visual range, the cylindrical test tube type is sufficiently accurate. These cuvettes are prone to variable reflection and refraction errors as well as lens effect.

Glass tubing (from which cuvettes are made) is rarely round and is not polished, so there are considerably more surface irregularities. For this reason, round tube-typed absorption cells require close calibration and segregation into matched groups. Because these cells are generally somewhat oval, it is necessary that they be marked in front for proper light-path orientation. Another solution to this problem would be to use a flow-through cuvette. Since the same cuvette is in the light path at all times, cuvette errors would be compensated for by the bank. Highly reproducible results can be obtained, provided the cuvette does not become dirty and no bubbles are introduced and/or trapped in the cell. A source of vacuum is needed to empty the cell. Filling is often a very real problem because of the small tops. This may lead to spillage into the instrument and resultant corrosion.

6.2.5 Related Optics

The range of transmittance of materials for construction of windows ad lenses is a critical factor. The absorbance of any material should be less than 0.2 at the wavelength of use. Ordinary silica glass transmits satisfactorily from 350-3000 nm. Special Corex glass will extend the ultraviolet range to about 300 nm. Quartz or fused silica must be used below this. Beam reduction is accomplished by condensers that can reduce the beam size by a factor of 25 without loss of appreciable energy.

In colorimeters and economically priced spectrophotometers, simple lenses are used to focus or collimate the light beams. In the more expensive instruments, front-surfaced mirrors are used to reduce energy loss. These mirrors are aluminized on their front surfaces as other metallic surfaces show selective absorption at certain

wavelengths. The surfaces of these mirrors are coated with a thin film of magnesium fluoride to reduce light scattering.

6.2.6 Detectors

Any photosensitive device may be of use as a detector of radiant energy, provided that it has a linear response in the part of the spectrum to be used and is sensitive enough for the task at hand. These devices must first convert electromagnetic energy to a different type of energy, namely electrical energy. The electrical energy produced can then be measured. The most common of these is the *photocell* (barrier layercell). These require no external voltage source, but rely on internal electron transfers to produce a current in an external circuit. These cells are composed of a iron back plate and a layer of crystalline seleium or cadmium on one surface.

Electrical contacts are made with the iron plate and an electrical conductive transparent film on the front active layer. The cells are then sealed to prevent physical or chemical damage. A photon striking one of the selenium or cadmium atoms transfers its energy to an electron and raises it into a conduction band. This electron now travels from the front to the back of the photocell and through the external circuit for measurement. These cells are simple, quite inexpensive, and in general are very dependable. However, they have certain faults that should be understood. These cells can become "light blinded," a condition similar to that of the human eye. They need time to rest. Their response is temperature sensitive, requiring some warm-up time for temperature stability. Their low internal resistance and low output of electrical energy are not easily amplified.

Phototubes are constructed of a negatively charged cathode and a wirelike positively charged anode. The cathode is coated with a photoemissive substance such as *cesium oxide*. When a photon strikes this layer, electrons are emitted and jump through the vacuum over to the anode, where they are collected and return via the external circuit. The output from these tubes can be amplified and then fed into external circuits for greater sensitivity. For this reason, this type of detector is used on all precise instruments that have a close restriction on the slit and thereby the wavelength presentation.

The photomultiplier tube combines photocathode emission with multiple cascade stages of electron amplification of primary photocurrent within the tube itself. The tube is constructed so that the primary photoelectrons from the cathode are attracted and accelerated to several succeeding dynodes. These dynodes are

constructed of a material that will give off several secondary electrons when hit by other high-energy electrons. Thus the photomultiplier tube can measure light intensities about 200 times weaker than the ordinary phototube.

6.2.7 Readout Devices

A galvanometer is an instrument used to detecter or measure currents. A flat suspended coil lies in the plane between the poles of a permanent magnet. When the current to be detected or compared passes through the coil, it sets up a magnetic field whose poles are at the front and back, or 90° away from permanent magnet poles. In trying to set its lines of force in line with those of the magnet, the coil turns as far as the magnetic forces can twist the wire. The stronger the current, the stronger the magnetic field and the farther the coil turns. The iron core makes a uniform field for the coil to turn in; a beam of light reflected from the minor on the coil serves as a pointer. When an arbitrary readout device, such as per cent scale or optical density scale, is incorporated into this system, we have what is called a *direct readout meter.*

Actually, this readout could be in millivolts, millamps, mg% mEq/liter, or any other arbitrary unitage, provided the scales take into consideration the linear relation of milliamps and $\%T$ and the log relationship between millivolts and absorbance. Therefore, this type of readout is generally considered to be the fastest, simplest, and the easiest to use. If, on the other hand, an electrical force governed by the action of a variable resister (potentiometer) is used to bring the meter to null point, the system is said to be a *null-point system.* In this case, an arbitrary scale is fitted to the potentiometer scale. This scale may carry the same arbitrary type of unitage as the first. This slide wire potentiometer may also be connected to a servomotor of a digital readout system or of a recorder.

6.3 TYPES OF INSTRUMENTS

Although the filter photometer is not to be considered as a spectrophotometer, because it does not give a continuous source of monochromatic radiant energy, they must be considered together when it comes to function and system design. Colorimters that use the regular glass filter will also use a simple low-energy light source. In fact, this type of filter dictates the use of this type of source, as well as eliminating the need for even a single lens. Therefore, the simplest system would require a power supply, light source, filter,

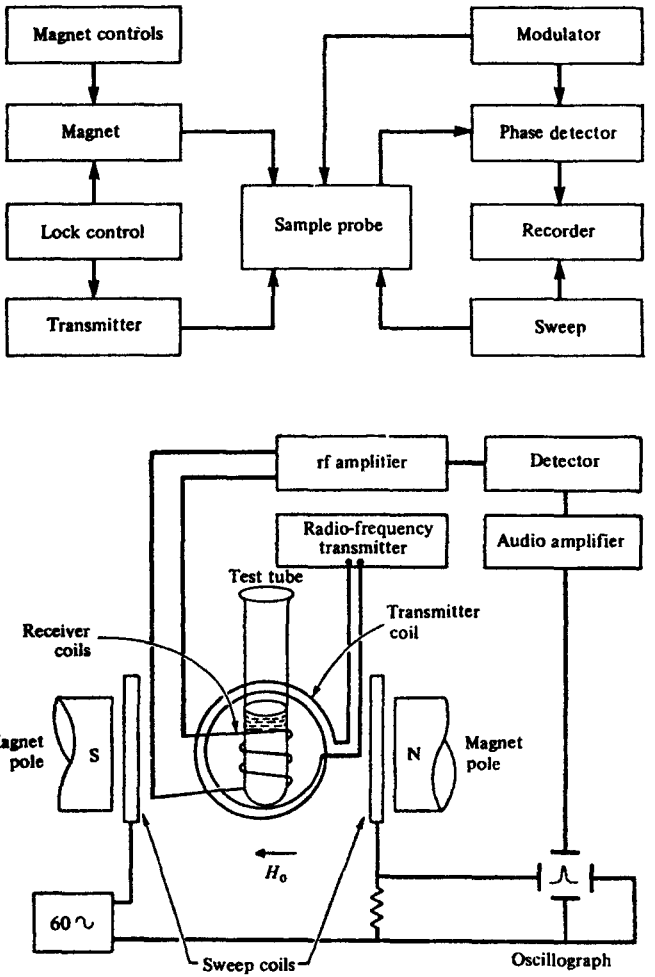


Figure 6.3 Top, block diagram of a high-resolution nmr spectometer. Bottom, schematic arrangement of components in vicinity of sample probe.

sample holder, detector, and a meter in that order. However, when an interference filter, prism, or diffraction grating is used, a more elaborate system is necessary. These types of monochromators require that radiant energy be focused into the system in parallel lines.

Diffraction gratings and prisms also require the use of entrance and exit slits to *limit* the spectrum and exclude stray light. In fact, most grating monochromators require that sharp cutoff filters be placed at the entrance to eliminate second-order spectrum. When

you add to this a sample holder, a detector, and a meter, you have then a single-beam spectrophotometer. In a double-beam system monochromatic light from either a single or double monochromamtor is focused through both a reference and a sample compartment. The intensity of these two beams of light is then measured by either one or two detectors, and the sample beam is compared to the reference side as a ratio. This ratio may then be fed into a meter or directly into a ratio recording and a beam splitting and reuniting system of choppers to overcome the problem of balancing two phototubes to an exactly equal sepctral response.

6.4 QUALITY CONTROL

The most insidious problems of instrument malfunction have to do with the certainty of wavelength, bandwidth, linearity, stray light, and photometric accuracy. These terms probably mean very little to the laboratory workers who rely completely on the colorimeters to perform, without a question as to how well they did or could function; but quality in instrument performance cannot be assumed, it must be proven.

The monitoring of instrument performance is even more important in the spectrophotometer. This is especially true in the narrow bandpass units. These systems with their greater potential accuracy and sensitivity demand closer attention to their functional performance. Electrical malfunction will not be considered here, even though it is a part of instrument quality control. This type of problem is usually covered in detail by the instrument handbook and requires a specially trained individual to make repairs.

6.4.1 Wavelength Calibration

Wavelength calibration is a means of insuring that the radiant energy being emitted from the exit slit of the monochromator is the same as the specified by the wavelength selector and within the tolerances of the instrument. This simple check should be made whenever a new lamp is installed and routinely thereafter. This is to insure that the slight changes in lamp filament position during usage are corrected and do not lead to gross analytical errors. Minor errors in wavelength setting are of relatively little importance in wide bandpass spectrophotometers, and these units may be checked less often.

Wavelength calibration of the more sophisticated narrow bandpass instruments are of the utmost importance, and no scan of any

substance is worth doing if the monochromator is not in calibration. Didymium and holmium oxide glass filters have been used for many years on the general laboratory instruments. More recently, the combination of the didymium glass filter and an acid solution of nickel sulfate has been used. This solution gives a peak transmission at 510 nm.

A solution of cobalt chloride has a peak absorbance at this wavelength. This solution may be helpful as a secondary check on the nickel sulfate when it is not in agreement with the didymium filter. Precision instruments require a more precise means of wavelength calibration. The emission bands of mercury arc or deuterium lamps are good sources of radiant energy with specific emission spectra. The quartz mercury arc lamp, with its multiple emission bands, provides the best source for accurate wavelength calibration from 205-1014 nm. The deuterium lamp in ultraviolet instruments has two major emission lines, 486 and 656.2 nm, that provide a good source of wavelength calibration in the visual range. This lamp provides the most practical means of wavelength calibration of a grating instrument, for only two lines are needed when light dispersion is linear.

6.4.2 Stray Light

Stray light is any radiant energy measured by the detector that is outside the spectral region isolated by the monochromator of the instrument. It can be caused by dirty optics, poor baffling, or faulty grating in the monochromotor, as well as fluorescence of the sample itself. Because stray light may be independent of sample concentration, its relative effect is usually greatest at low transmittance, where instrument errors are also greater. The American Society of Testing Material E-13 Committee Standards have been adopted for the measurement of stray radiant energy in spectrophotometers and provides a good means of instrument evaluation.

However, for ease of understanding and routine checking, a common sense approach may be more practical. Scattered light can be of any wavelength, and in grating instrument it may also be of second-order reflection. This may also be true of old and poorly constructed interference filters. Thus, this light may well be of a wavelength far removed from that selected and will not be absorbed by the test solution. From this information, it is easy to reason that scattered light can cause nonlinearity and insensitivity. Sharp cutoff

filters may be used to directly asses the amount of stray light at the extremes of the instrument range. These filters have the unique ability to stop all light above or below a certain range. Thus, by the proper selection of filters, it is possible to check the apparent stray light of any instrument. This may be done by setting the wavelength selector at the desired wavelength (400 or 700 nm) and then adjusting the instrument meter to read 100% transmission. Place a filter in the sample compartment and note the meter reading. This reading is an indication of stray light and must be considered as a instrumental error.

If one is performing time rate analyses, one of the simplest checks to make is to measure the absorbance of NADH at 340 nm with the tungsten and deuterium light source. If the absorbance of the solution is greater with the deuterium lamp, stray light was present when the tungsten lamp was used and this has been corrected by employing the deuterium radiation source. This is because the deuterium lamp is a discontinuous radiant energy source in the visual range, and therefore does not emit radiation at 680 nm, the wavelength of light that is of primary importance in second-order effect.

Also very little energy is emitted by this lamp above 400 nm, so that scattered-light problems are greatly diminished when working in the ultraviolet range. Nickel sulfate solution has also been used for this purpose over the visual range. This solution shows maximum absorption at 400 and 700 nm, and under certain circumstances no light should be transmitted at these wavelength. Any deviation from this on repeated checks is an indication of stray light.

6.4.3 Photometric Accuracy

Provided that the stray-light check is negligible, the linearity check tests the ability of the photocell to produce a signal proportional to the light intensity and of the instrument's meter system to measure this signal accurately. A linearity check can be made by reading the absorbance of standard solutions of various salt solutions that have major absorption peaks at certain wavelengths or by using neutral density filters that absorb energy over a broad range. Solutions of potassium chromate and potassium dichromate have been used for evaluating photometric accuracy in the ultraviolet range. A solution of 0.04 g/liter of potassium chromate dissolved in 0.05*N* potassium hydroxide will show maximum absorbance at 273 and 373 nm. The

absorbance range reported for this solution is 0.189-0.199*A* and 0.247-0.249*A*, respectively, for these two wavelengths. When measuring absorbance below 260 nm, this solution should not be more than 6 months old. This may be the reason for the wide range in absorbance values at 273 nm. A solution of potassium dichromate, 0.05 g/liter dissolved in 0.01*N* sulfuric acid, also exhibits two absorbance peaks that may be used. These are at 257 and 350 nm and should read 27.6% transmission at the later wavelength. Absorbance patterns of this solution are varied by changes in acidity: therefore, the pH should be carefully controlled. Cobalt ammonium sulfate and Thompson's solution have been used for this evaluation in the visual range. These solutions require great care in preparation and handing.

The absorbance cells used must be cleaned and matched before use. Although these materials are relatively stable, they may react with contaminants and deteriorate. The solvent solution must also be checked for absorbance in the ultraviolet range when these solutions are to be analyzed. Frings, *et al.*, has proposed using French's green food coloring for monitoring detector response at three wavelengths. This dye solution has absorbance maxima at 257 nm, 410 nm, and 630 nm. The advantage of this solution is that it provides a means of response verification in both visual and ultraviolet ranges with one solution at a very modest cost.

The National Bureau of Standards developed a set of neutral density glass filters for the specific purpose of checking the photometric scale of spectrophotometers. These filters are premounted in holders ready for use in the standard 10-mm cell compartment. They are not readily affected by temperature change nor dose they require a critical wavelength adjustment. These standard reference materials (SRM) come as a set of three individual filters calibrated and certified to ±0.5% transmittance over four different wavelengths of the visual range. These filters were the forerunners of other sets manufactured by Bausch and Lomb as well as Chemetrics Corporation. All of the later filter sets have a filter than will fit a standard 10-mm cell compartment. These sets provide a means of checking for stray light, bandpass, and wavelength as well as photometric accuracy.

6.5 STEPS IN SPECTROPHOTOMERTY

Since there are many different spectrophotometers, it would be of little use to describe the operation of any one of them. In this

section, therefore, some general pointers, which apply to all spectrophotometers, will be covered.

6.5.1 Care and Use

Although spectrophotometre cells come in a variety of sizes and shapes, those most commonly used are square cells with a light path of 1.0 cm. Silica or quartz cells are used for measurements in the ultraviolet region, since glass is not transparent below 330 nm. In colorimeters, test tubes are employed. No matter what type of cell is used, it is of primary importance that the cells are well-matched. The absorbance of a substance in solution is compared to a blank containing solvent alone. If the two cells containing solvent alone do not have identical absorbencies at the test wavelength, the measurement will be in error. To circumvent this problem, it is essential that the difference in absorbance at the test wavelength be determined with both cells filled with pure solvent. The value obtained (the cell correction) should be subtracted from the measured absorbance of the sample.

Cell corrections can be relatively high. For example, a small fault in the optical surface of a cell can result in an absorbance of 0.15. Furthermore, an invisible fingerprint or residue in a cell can have absorbencies in the ultraviolet region of up to 1.0! When the cell corrections are high (for example, greater than 0.025) it is usually more convenient to clean the cells (see below) to reduce the correction. In some cases, visual examination of the cells will reveal small scratches. Badly marred cuvetes are useless because the scratches cannot be removed readily. Only two of the four sides of a cuvette are of optical quality.

With most cuvettes the nonoptical sides are clearly marked. If the markings are removed, it is important that the nonoptical surfaces be remarked. A cell inserted the wrong way in a holder can lead to grossly erroneous results. After each use, cuvettes should be emptied immediately. An aspirator pump is frequently used for this purpose. The cuvettes are then rinsed three to four times with the solvent (usually water). The final rinse may be made with methanol and, after drainage of the methanol, the cuvettes may be blown dry with a stream of clean, filtered air.

The outer surfaces of the cells should be polished with good quality lens paper. Do not use ordinary tissue as it will eventually scratch the cells. On occasion, this mild cleaning procedure will not

reduce cell corrections. Other more drastic procedures may then be tried, such as soaking the cells in chromic acid for a brief period.

6.5.2 Slit Width

The light which emerges from the exit slit of a monochromator is not truly monochromatic. Instead, there is a distribution of a portion of the spectrum across the exit slit with peak intensity at the selected wavelength. Thus, the narrower the slit width, the greater the spectral purity of the light. In general, the best slit width to use is the narrowest. Since the dispersion of light by a prism is wavelength-dependent, and decreases with increasing wavelength, the control of the slit width is very important for measurements in the near infrared region of the spectrum. The magnitude of the slit width is also a valuable clue to the proper operation of a spectrophotometer. For example, a rapid increase in the width of the slit in the ultraviolet region may indicate a failing lamp or a dirty mirror.

6.5.3 Wavelength Calibration

A monochromator is only as good as its calibration. Although it is a relatively simple matter to calibrate a commercial spectrophotometer, calibration is not performed as frequently as it should be. Obviously, a small variation in the true wavelength setting can lead to erroneous results if the measurement is being made at the absorption maximum of the substance, particularly if the absorption band is very narrow.

6.5.4 Stray Light

In practice, it is very difficult to prevent reflections from various components of the monochromator from entering the exit slit. This light is called stray light. Filters are provided on most spectrophotometers to combat this problem. Stray light can cause false absorption maxima and even cause maxima to disappear. Furthermore, a malfunctioning spectrophotometer can give perfectly correct readings in the ultraviolet and blue regions of the spectrum while giving nonsense readings in the red region. There are several ways to check for stray light, and at least one of these measurements should be made as part of routine maintenance of a spectrometer.

6.5.5 Deviations from Beer's Law

A standard curve must be constructed for the spectrophotometric assay of any substance. Apparent deviations from Beer's law are

known, especially in the case of substances capable of dissociation. Furthermore, many spectrophotometric assays are based on the formation of colored complexes, and it is possible that the concentrations of the reagents may become limiting for complex formation. If at all possible, it is best measure absorbance in the region 0.1 to 1.0 The analytical error which is inherent in spectrophotometers is minimal over this rage.

6.5.6 Turbid Samples

A cloudy or turbid sample gives absorbance values which are too high. Some of the light which would have been transmitted by a clear solution of the sample is scattered and, therefore, does not reach the phototube. It is essential to have clear samples. Even a slight turbidity, which may be early invisible to the naked eye, can lead to serious errors in readings, especially in the ultraviolet region. A simple test for artifacts caused by light-scattering is to vary the distance between the sample and the phototube.

If the absorbance reading is independent of this distance, light-scattering is not a problem. Unfortunately, this test is impossible to perform with many spectrophotometers. In this case it is more difficult to assess the contribution of light-scattering. It should be mentioned that turbid samples are frequently used deliberately for quantitative or semiquantita-tive analysis. For example, measuring the increase in absorbance (due primarily to light-scattering) of a bacterial liquid culture is a convenient method for determining the rate of growth of the culture.

7

Physical Displacement

Successful racehorses gallop at 35 to 40 miles per hour. In this gait the two fore feet are set down in rapid succession, then the two hind feet, followed by the two fore feet again. The spring action of a sheet of tendon allows the back to bend and extend at appropriate stages, lengthening the stride.

7.1 RUNNING ADAPTATIONS

The animals, which pass their life time exclusively on the earth are known as cursorial forms. In these types, speed has developed in a very wonderful manner. The body of cursorial animals is moulded externally in such a way, as to offer minimum resistance to the air, through which they walk or run. The limbs are the main propelling organs of the terrestrial animals. As the hind limbs are the more efficient drivers, they undergo more modifications than the fore-limbs in an organism, whose cursorial adaptation has not

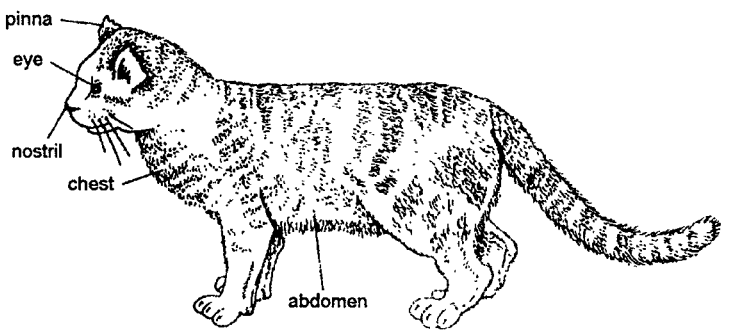


Figure 7.1 *Felis domesticus* (Cat).

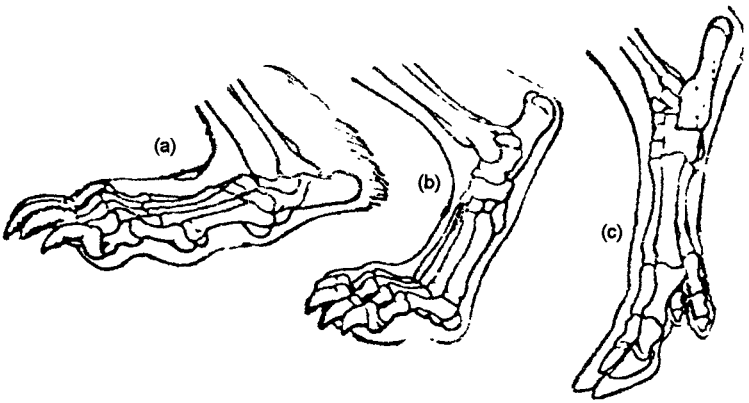


Figure 7.2 Foot postures—(a) plantigrade, (b) digitigrade, (c) unguligrade.

gone very far. The limbs have become larger, the foot posture has undergone changes from plantigrade to digitigrade in a large proportion of modernized mamm-als (specially in ungulates), which has occurred for speed adaptation.

7.1.1 Body Contour

All speedy animals, whether terrestrial, aquatic, or aerial, have the body moulded externally in such a way as to offer the least resistance to the medium through which they pass. Owing to the greater resistance of water, this is especially true of the aquatic, nevertheless a speedy cursorial type also shows it, though not always as well when at rest as it does when in action. A race-horse with head and neck extended, ears thrown back, and every tense muscle of its wonderful body working with machine-like precision, shows the beautiful contour of a perfectly adapted mechanism. The body spindle-shape, the stream-line contour, the lines extended into the neck and head without a break in their curves—all are calculated for swift passage through the air with a minimum of resistance.

7.1.2 Mechanism

7.1.2.1 Loss of general utility

The propelling organs in cursorial forms are the limbs exclusively, so that, aside from the resistance-lessening contour, this adaptation concerns itself chiefly with their modification, of which the first is the loss of general utility. This is especially true ot the hind limbs, because they are the more efficient drivers and therefore, especially in forms whose adaptation has not gone very far, are likely to be

somewhat is advance of the fore limbs in the degree of their revolution. For this several reasons may be assigned: (1) The extended hands pull the body forward in running, (2) the fore part of the body is usually heavier than the hind part and requires large limbs to support and propel it; (3) running is a sort of leaping on all fours, and the hands are larger and wider to take the implact when the animals falls forward; finally (4) the foe limbs, being nearer the mouth and hence perhaps somewhat concerned in food-getting, are the last to lose their generally utility. Two notable instances of this accelerated evolution of the hind limbs over the fore are the early fourtoed horses of the Eocene in which, while the hand still retains its four digits, the foot has but three. It is not until Oligocence time that the additional finger is lost and the evolution of both limbs becomes parallel. In another Oligocene form, *Protoceras*, a curious artiodactyl with remarkable excerescences upon the skull and dagger-like canine teeth, the hand is fourtoed while the foot has but two.

7.1.3 Change in Foot Posture

The primitive terrestrial foot is plantigrade (Lat, *planta*, sole, and *gradi*, to walk), which means that the entire palm or sole rests on the ground, neither wrist not ankle being raised. Almost the first step in speed adaptation is the lengthening of the limb and this may be accomplished without the actual elongation of a bone, merely by rising upon the toes. While the bear, raccoon, and the primates such as the baboons and man are plantigrade, probably secondarily so, a large proportion of moderinzed mammals have become digitigrade (Let. *digitus*, fingure, toe), walking

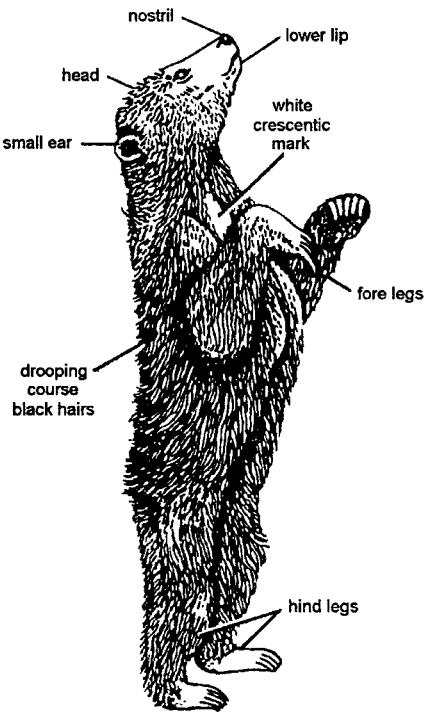


Figure 7.3 Foot of plantigrade bear.

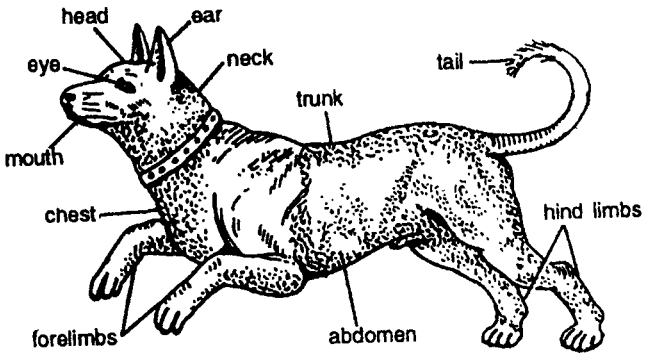


Figure 7.4 Digitigrade—Canis familiaris.

or running upon the digits themselves, with the bones of the wrist (carpal) and ankle (tarsal), the upper ends of the palm (metacarpal) and the sole bones (metatarsal) clear of the ground. Some of the speediest of animals—dinosaurs, birds, dogs, all mammals in fact but the ungulates—have merely perfected the digitigrade gait, developing special sole-pads for the absorption of the shock of impact, and have never gone beyong it. The ungulates or hoofed animals, on the other hands, walked on the modified nail or hoof (unguligrade, Lat. *ungula*, hoof), the distal toe-bones (unguals) being depressed or flattened and not, with rare exceptions, compressed or

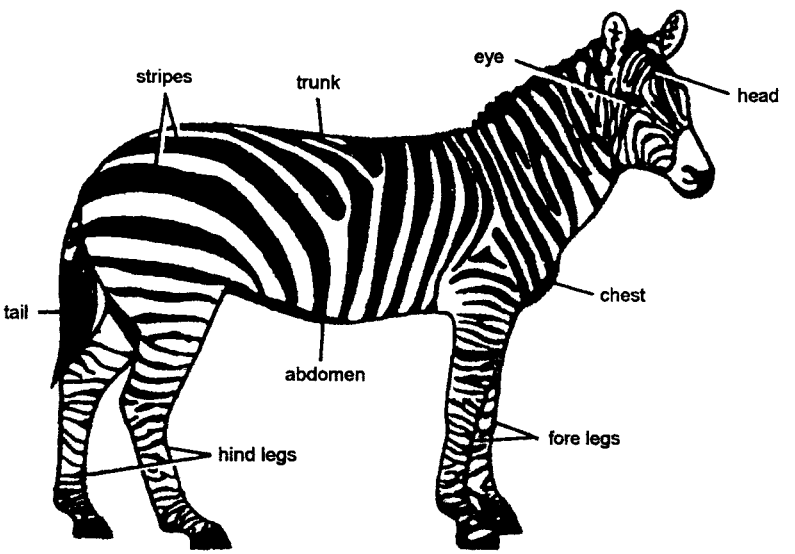


Figure 7.5 Unguligrade limbs of *Equus zebra*.

claw- like. The hoof has reached the highest degree of perfection in the horses; in other related but non-cursorial types like the rhinoceros the hoof bear little of the weight, as a broad cushion-like pad serves instead.

Ungulate animals show all gradations from the semiplantigrade condition of the slower forms to the high, stilted hoofs of certain of the African antelope, notably the Klipspringer (*Oreotragus saltator*). A truly unguligrade condition has never been attained among reptiles, although certain beaked dinosaurs has depressed instead of clawlike unguals. *Triceratops* and *Stegosaurus* were certainly far from speedy; hadrosaurs, on the other hand, which also bore this type of ungual, were bipedal and doubtless possessed a fair measure of speed when well under way, though littel celerity of movement is indicated.

Certain ungulates like the modern camels have become secondarily digitigrade, the foot having retrogressed as an adaptation to the yielding desert stands. This is not accompanied, however, by any material loss of speed, as the camels are among the most remarkable travellers of all terrestrial forms.

7.1.4 Loss of Digits

Plantigrade animals are generally five-toed; there are of course exceptions, but the elevation of the wrist generally carries with it digital reduction, digitigrade animals becoming four-toed, unguligrade four-, three-, two-, or even one-toed, two toes in the artiodactyles and one in the perissodactyls being the irreducible minimum.

The frilled of Australia, *Chlamydosaurus*, is five-toed but the lateral toes are shorter than the median ones, which is almost universally ture except in aquatic types such as the seal and otter. Hence when *Chlamydosaurus* runs on its hind feet, as it does when startled, the outer and inner toes are raised off the ground and the animals makes a three-toed track. If this were the habitual gait of the creature, the lateral digits would be rendered practically useless and would follow the course of all useless organs and become reduced, whatever the philosophical explanation of the mean whereby this is accomplished.

Environment as well as speed adaptation has its influence in determining digital reduction, for it will be accelerated if the ground is hard, as in the prairie-evolved horse with but one remaining digit, or the prong-horn antelope of similar environment with two. On the other hand, the Miocene forest horse, *Hypohippus*, retained and lateral toes as functional organs just as the reindeer and caribou

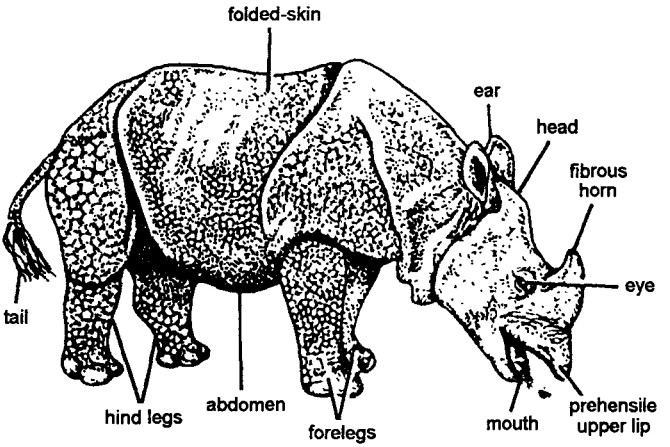


Figure 7.6 Rhinoceros foot.

(*Rangifer*) have today, as an adaptation to a yielding footing, while contemporary relatives had in each instance evolved much farther along the line of digital reduction.

Concurrent with the loss of digits, especially if the foot be lengthening after the manner to be described below, comes a compacting of the bones of the palm and sole (metapodials) and often this is carried so far as to give rise to actual fusion of these elements into a "cannon-bone". The dinosaurs, with one doubtful

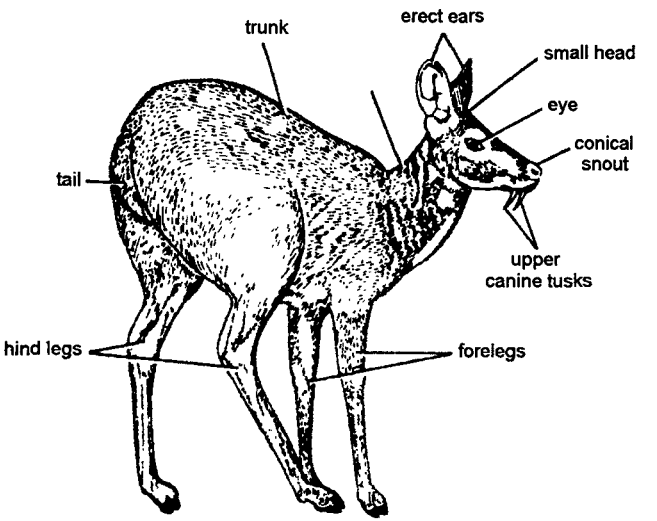


Figure 7.7 Legs of musk deer.

exception, never attained an actual fusion, although in many respects, especially in *Ornithomimus* of the Cretaceous, the foot is very bird-like. The birds, on the other hand, always show a fusion of the metatarsals. Among mammals the carnivores do not torm a cannon-bone nor does the marsupial wolf; but all the speed-adapted ungulates do. Ancient ungulates, however, had the metatarsals separate, and we can often witness the fusion in fossil series (camels, etc.) when the proper degree of speed adaptation has been reached.

Among rodents, the jerboa, a three-toed bipedal form, has a foot and metatarsus so wonderfully bird-like that one almost has to count the phalanges of the digits to be sure he has a mammal before him.

7.1.5 Reduction of Fibula and Ulna

The fore arm and shin, that is, the second segment from the body, have each typically two bones : in the arm, the radius and ulna, and in the leg, the tibia and fibula. These are both developed in slow-moving forms or where the fore limb still has considerable general utility, especially if the rotation of the hand on the arm is retained. On the other hand, cursorial forms, especially if the limbs are exclusively locomotor, tend to reduce the ulna of the arm, the proximal end onlyl being present in extreme cases to form the elbow joint. They also lose the fibula of the leg, which may be reduced to the merest vestige.

7.1.6 Loss of Universal Movement

The entire motion of the limbs becomes pendulum-like, that is, restricted to movement in but one plane, the exception being at the hip and shoulder, where universal movement is still retained through the development of a ball-and-socket articulation. The necessity for this is apparent, first to avoid interference between fore and hind feet when running, since a dog, for instance, at top speed brings his hind feet well in advance and outside of the fore. A second need is that of lying down and rising again, which would be practically impossible were the movement at hip and shoulder restricted to the fore-and-aft plane. With the other articulations, those of ankle and wrist, knee and elbow, and between the digits, the tendency is toward rigid limitation of movement in unguligrade, less so in digitigrade forms. This is accomplished by the development of tongue and groove joints such as were discussed under Kinetogenesis. These are very perfectly shown in the hind limb of a modern horse, as

well as at the elbow joint, forming in each instance an articulation permitting movement through a wide arc in one plane of space and none whatever in any other. These joints, while they may be broken, cannot be dislocated.

The limbs are compound levers, for not only is there motion of the limb as a whole but also between its component parts. The lengths of each of the several segments bear definite relations to the speeds developed and also to the loads they have to carry. those forms which, like the elephant, are mightly of frame, have a type of limb which is in marked contrast to that of a horse. To the former type has been applied the term graviportal (*i.e.*, weight-carrying), to the latter cursorial, although both are adapted to increase their owner's travelling powers. In the graviportal type such as the mastodon, the foot is short and the thigh and shin relatively long, whereas in a cursorial form the foot elongates and the thigh is conservative.

7.1.7 Lengthening of Limbs

The lengthening of the limbs in cursorial types is usually thus accomplished by a growth of the distal segments only, the foot and shin and hand and fore arm increasing, but rarely the thigh or upper arm. This increase is therefore both actual and relative. This is well shown in comparing the limb of a horse with that of a man, for while the human thigh may actually exceed the length of that of the horse, the horse's foot, which is measured from the end of the hoof to the hock, the equivalent of the human heel, may be two and a half times that of the man.

This lengthening of the distal segments, which is for the purpose of increasing the length of stride, would be of little avail were the muscles not concentrated at the proximal end of the limb, their power being transmitted by long slender tendons to the lower leg and foot. One effect of this may be best understood by comparing the limb with a pendulum. The length of the pendulum determines the scope of swing, but the position of its center of gravity controls its speed or rate of beat. To accelerate the beat, therefore, the bob is moved upward, to retard it, it is moved downward, the arc of pendular motion remaining constant. This concentration of the muscles at the proximal end of the limbs has the same effect as raising the pendulum bob, and by this device—lond slender limb and high but powerful muscles—the maximum length of stride and

speed of movement are obtained. While this may well be an important reason for the concentration of muscle, the parallel with the pendulum is not exact, for the opposing muscles not only serve to initiate the swing but also to damp it. Perhaps the chief reason, therefore, is that by raising the insertion points of the thigh-muscle focussing around the knee, the *angles of insertion* of many muscles are increased and this gives higher propulsive components, *across* the shaft of the femur; at the same time the muscles are shortened and made thicker, which increases their power and speed of contraction. It can readily be seen that a limit may be reached beyond which bone will not stand the strain to which it would be subjected, although bone is a wounderfully efficient material. Hence one would expect to find the greatest speed developed on the part of creatures of small to moderate size—the antelope of Africa or horses like the wild ass of Persia (*Equus onager*) the speed of which has been mentioned and which reaches a stature of but 11½ hands. The Mongolian wild ass or Kiang (*Equus Kiang*), according to Roy Chapman Andrews, who chased one for 29 miles in a motor car; averaged 30 miles an hour for the first 16 miles and then when it began to slow down, still ran four more miles at a speed of 20 miles an hour. The modern race-horse is relatively small comparedb with some other breeds, and the limit of weight, size, and speed consistent with safety seems to have been approximately reached.

7.1.8 Ratios

Lengthening of limbs also implies, at any rate in a quadruped, the concurrent lengthening of neck and skull in order that the animal may readily reach the ground for food and drink. Hence the various parts of an animals's frame bear definite ratios to one another and this may also extend to individual bone proportions, definite "speed index" being recognizable. This makes it possible through the law of correlation to gain some insight into the habits of extinct and little-known forms through the study of comparatively fragmentary remains.

7.1.9 Bipedality

A two-footed mode of progression as an adaptation to speed has been repeatedly evolved among vertebrates, as follows :

Reptiles :

Lizards, several occasionally bipedal.

Dinosaurs, two evolutions.

Birds :

One evolution.

Mammals :

Marsupials, one evolution.

Rodents, three evolutions.

In all, eight or more times.

The erect posture of man was probably not originally a speed adaptation, nevertheless speed has always been a vital factor in human evolution, in all offensives and defensive operations. The human foot, which was originally a climbing sturcture, has been readapted for bipedal walking and running. The long thigh and shin of modernized man increases the stride materially in contrast to those of the gorilla and chimpanzee. The Neanderthal man had short stocky limbs as compared with the existing species, but doubtless could outrun any of the anthropoid apes.

7.1.10 Counterpoise

Some sort of counterpoise is always necessary in a semi-errect biped and the tail usually assumes this function. In the Kangaroo and in the dinosaurs it is a powerful organ and serves as a prop, like a third limb, when the creature rests without coming down on all fours. The tail may be comparatively short and heavy in larger forms or extremely long and slender in more lightly built creatures, on the principle that an ounce at the end of a sixteen-inch lever is as effective as a pound on one but an inch in length.

Many dinosaurs and bipedal lizards have a long, attenuated tail. This is especially true of the dinosaur *Podokesaurus*, a Triassic form from the Con-nectucut valley, and of the Australian frilled lizard *Chlamydosaurus*. Among mammals the Kangaroos have a relatively short, heavy tail; the jerboa on the other hand has a very long one terminating in a tuft of hair, which through its resistance to the air adds effect to the counterpoise.

No existing birds have a long tail, that is, as regards the actual tail itself, although the feathers may be long. These, as in the pheasants, may subserve a balancing function. The true cursorial birds, Ratitae, are practically tailless, but maintain their balance with ease, the heard and neck sufficient. The ostrich raises its wings as an aid in running, but with the practically wingless cassowary or the emu the head and neck alone must serve.

7.1.11 Shortening of Neck

In bipedal mammals there is a tendency toward reduction in the length of the neck, especially in the rodents such as the jerboa, in which cursorial adaptation is extreme and there is a remarkable cervical reduction associated with the shortened skull. There is of course no diminution in the number of neck vertebrae, for the number, seven, is with two, or three exceptions (sloths and manatee) constant among mammals; but the vertebrae themselves are shortened and tend to coalesce into a rigid mass of bone. Thus in the rodent *Pedetes* cervicals 2 and 3 are so closely articulated as to eliminate motion, in *Perodipus* the axis (2d cervical) and next two vertebrae are fused, while in *Dipus* (jerboa) all of the cervicals expect the atlas (1st cervical) are coossified as in the whales. As we shall see, the shortening of the neck may be also an aquatic adaptation, since it occurs in the whales and sirenians.

7.1.12 Mental Precocity

Animals which depend upon speed for safety, as do the ungulates or the whales, cannot have helpless young. Such must either be brought forth in some secluded den or carried about by the mother. Carnivores and rodents have very feeble young, but they are kept hidden until able to shift for themselves. The Kangaroo, on the other hand, must carry her offspring with her and this undoubtedly proves a very heavy handicap to the race when competition with higher forms prevails, for the destruction of the mother means that of the young as well. With all other forms which depend upon speed for safety, the young animals must be able to keep up with the herd almost at once.

Hence there is no period of helpless infancy, but the new-born deer on horse or camel, with its grotesquely long limbs, has the relative mental alertness of a very much older dog or rat, although the ultimate mental attainments of the ungulate may not be so great. As an illustration of precocity, Andrews describes an experience with a ten-day-old baby antelope. For four miles he seldom went solwer than 25 miles per hour and for five more miles he averaged 15 miles per hour. He circled too quickly for the car in fact, which had to average about 40 miles to overtake him.

7.1.13 Significance of Cursorial Adaptation

Not only does speed adaptation give rise to some nature's most beautiful and perfect machines, but it seems to have a much deeper

meaning which has been summarized by Boom. He is speaking of Permian reptiles :

"The African, or more preferable the south Atlantic type, is chiefly remarkable for the great development of the limbs.....what may have been the cause we can not at present tell, but it was a most fortunate thing for the world. It was the lengthened limb that gave the start to the mammals. When the Therapsidan [mammal-like reptile] took to walking with its feet underneath and the body off the ground it first became possible for it to become a warm-blooded animal. All the characters that distinguish a mammal form a reptile are the result of increased activity—the soft flexible skin with hair, the more freely movable jaws, the perfect four-chambered heart, and the warm blood. It is further singularly interesting to note that the only other warm-blooded animals, the birds, arose in a similar fashion from a different reptilian group. A primitive sort of dinosaur took to walking on its hind legs, and the greatly increased activity possible resulted in the development of birds. Birds were reptiles that became active on their hind legs mammals are reptiles that acquired activity through the development of all four."

Back of all this lay the impelling natural cause. The earliest known mammals are late Triassie, the first recorded bird Middle Jurassic; the inference that both stocks arose in Permian time is justifibale from the degree of evolution which each class had attained by the time the actual record of their existence beings. Schuchert tells us that early in the Permian the climate of the lands seems everywhere to have been arid or semi-arid and that this condition lasted into Jurassic time. One characteristic of desert animals of to-day—the lizards, birds, gazelle, Persian ass—is *speed*, for the creature must fare widely for food and drink if he would fare well. Again we are told that during the Permian there was a period of extensive glaciation with a severity of climate, especially in the southern land masses, as great as, if not greater than, the polar one of Quaternary time, although, like the latter, the Permian glacial period had warmer interglacial intervals as well. The incentive for speed already given, rendering the development of warm blood possible, the devastating cold would soon place a premium upon such as did develop it and eliminate those which did not. From this fortunate relation of cause and effect might well have arisen on the one hand the primal mammal, making human evolution possible, and on the other hand the ancestral bird.

7.2 WALKING

Human walking is unique. No animals walks as we do; althouh birds walk on two legs and apes sometimes do, thier styles of walking are quite different from ours. Most birds use flight rather than walking as their principal means of travel, and apes do most of their walking on four legs. Apart from ostriches and other fllightless birds, no other animal depends as much as we do on two-legged walking. Is there something particularly good about our peculiar walking style?

7.2.1 Pendulum of Swinging Legs

The distinctive feature of human walking is that we keep each leg almost straight while its foot is on the ground. The sequence of photographs at the top of the page follows the course of movement through a single step. At stage (c) of a step, the supporting leg is straight and vertical, so the body is high. At stages (a) and (c), the legs are straight but sloping, so the body is lower. As a consequence of our leg's changing slopes, our heads bob up and down by about 40 millimeteres (1.6 inches) in the course of each step. Although this bobbing may seem like a trivial side effect, it actually may reduce the energy cost of walking. A comparison of speed and height over the course of a step shows why.

As we walk, our feet exert forces on the ground. The illustration shows the directions of the forces, which have been recorded by means of force plates, instrumented panels set into the floor that give electrical outputs indicating the downward, backward or forward, and sideways components of any force that acts on them. The records show that the force on each foot is always more or less in line with the leg. At stage (b) the foot is pushing forward as well as downward on the ground, so the body is not only being supported but is also being slowed down. At stage (d) the foot is pushing backward as well as down, so the body is being speeded up. thus the body is traveling relatiely fast at stages (a) and (c) of the stride and more slowly at stage (c). For example, the speed of someone walking moderately fast might fluctuate between 1.7 meters per second (3.8 miles per hour) at stages (a) and (c) and 1.4 meters per second (3.1 miles per hour) at stage (c).

To understand the implications of these movements and forces, we need to know about two kinds of energy : Potential energy and kinetic energy. Potential energy is the energy that matter has because

of its height. That energy changes with heights is illustrated, for example, by hydroelectric schemes : as water flows downhill it loses potential energy, which is converted to electrical power. Kinetic energy is the energy that moving objects or fluids have because of their speed. For example, wind loses kinetic energy as it slows while passing over the blades of a windmill, and that energy supplies the power that drives the mill.

Whenever the body is raised it gains *potential energy*. That energy must come from somewhere; when you climb a mountain, for example, the potential energy you gain is supplied as work done by your muscles. The body also gains energy whenever it speeds up, and this *kinetic energy* must come from somewhere as well. When you accelerate at the start of a strint, your muscles do the work that increases your kinetic energy. You might conclude that the muscles must do quite a lot of work in the course of each stride and that they consequently uses a lot of metabolic energy. In walking, however, the body is high while it is traveling slowly and low while it is traveling fast, so its potential energy is high while its kinetic energy is low, and viceversa. The same is true of a pendulum, which is highest when it stops at the end of a swing and lowest when it is moving fastest through the bottom of the swing. As the pendulum swings it converts potential energy to kinetic energy and back again. Energy is swapped back and forth between the two forms, and the pendulum will continue swinging for a very long time without any fresh input of energy. Similarly, very little work is needed from our leg muscles as we walk.

The pendulumlike quality of walking is a consequence of the straightness of our legs. It probably saves metabolic energy, b ıt possibly not very much. If the changes in kinetic and potential energy were less balanced, our muscles would have to do more work at some stages of the stride to increase the total (kinetic plus potential) energy of the body; at other stages they would have to work like brakes, doing negative work to reduce the (kinetic plus potential) energy. While doing positive work they would use *metabolic energy* faster, but while doing negative work they would use it more slowly. The two effects might fairly nearly caneel each other our, but our knowledge of physiology is not precise enough for us to be certain. The pendulumlike quality of human walking probably reduces the "cost of work" element of the food energy requirement, but possibly not by very much.

There is another, possibly more important consequence of our straight-legged style of walking: it enables us to support our weight without the need for large forces in our leg muscles, thereby reducing the "cost of force" element of the energy requirement. When you stand with your knees straight, the line of action of your weight passes close to the knee joints and little tension is needed in your muscles to prevent the knees from collapsing under the load. If you stand with your knees bent, however, the line of your weight is farther from them and your muscles must exert more force. Similarly for walking : the straighter your legs, the less force your muscles need exert. for a practical demonstration, try taking a walk with your knees bent. You will feel the extra tension in the quadriceps muscles (at the front of the thigh), and you will find your steps unusually tiring.

Although our straight-legged style of walking is so economical of energy, only humans use it. Even our closest relatives, the apes, walk with their legs bent. Chimpanzees usually walk on all fours, on the soles of their hind feet and the knuckles of thier hands, but they sometimes walk on their two hind feet only, especially when carrying things in their hands. Even when walking bipedally they walk unlike us, with their knees bent and their back sloping. Gibbons usually travel through the treetops by swinging from their long arms, which are too long for quadrupedal walking. They sometimes walk along the upper surfaces of thick branches, moving on their hind legs alone, with their knees bent. In the wild they seldom or never descend to the ground, but in zoos they sometimes do—and again walk bipedally on bent legs.

7.2.2 Running: Bent Legs for Speed

Humans also move with bent legs—when they run. At the stage of a running step when the force on the foot is largest (stage c in the illustration on the next page), the knee is rather bent and the muscles must exert large forces, at a cost in metabolic energy. This extra cost suggests that running will be expensive of energy. If running is so expensive, and our straight-legged style of walking is so economical of energy, why do we ever run? Why don't we just walk faster?

Despite the apparent extra cost, adult people of normal size change from walking to running whenever their speed exceeds a highly predictable rate, about 2 to 2.5 meters per second (4.5 to 5.6 miles per hour). To try to explain why, we will analyze a simple

model that seems to reduce walking to its essentials. The person shown in the diagram on the facing page is walking at a speed v, keeping the legs utterly straight while the foot is on the ground. The body moves along an arc of a circle centered on the foot, a circle whose radius is the length l of the leg. Here we need a simple formula from physics. An object moving in a circle is constantly changing its direction, turning toward the center of the circle. That change in direction affects the object's velocity, its speed in a specified direction . Because its direction is changing even though its speed may be constant, its velocity is also changing : in other words, the object has an acceleration toward the center of the circle. If the speed is v and the radius l, this acceleration is v^2/l. A force pulling toward the center of the circle is needed to give the object this acceleration and so prevent it from flying off at a tangent. In the case of the walker, the acceleration is downward toward the foot and the force causing it must be the weight of the body, so the acceleration cannot be greater than the gravitational acceleration g :

v^2/l cannot be greater than g.

so v cannot be greater than $\sqrt{gl}$

The gravitational acceleration g is about 10 meters per second squared, and the length of an adult human leg from hip to sole is about 0.9 meter. Thus $\sqrt{gl}$ is about $\sqrt{10 \times 09} = 3$ meters per second. We can conclude that it is physically impossible to walk like the person in the diagram at speeds greater than 3 meters per second. That is the maximum possible walking speed, and people actually change from walking to running at the slightly lower speed of about 2 to 2.5 meters per second. The maximum speed is lower for children because their leg length l is shorter—and small children often have to run to keep up with their walking parents.

That seems to be a convincing explanation of the change from walking to running, until you look at walking races. The rules require the leg to be straight while the foot is on the ground, but athletes nevertheless manage speeds of 4 meters per second. The secret is a peculiar movement of the hips, which lowers the body's center of gravity a little at the stage of the stride when the leg is vertical. The body rises and falls less in each step than we supposed in the simple theory, so the accelerations that are needed at any particular speed are less than we calculated.

If you are not constrained by the rules of a walking race, it is best to change from walking to running at a speed a little less than

the theoretical maximum. The reason is that the energy cost of walking rises steeply at speeds near the limit. Physiologists have shown how the advantage shifts from one gait to the other by measuring the rates of oxygen consumption of people walking and running at different speeds. The subjects walk on a moving belt that keeps them stationary relative to the laboratory, while wearing face masks from which air is sucked through a hose to oxygen-analysis equipment. Plenty of fresh air for breathing is drawn in around the edges of the loose-fitting masks, but all the air the subjects breathe out is sucked away for analysis. the equipment measures the volume of air that passes through and the concentration of oxygen that has been removed by respiration, and from that the rate at which metabolic energy is being used. When energy consumption is plotted against speed for walking and then against speed for running, the resulting curves cross each other at about 2 meters per second. Below that speed, walking is the less energy consuming way of getting around; above it, running is more economical. At any particular speed we use the gait that needs less energy unless there is some special reason (such as the rules of a walking race) for doing otherwise.

Notice that both graphs give rates of energy consumption only for low to moderate speeds. The reason is that at higher speeds muscles work anaerobically and we build up an oxygen debt. Measuring oxygen consumption is a good way of finding out how much energy is needed for locomotion, but only at speeds at which the muscles are working aerobically.

7.2.3 Spring in a Running Step

Runners bounce along in a motion that looks quite different from walking. Running also depends on a different energy-saving principle. When we walk, each foot is on the ground for more than half the stride, so there are stages when both feet are on the ground simultaneously. In contrast, when we run, each foot is on the ground for less than half the stride, so there are stages when both feet are off the ground. Thus running is a series of leaps, and the body is highest when the feet are off the ground. As in walking, the forces on the feet keep more or less in line with the legs, so the body is slowing down at stage (b) and speeding up at stage (d). Thus the body is higher and moving faster at stages (a) and (e) but is lower and moving more slowly at stage (c). Its kinetic and potential energies are highest at stages (a) and (e) and lowest at stage (c).

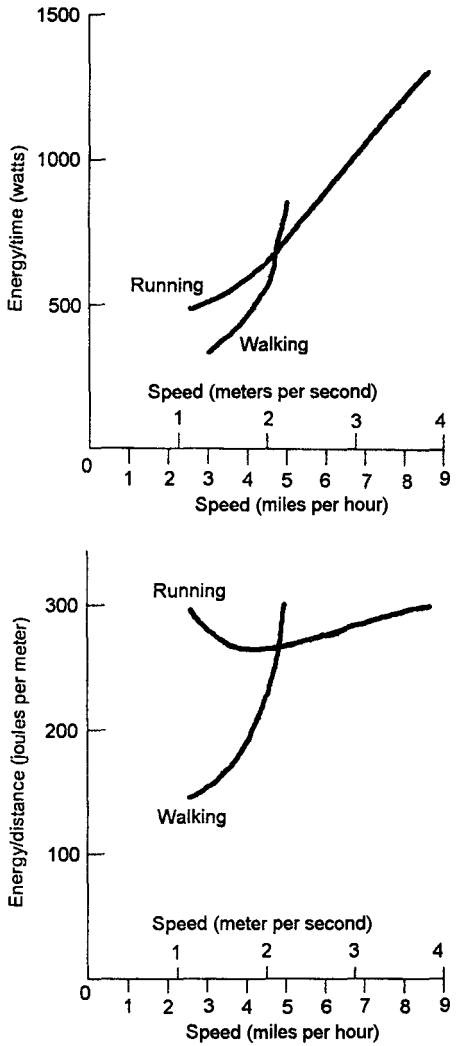


Figure 7.8 In plotting energy consumption against speed, whether we plot energy consumption per unit time or per unit distance, we find that walking uses less energy at speed below about 2 meters per second and running uses less energy above that speed.

There can be no question here of energy being swapped back and forth between the two forms in the pendulumlike manner of walking, but we will soon see how energy is saved by a different energy-swapping principle.

The basic principles of human running are the same as those of kangaroo hopping, even though the two forms of motion look very different. The kangaroo sets both hind feet down on the ground at the same time, but during each hop kinetic and potential energy rise and fall together as in running : they are highest in midair and lowest at the midpoint of the period of contact of the feet with the ground. Again, there is no question of a pendulumlike exchange of kinetic and potential energy. The same is true of the running gaits of quadrupedal mammals (dogs, horses, antelopes, and all the rest).

In all these gaits, the animal travels like a bouncing ball. When a ball hits the ground, it is brought rapidly to a halt, losing kinetic energy, which is largely converted into elastic strain energy as the ball is squashed out of shape. The ball then springs back to its original shape, and the elastic energy is converted back into kinetic energy as the elastic recoil throws the ball back into the air. Similarly, in the case of a running person or a hopping kangaroo, the kinetic energy and potential energy lost at each footfall are converted briefly into elastic strain energy and then returned in an elastic recoil.

Our bodies owe their bounce to springs that stretch to store elastic strain energy and recoil to return it. The most important of these springs are tendons, especially the tendons of muscles in the lower parts of the legs. As the connection between muscle and bone, tendons transmit the force from muscles to the moving joint. The force stretches the tendons whenever it increases and allows them to shorten whenever it falls. Tendons are not very obviously elastic : they stretch only a little before they break. In this respect they are more like ropes than like rubber bands. Yet the small amount of stretch is enough to save the muscles a considerable amount of work.

Tendons from different mammals and different parts of the body all have very similar properties, and all will stretch by about the same percentage of their length when under the same stress. Their elastic properties are best investigated in machines of the kind that engineers use to test the strength and elasticity of metals and plastics. At the top of the machine shown in the picture on this page there is a load cell, an electrical device that measures any forces that are exerted on it. At the bottom is a hydraulic actuator that can be made to move up and down and that is capable of exerting large forces. A tendon dissected from an animal carcass is fixed in the machine, one end held in a clamp attached to the load cell and the

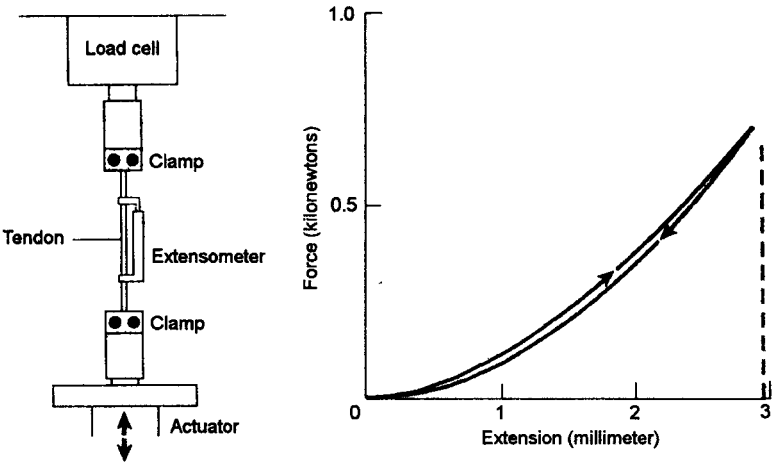


Figure 7.9 A dynamic testing machine used to measure the elastic properties of tendons (left) and the record of a test on a tendon from the hind leg of a wallaby (right).

other in a clamp attached to the actuator. Moving the actuator down stretches the tendon and moving it up allows the tendon to recoil. The machine can be adjusted so that the forces and the rates of stretching are about the same as they would be in running. The tendon would be moist inside the body and must be kept moist in the experiment because its elastic properties would change if it were allowed to dry out. The tendon amy be kept at the temperature of the living body, but the extra bother generally seems unnecessary because the properties of tendons at normal room temperature are almost exactly the same as at body temperature.

The graph shows the result of a test on a tendon from the hind leg of a wallaby, a small kangaroo. The record shows the force increasing as the tendon is stretched and falling as it is allowed to shorten, just as if a rubber band or a steel spring were being stretched and allowed to recoil. The line showing the force during stretching is slightly above the line for the recoil, so the record forms a narrow loop rather than a straight line. The loop in the record shows that the energy returned in the recoil is a little less than the work needed to stretch the tendon, which is inevitable because no material is perfectly elastic. (The energy that is not returned is lost as heat.) A tendon returns 93 percent of the energy, losing only 7 percent as heat; this percentage is good in comparison to rubbers and plastics. If it were a less good elastic material, more

energy would be needed for running, because muscles would have to do work to replace the lost energy. Moreover, our tendons would heat up when we ran and might even get cooked.

The most important tendon for human running is the *Achilles tendon,* which you can feel through your skin, running behind your ankle to attach to the heel bone. To find out how much energy it can save, we need to know how much force acts on it and how much it stretches, because the strain energy stored in a spring is proportional to the force multiplied by the amount of stretch. It would be difficult to measure how much the tendon stretches when we run, but it is fairly easy to calculate that stretch.

From records of people running across a force plate at middle distance speeds, we find that each foot exerts a peak force on the ground of about 2.7 times body weight. The foot presses down with this force on the ground, and the ground pushes up on the foot with an equal, opposite force. This force is distributed over much of the sole of the foot, but for the purpose of calculation we can think of the entire force as acting at a single point, called the center of pressure. Sophisticated force plates tell us not only the size and direction of any force that acts on them, but also the position of the center of pressure. At the stage of a running stride the center of pressure is on the ball of the foot, close to the bases of the toes. The ankle joint is a freely movable pivot, so this force is 2.7 times body weight acting in front of it needs a balancing force behind it. (Similarly, the weight of a child on one side of a seesaw must be balanced by another child on the other side.) The balancing force on the foot is supplied by the muscles of the calf pulling upward on the heel bone through the Achilles tendon. The *Achilles tendon* is about 47 millimeters from the ankle joint, and the line of action of the ground force is about 116 millimeters from the ankle joint, so by the pricniple of levers the force in the tendon must be (116/47) × 2.7, or almost 7 times body weight. This force is about 3500 newtons, or a third of a ton, for a typical (50 kilogram) woman and 5000 newtons, or half a ton, for a 70-kilogram man.

The cross-sectional area of the Achilles tendon in adult men is about 90 square millimeters, so the 5000-newton force sets up a stress in the tendon of about 56 newtons per square millimeter (8000 pounds force per square inch). This stress is about half the stress that would be needed to break the tendon, and it is enough

to stretch it by about 6 percent of its length. If we include the part of the tendon that runs up into the flesh of the calf muscles, the tendon is about 250 millimeters long. When stretched by 6 percent, it extends about 15 millimeters, enough to allow the ankle to bend through 18 degrees. If the tendon were not extensible, the muscle fascicles would have to lengthen and shorten this much more to allow the ankle to make the movements that are needed for running. About one third of the negative and positive work that would otherwise have to be done by the muscle lengthening and shortening is done by the passive stretch and recoil of the tendons. If the tendon were inextensible, all the kinetic and potential energy lost by the body in the first half of the step would have to be removed by the muscles doing negative work and lost as heat. It would then have to be replaced by the muscles doing positive work in the second half of the step. Because the tendon is extensible, however, one third of the energy that would otherwise be lost is stored and returned.

Because the muscles do less work, metabolic energy is probably saved, but even further savings are possible. If the muscle fascicles need not lengthen and shorten so much, the person or animal can make do with muscle fascicles that are shorter so fast, and the leg may be moved by slower, more economical muscle fibers. In either case, the "cost of force" elements of the metabolic energy consumption will be reduced.

In antelopes, horses, and related mammals, the tendons that serve as springs in running are very long and the muscle fascicles exceedingly short. Almost all the movement at the ankle joint, while the foot is on the ground in running, results from the stretching and recoil of the tendons, and the muscle fascicles lengthen and shorten very little. The most extreme example is the plantaris muscle of the camel. The plantaris is rudimentary or even absent in people, but in most other mammals it is one of the strongest muscles of the hind leg. It runs from behind the knee, down the shank, around the heel, and along the foot to the toes. In the camel, its muscles fascicles have almost disappeared. Those that remain are only about 2 millimiters long, buried in the tendon, and they can surely have no significant function. The tendon itself is about 1.3 meters (51 inches) long, continuous from the knee to the toes, and like all tendons must serve as a passive spring. Here is a "muscle" that can exert forces and allow joints to move without any metabolic energy cost.

An arrangement so economical must have a drawback; otherwise all mammals would take advantage of it. In this case, the penalty for saving energy is loss of agility. If a camel's leg is positioned so that the plantaris tendon is taut, the camel cannot bend its ankle joint unless it also bends the knee or the toe joints to slacken the plantaris. This loss of freedom of movement is one of the many reasons that camels are not good at climbing trees.

The Achilles tendon is the most important spring in the human leg, but films of barefoot runners suggest that there is another spring in the foot. The films show that while pressed against the ground the foot is considerably squashed : the ankle is forced about 10 millimeters nearer the ground than if the foot were resting lightly. The flattening of the foot is a consequence of its arched structure. We have already seen that large upward forces act on the ball of the foot and (through the Achilles tendon) on the heel. These upward forces are balanced by a downward force of nine times body weight at the ankle joint, where the tibia (the pricnipal bone of the lower leg) presses down on the joint. The two upward forces and the downward force is between partly flatten the arch of the foot, stretching some of the ligaments that connect the foot bones to one another.

My colleagues Robert Ker and Mike Bennett and I suspected that the arch of the foot might be a spring and wanted to test this idea experimentally. The machine that we had used for stretching tendons was also suitable for squeezing feet, but we could think of no way to use the machine safely on feet that were still in place on people's legs. Instead we used feet that had been amputated by surgeons becauses they were diseased or, in one case, because the knee had been damaged beyond repair in a traffic accident. The tibia was attached to the machine's load cell and the foot rested on two steel blocks, which in turn were supported by the actuator. When the actuator was made to rise, the foot was squeezed and the arch flatened. Rollers below the uppermost steel blocks allowed the slight lengthening of the foot that occurred when the arch flattened.

This simple experiment imitated quite well the forces that act on the foot during running. The upward pressure from the block under the ball of the foot imitated the force from the ground. The other block pressd directly on the heel bone (we had removed the fatty pad of the heel, to expose the bone), and its upward push

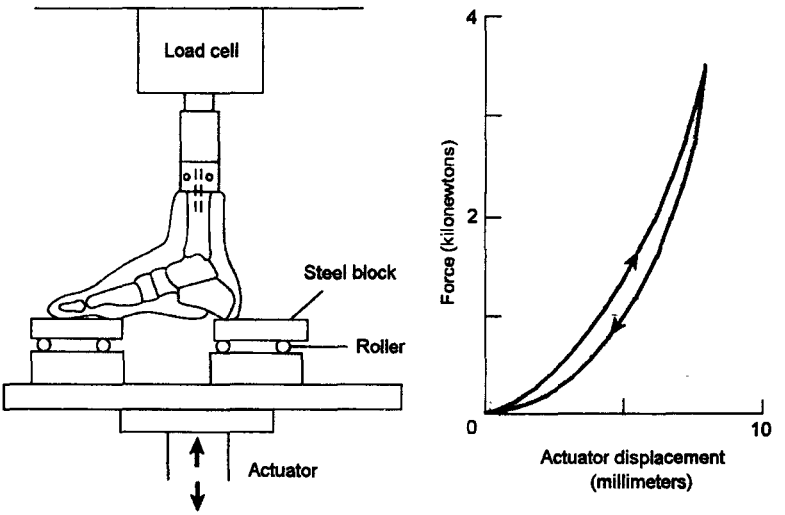


Figure 7.10 The experiment that demonstrated the spring in the arch of the human foot (left) and a typical result (right).

imitated the upward pull of the Achilles tendon. Finally, the downward push from the load cell imitated the force at the ankle joint of the living foot.

The experiments confirmed that the foot is indeed a reasonably good spring. It compressed under load and recoileds immediately when the load was reduced. The records showed loops that were wider than those in the experiments on tendon, indicating that more of the energy was being lost as heat, but most of the energy (78 percent) was returned.

It has been estimated that one third of the kinetic and potential energy that the body loses and regains in a running step is stored in the Achilles tendon and returned in its elastic recoil. The experiments on feet showed that they would store and return a further one sixth of the kinetic plus potential energy. Together, these two springs halve the work that the muscles have to doThe tendon springs of animals such as horses, camels, and antelopes are probably even more effective.

7.2.4 Four-Legged Gaits

Most animals that travel on two legs have a slow gait and a fast one. People and some birds (chickens, for example) walk to go slowly and run to go fast. Other birds, such as American robins, walk at low speeds, but at higher ones they hop. Kangaroo also

have two gaits (the hop is the faster one), but in their cases the solw gait is an awkward shuffle on all four legs and the tail. The walk, the run, and the hop exhaust the possibilities for two legged gaits, but much more variety is possible when four legs are used. Most quadrupedal mammals of the size of cats or larger switch back and forth between three different gaits : the walk, the trot, and the gallop (although really small mammals such as mice seldom walk).

In the quadrupedal walk, as in the human one, each foot is on the ground for more than half the time. The four feet are set down in turn, generally at roughly equal intervals of time and almost always in this order : left fore, right hind, right fore, left hind, left fore, and so on. The legs are not kept as straight as in human walking, and the pendulum effect is less pronounced.

Trotting and galloping are running gaits : each foot is on the ground for less than half the stride. There may or may not be a stage in the stride when all four feet are off the ground, but at least there are stages when both fore feet, or both hind feet, are off. In trotting, the feet move in diagonally opposite pairs, the left fore with the right hind and the right fore with the left hind. Camels and some longlegged breeds of dog use the pace, a gait that is superficially trotlike but in which the two left feet move together and the two right feet move together, instead of the diagonal pattern. Perhaps these animals prefer the pace because if they trotted, the long legs might get in each other's way at the stage when the fore leg swings back and the hind leg of the same side swings forward.

In walking and trotting the left and right feet of a pair are set down at equal intervals; for example, if the right fore foot is set down a quarter of a second after the left, the left is set down a quarter of a second after the right. In a gallop, the intervals are unequal. The two feet of a pair are set down in rapid succession, and then a longer interval follows before the first is set down again. In a full gallop, the two hind feet are set down and then the two fore. The canter, which is sometimes recognized as a distinct gait, is a slow gallop in which the first fore foot is set down at the same time as the second hind.

The footfall pattern of galloping makes it possible for the animals to lengthen the stride by bending and extending the back. While only the fore feet are on the ground, the back bends, pulling the hindquarters forward. While only the hind feet are on the ground,

the back straightens again, pushing the forequarters forward. Thus the body moves farther forward while each pair of feet is on the ground than it would if the back remained rigid. That enables the animals to travel faster or to travel more economically at the same speed.

The principal muscle that straightens the back is connected to the skeleton of the hindquarters by a sheet of tendon (the technical term is *aponeurosis*). This aponeurosis has elastic properties like other tendons and serves as an energy-saving spring, but only in galloping. To see why a spring might be useful, we have to think more about the animal's kinetic energy. It has seen how the body as a whole decelerates and reaccelerates while the feet are on the ground, but we have ignored the movements of the legs, which swing back while their feet are on the ground and forward again while they are off. Kinetic energy is associated with this movement whenever the legs are moving *relative to the body's center of gravity*. At the end of its forward swing and again at the end of its backward one, each leg has to stop and start swinging the opposite way : it has to lose and regain kinetic energy twice in each stride. The faster an animal runs, the faster the legs have to swing and the larger the swings in the kinetic energy. At the stage of the stride when the back is most bent, the fore legs have been swinging back and are about to swing forward and the hind legs have been swinging forward and are about to swing back. Both pairs of legs have to be stopped and started moving again, so kinetic energy has to be lost and regained. That could be accomplished entirely by muscles doing negative work to stop the legs and then positive work to reaccelerate them, but less metabolic energy is needed if some of the kinetic energy is stored as elastic strain energy in the aponeurosis and returned by its elastic recoil. That is what seems to happen.

It has already been observed how, for people, walking is more economical than running at speeds below 2 meters per second, whereas running is more ecoknomical at higher speeds. Dan Hoyt and Richard Taylor of Harvard University showed similarly for ponies that each of the three gaits—walk, trot, and gallop—was the most economical in the range of speeds at which it is used. Just as in the experiments with people, they had the ponies run on a moving belt while the air they breathed out was collected through face masks and analyzed. It is fairly easy to to train ponies (and many other animals) to run on a moving belt, but Hoyt and Taylor achieved

the more difficult feat of training the ponies to walk, trot, or gallop on command. These ponies could be made to gallop at speeds at which they normally would have trotted and to trot at speeds at which they normally would have trotted and to trot at speeds at which they would have preferred to gallop. By analyzing the use of oxygen Hoyt and Taylor obtained a graph of energy consumption per unit distance plotted against speed. The walking and trotting curves cross at 1.7 meters per second, telling us that walking is more economical below that speed and trotting above. Similarly, the trotting and galloping curves corss at 4.6 meters per second; above that speed the advantage shifts to galloping. To travel as economically as possible, the ponies should have changed from walking to trotting at 1.7 meters per second and from trotting to galloping at 4.6 meters per second.

To find out whether the ponies did that, Hoyt and Taylor filmed them moving around their paddock. The two men did not chase the ponies or disturb them in any way, but simply allowed them to move as they chose. They found that the ponies did indeed select the most economical gait, walking below 1.5 meters per second, trotting at 2.8 to 3.8 meters per ssecond, and galloping above 5 meters per second. Furthermore, the ponies generally avoided speeds near the intersections in the graph : they accelerated quickly from walking well below 1.7 meters per second to trotting well above it, and from trotting well below 4.6 meters per second to galloping well above it.

The reason is that ponies need to use less energy per unit distance near the middle of the speed range for each gait and must use more near the transition speeds. For example, a pony that travels at 4.6 meters per second would use 340 joules per meter if it moved steadily at that speed (whether trotting or galloping), but it could cover the same distance in the same time for less energy if it alternated between trotting at 3.5 meters per second and galloping at 6 meters per second using (in each case) only about 300 joules per meter.

People similarly avoid speeds near the walk-run transition. Ultrarunning is the sport of racing over a distance of 100 miles. The best competitors cover the distance in about 13 hours, at a speed of about 3.4 meters per second, running all the way, but many others average speeds of around 2.2 meters per second. These latter competitors walk part of the way at lower speeds and run the

rest at higher speeds, avoiding the uneconomical transition speed. Not surprisingly, different animals change gaits at different speeds: for example, short-legged cats make the changes at lower speeds than do long-legged giraffes.

7.2.5 Walking, Running, and the Design of Ships

Imagine that small animals were exact scale models of large ones. To say the same thing in more technical terms, imagine that animals of different sizes were geometrically similar to one another. If one animal were twice as long as another, it would also be twice as wide and twice as high, and all its bones would be twice as long and have twice the diameter. That is obviously not the case: a 2-kilogram cat is not an exact scale model of a 250-kilogram tiger, nor is a 20-kilogram gazelle an exact scale model of a 800-kilogram buffalo. However, it is more nearly the case than you might suppose, as a simple argument will show. The masses of geometrically similar animals would be proportional to the cubes of their lengths (our imaginary animal that was twice as long, twice as wide, and twice as high as another would have $2 \times 2 \times 2 \times = 8$ times the volume and so be 8 times as heavy). In other words, the lengths of geometrically similar animals would be proportional to the cube roots of their masses, or to $(\text{body mass})^{0\,33}$.

From similarity of shape we move to similarity of movement, for by extending the idea of geometric similarity we can arrive at the idea of dynamic similarity. Two shapes are geometrically similar if they could be made identical by uniform changes in the scale of length. Two motions are dynamically similar if they could be made identical by uniform changes of the scales of length, time, and force.

To understand what that means, imagine you had a film of a small cat running and another of a tiger running. You could change the sizes of the images on a screen by moving the projectors closer or farther away. You could also change the time taken for each stride by running the pojectors faster or slower. If the animals were running in dynamically similar fashion, you would be able to make the two films seem identical.

The movements of cats are remarkably like those of tigers using the same gait. Indeed, there are close similarities of movement even between less similar animals; for example, between cats and rhinoceroses. Would it be reasonable to suggest as a rough approximation that different mammals using the same gait move in dynamically similar fashion?

There is a physical principle that says that movements that are affected by gravity cannot be dynamically similar unless their Froude numbers are equal:

$$\text{Froude number} = \frac{(\text{speed})^2}{\text{gravitational acceleration} \times \text{length}}$$

The principle was first used by William Froude, a Victorian navel engineer who was making tests on small-scale models before building ships to new designs. He wanted to know how much power is needed because ships push waves along in front of their bows. To find out what the waves would be like around the real ships, Froude needed to know how fast he should propel the models to produce the same pattern of waves. He showed that the model speed should be adjusted to make the Froude number the same as for the ship. The wave patterns around model and ship would then be dynamically similar. As always when the Froude number is applicable, gravity is important here—in this case because it tends to flatten the wayes. The same principle has since been applied to other situations in which gravity is important.

The speed and the length that are used to calculate the Froude number can be defined in various ways to suit different kinds of movement. When considering walking and running, it seems sensible to use the forward speed of the body and the length of the legs:

$$\text{Froude number} = \frac{(\text{speed of locomotion})^2}{\text{gravitational acceleration} \times \text{leg length}}$$

For example, camel's legs are about nine times as long as those of cats, so the Froude numbers of these animals are equal when the camel is travelling three times as fast as the cat (the 3 squared in the dividend will cancel the 9 in the divisor).

Many observations have confirmed that the movements of different-sized mammals are fairly nearly dynamically similar when the mammals are traveling with equal Froude numbers. For example, in dynamically similar movement, animals of different sizes would have equal relative stride lengths (stride length divided by leg length). Measurements from films confirm that a graph of relative stride length against Froude number is more or less the same for mammals as different as dogs, camels, and rhinoceroses—and even for bipeds such as people and kangaroos. Recall that the camel and the cat differed in their speed of movement at equal Froude number: the camel traveled three times faster. We can expect a camel to change

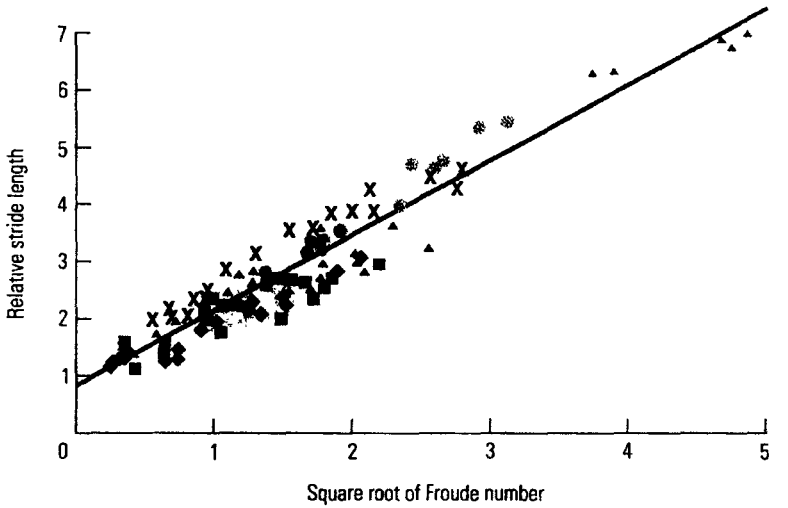


Figure 7.11 Animals of different sizes have about the same relative stride lengths when traveling with equal Froude numbers.

gaits at about three times the speed at which a cat makes the corresponding change. Speed measurements from films show that this is aproximately true if you count the peculiar pace of camels as equivalent to the trot of cats. More generally, mammals change from walking to trotting of pacing at a Froude number of about 0.5 (1.0 meter per second for a cat, 2.9 for a camel) and from trotting to galloping at a Froude number of about 2.5 (2.2 meters per second for cats, 6.5 for camels).

This generalization fits well with a conclusion reached earlier in the chapter, that straight-legged walking is impossible when the speed v is greater than the square root of the gravitational acceleration g multiplied by leg length l —when v is greater than $\sqrt{gl}$. This is equivalent to saying that stiff-legged walking is impossible at Froude numbers v^2/gl greater than 1. The change from walking to running is actually made at a rather lower Froude number, but the example may help to show why Froude numbers are important.

7.2.6 Energy Costs and Size

Although animals as different in size as cats and camels move in a similar manner, very tiny and very large mammals do not. If you look at the whole rangde of land mammals from shrews to elephants, you will see that their movements are not quite dynamically similar; smaller mammals run on bent legs and larger ones run with

their legs much straighter. This has major consequences for the forces the muscles have to exert and for the energy they use.

The definition of dynamic similarity says that dynamically similar movements can be made identical by adjusting the scales of length, time and *force*: in dynamically similar movements all forces are scaled up or down in the same proportion. An animal's weight is one of the forces that acts on it when it runs. Thus if different-sized animals moved in dynamically similar fashion, their muscles would exert forces proportional to their body masses. Each of the muscles of a 200-kilogram lion, for example, would have to exert 100 times as much force as the corresponding muscle of a 2-kilogram cat. The masses of geometrically similar animals are proportional to $(\text{length})^3$, but the cross-sectional areas of their muscles are proportional only to $(\text{length})^2$, or $(\text{body mass})^{2/3}$. The cross-sectional areas of their muscles would be only $100^{2/3} = 22$ times as much in the lion as in a geometrically similar cat. We divide force by cross-sectional area to obtain the stresses in the muscles, which would be $100/100^{2/3} = 100^{1/3} = 4.6$ times as much in the lion as in the cat. That seems bas enough for lions, whose muscles would have to work much nearer their limits of strength than the muscles of cats have to do, but cats and lions are far from the extremes of mammal size. If a 3-gram shrew were scaled up to the size of a 3-tonne elephant and still ran on bent legs like a shrew, the stresses in its muscles would be increased 100 times.

Elephants and other large mammals avoid the need for impossible muscle stresses largely be keeping their legs much straighter than small mammals. Andy Biewener of the University of Chicago has studied the posture of different-sized mammals and the dimensions of the muscles and their positions of attachment; he has concluded that when mammals ranging at least from 90-gram chipmunks to 300-kilogram horses use similar gaits, the peak stresses in their leg muscles are about the same.

It seems obvious that elephants cannot run on bent legs, but why don't shrews run on straighter ones? The forces in their muscles would be less, so they would use less metabolic energy. The most plausible reason suggested so far is that an animal with its legs bent is immediately ready to accelerate or jump, but one standing on straight legs cannot pounce or jump out of the way until it has first bent its legs. Small mammals run on bent legs, and large ones keep their legs straighter; the difference is a matter of size.

Richard Taylor's laboratory is an old missile site in the words outside Cambridge, Massachusetts. There is plenty of room in the surrounding paddocks for the ponies that trotted or galloped on command in the gait experiments, and there is also room for more exotic animals. At various times Taylor has kept kangaroos, gazelles, cheetahs, and even young lions. Indoors he has moving belts to suit animals of all sizes, from 100-kilogram ponies down to tiny chipmunks less than a thousandth of that mass. There he and his colleagues have measured rates of oxygen consumption of a very wide range of mammals (and also of birds and lizards) walking and running at different speeds. They also shipped moving belts and oxygen analyzers to kenya, where they measured mammals ranging from a 600-gram mongoose to a 240-kilogram eland (a large antelope). One of their most ambitious projects was to measure the oxygen consumption of a moving elephant, but even they were daunted by the prospect of building a moving belt strong enough to support the animal. Instead they made the measurements in a zoo while the elephant walked along paths and the oxygen analyzer traveled alongside on a golf cart.

Not surprisingly, this ambitious program of research showed that animals in general (like people and ponies) use oxygen faster when running fast than when walking or running slowly. Also (again not surprisingly), large animals use oxygen faster than small ones traveling at the same speed. More interestingly, the research produced a general equation relating the energy used by a running animal to its speed and body mass. It can be described by an equation,

$$P_v = P_o + Cv$$

P_v is the rate of consumption of metabolic energy when running at speed v and P_0 is the rate of consumption when standing still. C_v is the extra rate of energy use (over and above the standing rate) for running at speed v; therefore C is the extra power (rate of energy use) divided by speed. In other words, C is the energy cost per unit distance:

$$C = \frac{\text{energy cost}}{\text{distance}}$$

C is a constant for each individual animal, but it is different for different species, especially for species of different sizes. You might expect locomotion to be twice as costly for a large animal as for a smaller one hald its mass; in other words, you might expect C/m to be constant (where m is body mass). However, the real situation is

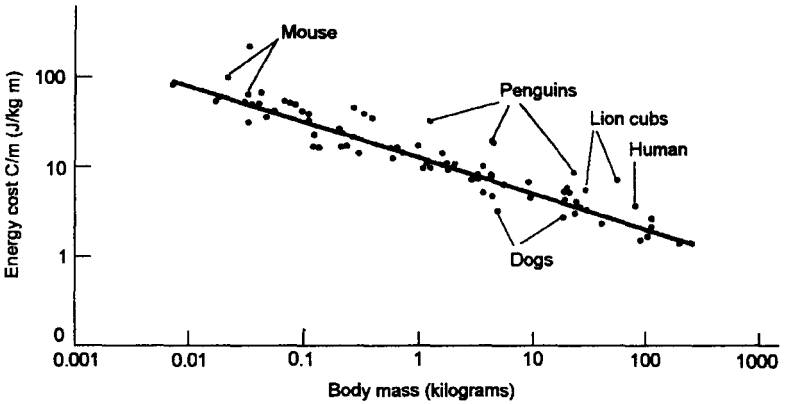


Figure 7.12 The energy cost of walking or running for various mammals plotted against body mass.

more complicated. C is larger for larger animals, but not in proportion to their masses: C/m is *smaller* for larger animals.

The graph on this page shows C/m plotted against body mass. The scales have been made logarithmic so that the distance along the x-axis from 10 grams, to 100 grams, for example, is the same as from 100 grams to 1 kilograms or even from 10 kilograms to 100 kilograms. Plotted in this way, the data points all lie close to a line that has a slope of –0.32. This tells us that C/m is proportional to (body mass)$^{-0.32}$, so the cost of travelling a unit distance, C, is proportional to (body mass)$^{0.68}$. Every time you increase body mass by a factor of 10, for example, you increase the energy cost per unit distance by a factor of $10^{0.68} = 4.8$. Some obviously ungainly animals such as penguins have a higher value of C than the line predicts, but the data are probably not accurate enough to say much about which animals move economically and which do not. Their main value is that they show the general relationship between C and body mass.

Imagine similar animals of different sizes, moving with dynamically similar gaits. The forces involved would be proportional to their body weights, and the distances covered would be proportional to their leg lengths. Thus the work (force times distance) done in corresponding movements would be proportional to body weight multiplied by leg length. The distance traveled in a stride would also be proportional to leg length, so work per unit distance would simply be proportional to body mass—that is, to (body mass)$^{1.00}$. If muscles of different-sized animals worked with the same

efficiency (and there is no very obvious reason why they should not), metabolic energy used per unit distance would be proportional to (body mass)$^{1.00}$. It is actually approximately proportional to (body mass)$^{0.68}$. The discrepancy is serious because we are dealing with a very wide range of body masses. If energy cost were proportional to (body mass)$^{1.00}$, a 1.5-tonne elephant would use 75,000 times as much energy per unit distance as a 20-gram mouse. The energy cost is actually proportional to (body mass)$^{0.68}$, and the elephant uses only about 2000 times as much as the mouse. The different energy costs of running for different-sized animals cannot be explained simply in terms of the work that their muscles have to do.

In 1990 Richard Taylor and his student Rodger Kram offered a new explanation for the energy costs of running of different-sized animals: they asumed that the energy cost of exerting force is much more important thant the cost of doing work. This assumption seems reasonably plausible, because runners need a lot of energy but may not do much work. Running at constant speed on level ground requirs merely enough net work to overcome the air resistance and the friction in the joints. Nearly all the positive work done by muscles at some stages of the stride is cancelled out by negative work done at others. The metabolic energy consumption is increased while the muscles are shortening (doing positive work) and reduced while they are lengthening (doing negative work) and reduced while they are lengthening (doing negative work), but the total energy used in a complete stride may not be much more than if the muscles had exerted the same forces, without either lengthening or shortening. It may be possible to explain the energy cost of running almost entirely in terms of the cost of force production.

The energy cost of force production is expected to be

$$\text{cost of force} = \frac{\text{force} \times \text{fascicle length} \times \text{time}}{\text{economy}}$$

This esquation tells us the energy used in a given time, but we would like to know the energy per unit distance, the qunanity C that we have been discussing. Divide both sides of the equation by distance and remember the speed is distance/time:

$$\frac{\text{cost of force}}{\text{distance}} = \frac{\text{force} \times \text{fascicle length} \times \text{time}}{\text{economy} \times \text{distance}}$$

$$= \frac{\text{force} \times \text{fascicle length}}{\text{economy} \times \text{speed}}$$

Kram and Taylor argued that the less time the feet stay on the ground at each footfall the faster (and so less economical) the muscles must be. For example, a mouse whose paws each stay on the ground for one twentieth of a second in each step needs faster, less economical muscles than a horse whose hooves each stay on the ground for half a second. They suggested that economy might be proportional to ground contact time:

$$\frac{\text{cost of force}}{\text{distance}} \text{ is proportional to } \frac{\text{force} \times \text{fascicle length}}{\text{ground contact time} \times \text{speed}}$$

Ground contact time multiplied by speed is the distance the animal's body travels while one particlar foot is on the ground; this is called the step length:

$$\frac{\text{cost of force}}{\text{distance}} \text{ is proportional to } \frac{\text{force} \times \text{fascicle length}}{\text{step length}}$$

If animals of different sizes were geometrically similar to each other and used dynamically similar gaits, fascicle length and step length would be proportional to each other. As a consequence, the cost per unit distance of generating the required muscle forces would simply be proportional to the forces themselves. The forces would be proportional to body mass, and so the cost per unit distance would be proportional to body mass. That conclusion is identical to the one we reached when assuming that the cost of running was the cost of work. We do not seem to have made much progress toward understanding the energy costs of different-sized animals.

Animals of different sizes, however, do not use gaits that are precisely dynamically similar: small mammals like mice run on strongly bent legs, and large ones like elephants run on much straighter legs. In geometrically similar animals, cross-sectional areas would be proportional to $(\text{length})^2$ or to $(\text{body mass})^{0.67}$. This argument suggests that the forces in leg muscles, and so the energy cost per unit distance (C), should be proportional to $(\text{body mass})^{0.67}$. This conclusion is almost exactly right: measurements of oxygen consumption already described show that C is approximately proportional to $(\text{body mass})^{0.68}$. This explanation of the energy cost of running is impressive and very attractive. It assumes that the energy cost of running is predominantly the cost of force rather than of work, and it assumes that economy is precisely proportional to ground contact time. Our knowledge of muscle physiology is not yet good enough for us to be sure whether these assumptions are sound.

The Kram and Taylor theory seems to explain why many animals that run well have rather long legs for their body mass. For example, a 60-kilogram gazelle has hind legs 0.80 meter long, whereas a 60-kilogram warthog has hind legs only 0.55 meter long. The longer legs of the better runners enable them to increase step length, and the theory tells us that the cost per unit distane should be proportional to 1/step length.

A similar argument provides a further explanation of the advantage of galloping. It has already seen how the work that the muscles have to do, swinging the legs backward and forward, is reduced in galloping by the spring action of an aponeurosis in the back. In addition, bending and extending the back enables the body's center of gravity to travel farther while a foot remains on the ground: it increases the step length and thereby (by our new argument) should reduce the energy cost of running. Greyhounds and cheetahs are fast runners with shorter legs than antelopes of the same mass, but with exceptionally flexible backs. This flexibility suggests that they should be economical runners, but measurements of oxygen consumption while running do not show much difference between cheetahs and other mammals of similar size.

Antelopes, horses, greyhounds, and cheetahs seem to be among the fastest running animals, but again it is difficult to be sure. Many of the animals running speeds that can be found in books are merely subjuctive impressions based on the observer's experience of road traffic. Others are speedometer readings made by driving alongside a running animal in a vehicle, but they too are often unreliable. For example, if the animal tries to swerve away from the vehicle it will be traveling on the inside of a bend and the vehicle on the outside: the vehicle therefore must travel faster than the animal to keep alongside. The only large animals for which maximum speeds have been reliably measured are greyhounds and racehorses. Times given in the sporting pages of newspapers show that most greyhound races are won at 15 to 16 meters per second (34 to 36 miles per hour) and that most horse races are won at 16 to 17 meters per second (36 to 38 miles per hour). These animals have been bred for speed and seem likely to be faster than most wild animals, which have been selected for in the course of evolution not only for speed, but also for other qualities. A good athlete who runs 100 meters in 10 seconds is averaging only 10 meters per second, and the peak speeds of the world's best sprinters are only about 12 meters per second.

Many books will tell you that cheetahs can run at 70 miles per hour (31 meters per second). Such claims seems to be based on a popular article that described a tame cheetah running 80 yards in 2¼ seconds. Unfortunately, according to a later article, the enclosure where the test was made is only 65 yards long, so the speed that is claimed is probably much too high. A later record of 56 miles per hour (25 meters per second) is still astonishingly high, although more believable.

7.2.7 Sprawling Gaits

Birds and mammals run with their feet close under the body, so the lines of footprints made by their left and right feet are close together. Fossil footprints of dinosaurs show that they walked in the same way, but modern reptiles run with their feet well out on either side of the body, so that the lines of left and right footprints are well apart. This manner of running is usually described as a "sprawling" gait. Reptiles move diagonally opposite feet together, the left fore with the right hind and the right fore with the left hind. This sequence is the same as in mammalian trotting, but lizards also use it for slow gaits. They bend their bodies from side to side as they run, timing the bending so that it increases the length of their steps. Until recently, zoologists assumed that the sprawling style of running was uneconomical because it required large forces in muscles. Only when the rates of oxygen consumption of running lizards were measured was it realized that they run as economically as mammals. The energy cost per unit distance is about the same as for mammals of equal mass, and the metabolic rate while standing still (P_o) is considerably less than for mammals. Why then did posture change in the course of evolution, from the sprawling stance of early reptiles to the mammalian stance with the feet tucked in under the body?.

Dave Carrier of the University of Michigan showed that they stop to breathe; running and breathing use the same muscles in different ways, and so the two activities cannot be performed simultaneously. In running, the side-to-side movements of the body require the muscles to the left and right sides of the trunk to contract alternately. In breathing, however, the muscles of the two sides must act together to enlarge the ribcage and draw air into the lungs, or to compress it and drive air out. Mammals have no such difficulty: indeed, the movements of galloping seem to help breathing. The mammalian back, bending up and down instead of from side

to side, works like a bellows. When it bends it squeezes the body cavity, driving air out of the lungs, and when it straightens it draws air in. Records of the flow of air through galloping horses nostrils show that they take one breath per stride, breathing out as the back bends and in as it extends again. The sprawling gait can be fast as well as economical. As a general rule, lizards can sprint about as fast as mammals of equal mass. Quite small (50-gram) lizards have been timed at the remarkably high speed of 8 meters per second (18 miles per hour).

In contrast, tortoises are notoriously slow. Their slow muscles are very economical. If you try to ride a bicycle very slowly, you will probable wobble and fall off. Similarly, walking very slowly requires more precise control of the forces on the ground than walking faster or running. The reasons are not the same as for slow bicycling (there is nothing in walking comparable to the stabilizing gyroscope action of bicycle wheels), but the effect is the same: slow movement requires more precise control. Whatever the speed of walking, the feet exert fluctuating forces on the ground, and the body is generally not in equilibrium. At one stage the forces on the feet may total more than body weight and accelerate the body upward, but at another they may be smaller and let the body fall. The body will rise and fall in the course of a stride but this may not matter, unless the vertical movements are so big that the belly bangs on the ground. Similarly, temporarily unbalanced horizontal forces may make the animal speed up and slow down during each stride, or veer from side to side. Imbalance between the forces exerted by different legs may tilt the animal, causing it to pitch and roll. One can think of the rising and falling, speeding and slowing, pitching and rolling as unwanted movements, but walking remains effective if they are not too large.

When an animal goes fast, footfalls follow each other in rapid succession, and an unwanted movement started by an unbalanced force at one stage of a stride can soon be corrected. If the animal is moving slowly, however, the unwanted movement movement will continue for longer before there is an opportunity to correct it, and it may go too far. If the animal is moving very slowly, force fluctuations must be kept small. The problem is especially severe for a low-slung animal, such as a tortoise that supports its shell close above the ground. A small loss of height or a tilt through a small angle may make it hit the ground, which presumably would be unsatisfactory. A quick calculation will make the problem more

obvious. Imagine that an animal's legs suddenly stopped exerting any force, so that the trunk was completely unsupported. If the animal were a tortoise with its belly initially 5 centimeters from the ground, it would hit the ground after only 0.10 second. For a tortoise of this size, each stride would last about 2 seconds, so the shell would hit the ground if the animal was unsupported for a mere twentieth of a stride. However, if the animal were a dog with its belly 40 centimeters from the ground , it would take 0.28 second to fall. The dog's strides would each last about 0.5 second, so in its case the falling time would be half a stride period. The tortoise could tolerate only a tiny unsupported fraction of a stride (or smaller force fluctuations for larger fraction), but the dog's feet could be off the ground for quite a lagre fraction of a stride without ill effects.

In theory it is possible for a tortoise or other quadruped to keep perfect equilibrium through each stride without rising, falling, pitching, or rolling at all. Three feet are enough to support a body in stable equilibrium: a three-legged stool is stable, but a two-legged one is unstable and will fall over. The tortoise could remain stable by moving one foot at a time, always leaving three on the ground for support. If you sit on a three-legged stool and lean over too far to one side, it will fall over. All is well so long as the center of gravity (of you and the stool combined) is vertically over the "triangle of support" formed by the stool's three feet, but if the center of gravity moves outside the triangle, you will topple. Similarly, an animal standing on three feet cannot remain in equilibrium if its center of gravity is not over the triangle of support. The problem disappears if the feet are big enough and can be placed directly under the center of gravity (birds and people can stand on one foot), but animals with small feet need three, suitably placed.

It turns out that it is enough for a quadruped to move one foot at a time: to keep the center of gravity always over the triangle of support, it must move its feet in a particular order. This order, shown in the diagram on this page, is actually used by dogs, horses, and most other quadrupeds whenever they walk. Tortoises, which have far more need to keep close to equilibrium, also move their feet in this order, but unlike large mammals they do not move their feet at equal intervals. Each fore foot is set down only slightly before the diagonally opposite hind foot, so the footfall pattern is very like that of lizards (which move diagonally opposite feet together), and there are times when only two feet are on the ground.

At first sight this gait seems faulty, but Alan Jayes and I were able to show that, for an animal with slow-acting leg muscles, it would allow steadier walking than if the feet moved one at a time. The reason is that the "ideal" gait in the diagram can keep the animal in constant equilibrium only if the forces exerted by the other feet can be changed instantaneously, whenever a foot is lifted or set down. If the muscles are incapable of rapid adjustments, the animal can keep closer to equilibrium by moving its feet in diagonally opposite pairs. Keeping close to equilibrium is easier with six or more legs. Insects generally move their six legs in two groups of three, the two groups taking turns to provide the triangle of support. Each group consists of the front and rear legs of one side of the body and the middle leg of the other.

Centipedes may have to use gaits that keep their many legs out of one another's way. *Scolopendra* is a short-legged centipede that moves the legs of each side of the body in sequence from front to back: leg 1 before 2 before leg3, and so on. In a group of adjacent feet that are on the ground, the ones in front are at a later stage of the step than the ones behind, so the feet on the ground form clumps, as the diagram on this page shows. This gait presents no problem for *Scolopendra*, but if the legs were much longer they would cross over each other so that the foot of leg 2 was set down in front of foot 1, and foot 3 in front of foot 2. It is hard to see how the centipede could do that without getting into a tangle. Another diagram shows how *Scutigera*, a long-legged centipede, avoids the difficulty. It moves its feet in reverse sequence, starting at the rear. The result is that the feet are spread out instead of being clumped or crossed.

Insects, centipedes, and other arthropods use sprawling gaits, keeping their feet well out on either side of the body. For reasonably large land animals, sprawling gaits are optional: lizards and crabs use sprawling gaits, but birds or mammals of similar mass do not. For small animals such as insects, sprawling is probably essential: if they did not stand with their feet well out to either side of the body, they would be apt to be blown over by gentle breezes. The reason is that the force of the wind on an animal is proportional to its surface area, but the weight that helps to stabilize the animal is proportional to its volume. For geometrically similar animals of different sizes, smaller ones have larger ratios of area to volume and so are more likely to be blown over.

8

Specific Physical Displacement

8.1 CLIMBING

The animals that inhabit trees are said to lead an arboreal or scansorial mode of life. Climbing on the part of the arboreal animals is because of the necessity to procure food and safety. Relatively feeble and defenceless creatures may take to the trees for safety and retreat. There, they get abundant and easily obtainable food. From the stand-point of view of adaptation of arboreal forms, three different types of adaptations can be distinguished in them, as follows:

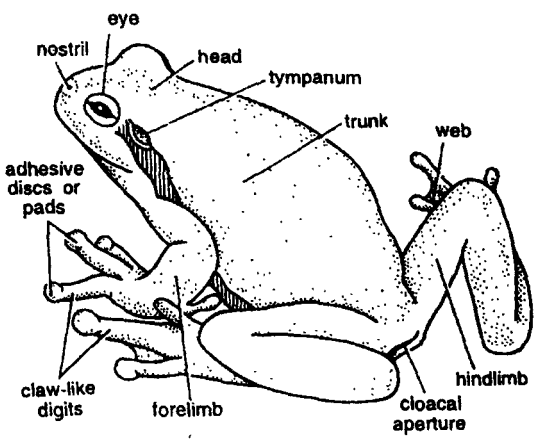


Figure 8.1 *Hyla.*

8.1.1 Categories of Scansorial Animals

On the basis of bionomic classification (*i.e*, habitat) scansorial animal may be of three types:

8.1.1.1 Wall and rock climbers

The classification of scansorial animals from the stand point of their adaptation groups them into three subdivisions, of which the first are the wall and rock climbers. These are not necessarily tree-inhabiting at all, but are, like the gecko lizards, well suited for climbing on the walls of buildings as well as on similar surfaces in nature. The geckoes are, however, a very old and widely distributed group, and the range of their individual adaptation is great, hence it may well be that their scansorial adaptation is after all a response to arboreal life, and that the peculiar structure of their climbing organs rendered subsequent rock-climbing possible.

Among mammals there is a genus of flying squirrels limited to high altitudes at Gilgit and perhaps in Thibet, and thought to live on rocks, perhaps among precipices. Here, again, we have a form whose ancestry may have been aboreal, but if not, it would afford an interesting instance of volant adaptation without an intermediate arboreal habitat.

8.1.1.2 Terrestrio-arboreal forms

The second category, the terrestrio-arboreal, embraces a number of carnivores, rodents, and insectivores which while capable of climbing, nevertheless are still perfectly at home upon the ground beneath the trees. They may nest in the trees with more or less extensive terrestrial excursions during the daytime, or they may climb for food and live on the earth unless impelled by hunger. Their climbing adaptations are very marked.

8.1.1.3 Arboreal forms

Still a third group embraces the wholly arboreal types, creatures which make the trees their home, and while some occasionally descend to the ground as in certain primates (gibbon), their terrestrial progression may be slow and laborious compared with that in their true habitat. Arboreal forms, according to their mode of locomotion, may be grouped in the following subdivisions:

8.1.1.3.1 Branch runners

Like the squirrels, marsupials, lemurs, and chameleons, which live and progress on all fours on the upper surface of the branches. The group embraces, nevertheless, some instances of very perfect arboreal adaptation, as the great majority of tree-dwellers are here included:

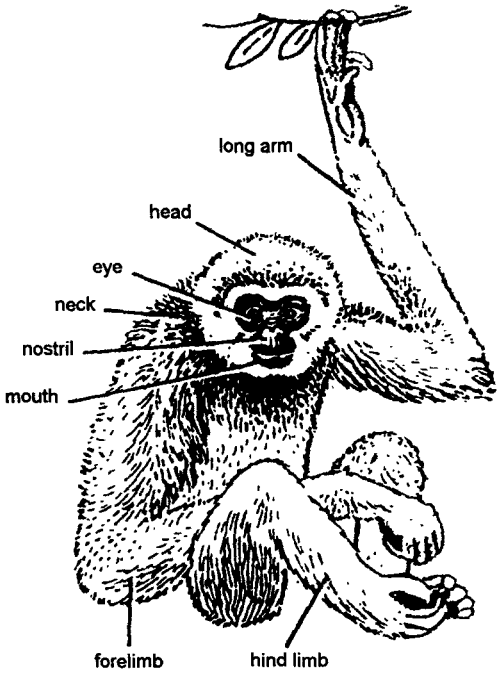


Figure 8.2 *Hylobatus* (Gibbon).

8.1.1.3.2 Forms suspended beneath branches

The sloths for instance, are so constituted that they cannot walk upon the branches but rest and move suspended from them by the powerful recurved claws of all four limbs. Sometimes when quiescent, if a convenient branch lie sufficiently near, the sloth may rest his back on it and relax the hold of one or more of his feet, but the inverted position is rarely reversed. On the ground the animal progresses with the utmost difficulty. The bats should perhaps also be included under this head, as they rest suspended by the hind limbs, head down. The same position of rest is assumed by the so-called flying lemur, *Galeopithecus*, really not a lemur at all but an insectivore.

8.1.1.3.3 Forms swinging by the fore limbs

These forms show a very remarkable method of progression by means of the fore limbs, swinging with great speed and accuracy from limb and from tree to tree. The hind limbs are comparable to those of the tree-dwelling marsupials and the creatures rest and progress on the tops of the branches at times, although the fore

limbs are almost the sole organs of more rapid locomotion. Here long many of the primates, more especially the great or manlike apes, and our pre-human ancestors.

8.1.2 Modifications

8.1.2.1 Body

Climbing adaptation, as in the other lines of adaptive radiation, implies certain body modifications as well as those of limbs. Body contour is of little moment in climbing, but strengthening of chest and ribs and of shoulder and hip girdles is of importance. Nevertheless, in thoroughly arboreal types the section of the thorax anteriorly is subcircular, and the ribs are much curved, in contrast with the compressed thorax and flat anterior ribs of quadrupedal running types. The ribs, especially in the sloths, are numerous and afford ample support to the contained viscera in their inverted position. The dorsolumbar series of vertebrae is often elongated, especially in the tree-sloths of the genus *Cholaepus*, where the number has apparently been increased from about nineteen (normal for the order) to from twenty-five to twenty-seven as a response to arboreal need. The same is true of other forms. *Capromys*, an arboreal rodent, has twenty-three as compared with the normal nineteen, and *Dendrohyrax*, the only arboreal ungulate now alive, has six more than its terrestrial, hoofed relatives.

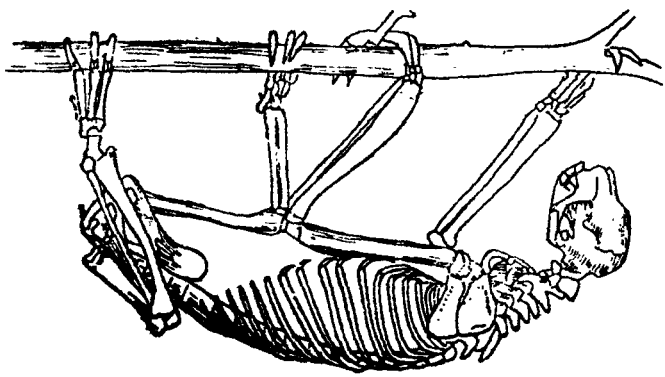


Figure 8.3 Skeleton of Sloth.

8.1.2.2 Limb girdles

The shoulder girdle especially is strong in that both elements, the clavicles and scapulae, are well developed although in terrestrial types the clavicles tend to diminish, even in closely related forms,

and may entirely disappear as in cursorial quadrupedal forms. The fore-and- aft swing of the limb of a deer or horse would be distinctly limited by a clavicle, but in a climbing type whose arms are subjected to much more varied and violent strain the clavicle is very essential, as it withstands the compression of the powerful breast muscles. The scapula is also well developed, but not exceptionally so.

8.1.2.3 Pelvic girdle

The ilium or hip-bone especially shows modification in such types as the sloths and primates as it is broadened out as a support for the viscera. This is markedly true of the sloths, whose inverted posture necessitates additional support, since the mesenteries or membrances which sling the intestine to the dorsal wall lose much of their efficiency when the body is erect or inverted.

8.1.2.4 Limbs

In contrast with the cursorial types, it is the proximal limb segments which now elongate, especially in the suspended and brachiating forms, those of the sloths again being very long, while in the great apes the relative length bears a direct ratio to the creature's climbing powers, reaching the extreme in the gibbons (*Hylobates*) whose arms are so long that the knuckles of the hand touch the ground when the animal stands erect. The progress of the gibbon from branch to branch and tree to tree is remarkable. Climbing forms are generally plantigrade, some of the raccoons secondarily so. In certain lemurs (*Tarsius; Galago*) the tarsus may be elongate, but this is probably due to the fact that the creatures leap as well as climb and the elongation of this segment is a response to the former need rather than to the latter.

8.1.2.5 Feet

The feet of arboreal animals may be either prehensile, that is, grasping, with more or less opposable digits, or nonprehensile. In the *non-prehensile type* the claws may to be well developed as in the squirrels or the cats, giving a fairly tenacious hold. In the Canada tree-porcupine (*Erethizon*) the plantigrade feet are armed with long curved claws, in addition to which the soles bear spines and tubercles which aid in climbing.

Adhesive pads either on the tips of the digits or on the soles of the feet occur several isolated instances, such as the tree-frogs, geckoes and *Dendrohyrax* among mammals. The frogs are aided by a stickily secretion of their pads. "Tree frogs, when hopping on

to a vertical plane of clean glass, slide down a little probably until the secretion stiffens, or dries into greater consistency. Wet leaves or moist glass-walls afford no hold. The adhesion of these frogs is assisted in most cases by their soft and moist bellies, just as a dead frog will stick to a pane of glass".

The geckoes, by means of their adhesive digits, climb up absolutely smooth and vertical surfaces, or, back downward, along a whitewashed ceiling. The apparatus, Gadow says, is complicated in its minute detail, but very simple in principle. The adhesion is effected not by sticky matter, but by small and numerous vacua.

Dendrohyrax, the tree-hyrax, is allied to the coney of Scripture. The tree-hyraxes frequent the trunk and larger branches of trees, sleeping in holes high up in the big trees, especially, according to Roosevelt's observations, the cedars. The adhesive organs have been described by G.E. Dobson, who says that these animals are enabled to climb perpendicular walls and trees without the use of claws. The thickly padded, tuberculated soles are drawn up by certain flexor muscles, thus leaving a partial vacuum by means of which the animal retains its hold.

The primitive type of prehensile foot has been developed in the two great mammalian groups, that of the marsupial being represented by the opossum (*Marmosa*), that of the early placentals by the creodonts, the archaic flesh-eating mammals, the foot of which has been shown to be a terrestrial modification of a grasping type.

Feet of the *prehensile type* are found today in the marsupials and primates. In the former group it is the hallux or great toe which is offset so as to oppose the fourth digit, the second and third being bound together in a common integument (syndactyly) and so slender that their combined strength about equals that of the outermost of *fifth* digit. In marsupials which have become terrestrial the offset great toe has become vestigia or may entirely have disappeared, as in the kangaroos. In the primates, while the foot is perhaps most apt to show this opposable first digit, it also exists in the hand, although it is nowhere developed to the degree shown in mankind, wherein the final perfection of the hand as a organs of prehension has developed since its release from the necessity of arboreal locomotion.

Syndactyly has already been referred to as occurring in the marsupials, and even such as are no longer tree-inhabiting, like the kangaroo, still exhibit this feature in unreduced condition. It doubtless

arose primarily, however, as an arboreal adaptation. The koala shows a rather remarkable modification for climbing, for the foot has a long, widely offset, clawless great toe, syndactylous second and third toes, which are clawed, and powerful clawed fourth and fifth toes, the former being the longer. The hand, on the contrary, has five subsequal digit, all of which bear sharp claws; but two digits, numbers 1 and 2, oppose the other three. Its clinging powers are so great that event death will not dislodge the creature from the tree in which it is shoot.

Among reptiles, the true African chameleons (*Chamaeleon*) exhibit remarkable syndactyly, as it extends to both fore and hind feet. On the hand the first three fingers form the inner bundle and are opposed to the outer two which are likewise syndactylously bound. The foot is similar but reversed, in that the inner bundle contains two, the outer one three digits. These very admirable grasping organs are supplemented by a prehensile tail, so that the creature is very firmly anchored in position, which is rendered necessary perhaps in part by its method of securing insect prey by the unerring aim of the enormously extensile tongue.

In the so-called scansorial birds such as the parrots. woodpeckers, and the like, the outermost toe has been rotated backward in such a way that it and the hallux oppose the second and third toes, the fifth, as in all birds, being absent. This gives a very firm grasp for the actual grip of a branch as in the parrots, or, reinforced by strong claws, enables the animal to cling to the roughened bark of a tree trunk. In the parrots, woodpeckers, and cuckoos the rotation of the outer toe is permanent and the foot is called *zygodactylous*; certain other owls, etc, may turn it backward or not at will.

8.1.3 Digital Reduction

While arboreal forms usually have need of all of their digits, occasionally one sees *digital reduction*. The foot of the koala, with syndactylous second and third toes, functions as fourtoed, even though consisting structurally of five; certain of the primates (lemurs), on the other hand, some of which resemble the koala superficially very much, have actually lost the second digit so that the opposability of the first in grasping a limb is unimpeded. In the lemur (potto, etc.), the fourth digit is the largest as in the koala. Digital reduction is also seen in the tree-sloths, the two-toed sloth *Cholaepus* having but two in the hand and three in the foot, while in the three toed

sloth *Bradypus* there are three in each, and the hand and foot are both somewhat elongated, especially in the powerful hook-like claws which, like the feet of the koala, retain their grip on the length even after the animal has been shot.

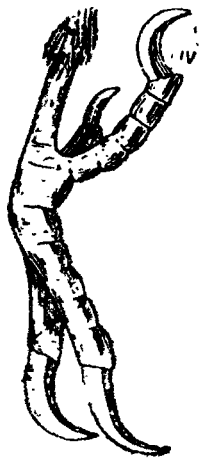


Figure 8.4 Foot of wood-pecker.

8.1.3.1 Tail

The tail may be prehensile or not as in the case of the feet. If non-prehensile, there are ectodermal spines or scales on the under side, as in the flying squirrel *Anomalurus*, which prevent the animal from slipping down. The same effect is produced in the woodpecker by stiff spiny feathers which are braced against the tree trunk to which the creature clings. The posture is familiar, and enables the bird to drill into the wood for the grubs upon which it feeds, or to excavate cavities for its nest or for the storage of food.

Prehensile tails are found in a number of unrelated instances, as, for example, the chameleon lizards which have been mentioned, the opossums, the tamandua which is one of the anteaters, and certain of the New World monkeys (Cebidae) such as the spider monkey, the howlers, and the capuchins. Where the prehensile powers are well developed the tail is naked on the under surface near the tip. One of the most perfectly adapted of these forms is the spider monkey, *Ateles*, in which the tail is highly prehensile and functions as a "fifth hand". Perhaps as a correlation with this excellent grasping organ the real hands have lost the thumb, but the four long digits which remain form a splendid hooklike device for suspending the body. Not all South American monkeys have a prehensile tail; on the other hand, none of the Old World forms do, so that its presence is diagnostic of a New World ape.

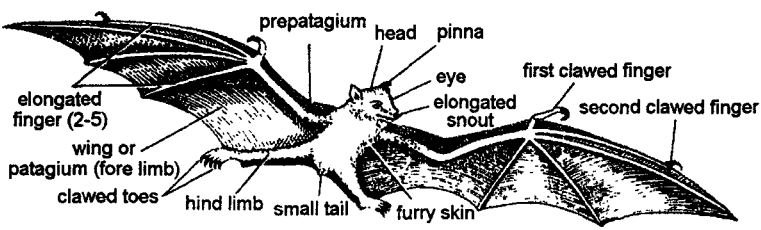


Figure 8.5 *Cynopterus*.

8.1.4 Development of Accessory Organs

Due to scansorial habits various accessory climbing organs might be developed. These include spines and tubercles on the forearm in some lemurs (*Hapalemur griseus*). In *Lemur catta* there is a specific climbing organ, which is composed of a patch of hardened skin on the forearm, which projects to a large extent. Both these organs have glands connected with them.

8.2 JUMPING

Animals jump for many different reasons. Lions jump onto prey to bring them down. Frogs and grasshoppers jump in directions that are hard to predict, to escape from danger, Squirrels and bushbabies jump between the branches of trees. Fleas resting on leaves jump when they senses the warmth of the body of a passing mammal of bird: with luck, they may be able to attach themselves to the potential host.

The secret of a good jump is speed. A fast takeoff will carry an animal closer to its target or farther from its enemy. As we will see, it turns out that the smaller the animal, the more difficult it is to reach a good jumping speed by the action of muscle alone. The muscles of smaller creatures need an assist, and for this reason we will find the most varied, and the most bizarre, jumping mechanisms in insects. Indeed, there is one type of beetle that manages to jump without even using its legs.

What ever the function of the jump, there is a simple relationship between the speed and angle at which the animal takes off and the distance it covers or the height it clears. It is particularly easy to calculate the height of a vertical jump. The kinetic energy of a body of mass m travelling at speed v is $\frac{1}{2} mv^2$. If the body rises

Figure 8.6 Bushbabies (*Galago senegalensis*) leap between branches in their forest habitat.

through a height h, it gains potential energy mgh (g is the gravitational acceleration). In a vertical jump all the kinetic energy at takeoff is converted into potential energy at the top of the jump.

$$mgh = \frac{1}{2}mv^2$$

The height is therefore

$$h = \frac{v^2}{2g}$$

In these equations, h is the increase in height to the body's center of gravity from the instant of takeoff to the top of the jump. For example, in a standing jump a human athlete might leave the ground with a vertical velocity of 3 meters per second, and the center of gravity would rise about $3^2/(2 \times 10) = 0.45$ meter. (The gravitational acceleration is about 10 meter per second squared.) The body's center of gravity is a little above the hips, so it would be about 1 meter from the ground at takeoff and would rise to 1.45 meters at the top of the jump. Senegal bushbabies are very much smaller than we are, weighing only 250 grams, but these tree-dwelling primates are very much better jumpers. A pet bushbaby was once observed jumping from the floor to the top of a 2.26-meter door. Its center of gravity could have been no more than 0.25 meter from the floor at takeoff, so the height gain h was 2 meters, which requires a takeoff speed of more than 6 meters per second.

People often rates fleas as the most remarkable of animal jumpers. Though they are so small, fleas can jump heights of at least 130 millimeters (5 inches). In absolute values their jumps are low in comparison to human jumps, but relative to body length they are much higher. Excluding the legs, flea's body is only about 2 millimeters long and a man's body (again excluding the legs) about 1 meter. Therefore it might be argued that a flea's 130-millimeter jump is equivalent to an impossible 65-meter human jump.

The fallacy of that argument was pointed out in 1950 in an article by the famous muscle phyiologist A.V. Hill. Hill wanted to figure out how the height of a jump depended on the size of a jumper. He would then be able to compare the expected jump heights for different-sized animals, including a flea and a man. He first argued that the work available for a jump should be proportional to body mass. The work done in a single muscle contraction is the

force multiplied by the shortening distance. That force is proportional to the cross sectional areas of the muscles, and we can expect the muscles to be able to shorten by amounts proportional to their initial lengths. Therefore, the work done in a single contraction is proportional to the cross-sectional area multiplied by the length—that is, to the volume of the muscle. If the animals being compared have equal proportions of muscles in their bodies, we can expect the work available for a jump to be proportional to body mass.

The work required for a jump to height h equals the potential energy gain, mgh. If this work is proportional to body mass m in animals of different sizes, all should be able to jump to the same height h. The flea that raises its center of mass 0.13 meter in a jump is doing much less well than a human who rises 0.45 meter. We should not ask why fleas jump so high, but why their jumps are so feeble.

Hill's theory said that different-sized animals should jump to equal heights, but ignored a problem faced by very small jumpers. People, bushbabies, and fleas all prepare for a jump by bending their legs. To take off, they extend the legs rapidly, accelerating over a distance of about 0.4 meter for people, 0.16 meter for bushbabies, and only about 0.5 millimeter for fleas. Humans accelerate from rest to about 3 meters per second at takeoff, so the average speed, as the legs extend, is about 1.5 meters per second; thus the 0.4-meter acceleration distance is covered in about 0.27 second. Films confirm that this estimate is about right. The simple equation on page 59 tells us that a flea jumping to a height of 0.13 meter would have to take off at 1.6 meters per second. (It would actually have to take off rather faster because the equation ignores the effect of air resistance, which slows small jumpers much more than large ones). As the flea accelerates during takeoff from rest to over 1.6 meters per second, its *average* speed would be a little over 0.8 meter per second: it would cover the 0.5-millimeter takeoff distance in only 0.6 milliseconds.

No muscle can complete a single contraction in so short a time. Some midges beat their wings at 1000 cycles per second; a midge has only 0.5 millisecond for each up or down stroke. Such frequencies depend on the wings and their muscles going into a state or resonant oscillation. A jump requires a single contraction to extend the legs, and no known muscle could do that in as little as 0.6 milliseconds.

The flea's jump is possible only because the animal has a built-in catapult, a tiny block of rubber like protein at the base of each in leg. These blocks are too small for mechanical testing but seems to be made of a protein called resilin that is also found in the tendons of some dragonfly wing muscles and in the hinges that join locust wings to the body. Tests on these larger (but still very small) pieces of resilin show that its properties are very like those of soft rubber. It can be stretched to three times its initial length and returns nearly all the energy in its elastic recoil.

When children use catapults they first stretch the rubber, storing up elastic strain energy in it. If the catapult is a strong one the child may be unable to stretch it fast, but a slow stretch will store the elastic strain energy equally well. When the catapult is released, the rubber recoils very rapidly indeed, projecting the stone much faster than it could have been thrown. Work done slowly on a catapult is returned much faster. Fleas store up strain energy in their resilin by a relatively slow contraction of their leg muscles lasting, it seems, about 100 milliseconds. A trigger mechanism releases the resilin, which recoils, extending the legs in only 0.6 millisecond and throwing the insect into the air.

The strong hind legs of a locust launch it into the air at a speed of up to 3.2 meters per seconds, fast enough to propel the insect about 0.5 meter (20 inches) if it jumped vertically. In fact locusts jumps at an angle and rise less high but can clear horizontal distances of about 1 meter. Like fleas, they must rely on elastic recoil, but instead of using a tiny block of resilin as a catapult they store elastic strain energy by means of a tug-of-war between two muscles.

About halfway along the locust leg is a joint called the knee, by analogy with the human leg. Kept bent when the animal is standing, the joint extends very rapidly to throw the animal into the air. A large extensor muscle filling most of the locust's fat thigh extends the knee and does most of the work needed to power the jump. A much smaller flexor muscle bends the knee and prevents the joint from extending too soon. Like many other insect muscles, these are pennate; the fibres converge on a central tendon made up of material very like the exoskeleton that encloses the whole body. When the insect is preparing for a jump, both muscles contract. The extensor is much the stronger but, while the knee is bent, its tendon runs much closer to the axis of the joint than does the tendon of the

flexor. This difference of lever arms enables the relatively weak flexor to hold the knee bent, against the stronger pull of the extensor. The large forces that are developed at this preparatory stage compress the semilunar processes (flexible parts of the leg skeleton, close above the knee) and stretch the extensor tendon, storing elastic strain energy in both. To jump, the animal suddenly relaxes the flexor muscle, allowing the knee to extend, which it does very rapidly, driven by the elastic recoil of the semilunar processes and the extensor tendon.

One of the most curious of jumping insects is the click beetle, which can jump twice as high as a flea without so much as bending a leg. The click beetle flexes its back instead. These beetles are often found crawling on blades of grass, but drop to the ground if disturbed. If they land upside down and cannot right themselves quickly, they jump by a sudden jackknife movement of the back that may throw them as high as 0.3 meter (12 inches) off the ground. The jump has been filmed using a special camera running at 3100 frames per second. (Ordinary films are taken at only 18 to 24 frames per second). Even at this exceedingly high rate, the films showed the beetles accelerating to takeoff in just two frames, or 0.6 millisecond. As in flea and locust jumping, the rapid acceleration of a click beetle jump is made possible by a catapult mechanism.

All these insects take off so quickly that there is scarcely time for muscles to start shortening during takeoff; the jumps are almost entirely powered by elastic recoil. Larger jumpers such as bushbabies and human beings have time to shorten their muscles during takeoff, but these more muscle-dependent jumpers also benefit from tendon elasticity. If you ask people to do a standing jump, they will generally bend their knees and then immediately extend them. If they pause with knees bent and then jump, they cannot jump so high. The reason is that muscles can exert more force while being stretched than while shortening or holding constant length. The muscles that extend the legs at takeoff also stop the downward preparatory movement; as they stretch to counteract the binding of the knee, they are able to develop large forces. These forces stretch the tendons in turn, storing strain energy that is returned in elastic recoil, increasing the height of the jump.

8.2.1 Swinging Through Trees

Squirrels and many monkeys travelling through the forest run along the tops of horizontal branches much as they run on the

ground, but leap as necessary from one branch to another, often moving from tree to tree; Bushbabies and lemurs move through trees in a different way, clinging to a vertical branches and leaping from one to the next, But some tree dwellers have found an efficient way to move through the forest while barely using their legs at all—these animals propel themselves by the strength of their arms.

Apes and a few monkeys swing by their arms below branches, using the technique called brachiation. Particularly good at brachiation are gibbons, Southeast Asian apes with very long arms and relatively short bodies and legs. Gibbons often have to travel along slender branches to reach fruit. Their method of travel is especially effective for traversing a slender branch, because walking on top of a thin branch is a feat of balance like tightrope walking, but swinging below a branch presents no such problems.

Children make swings go higher by "pumping:—bending and extending their legs at appropriate stages of the swing. Gibbons make use of the same technique, but instead of using it to go higher, they speed up their swinging through the trees. As they swing under a branch they bend their legs, raising their potential energy. At the top of the swing, as they reach out to catch the next branch, they extend their legs again. If at this stage they are flying through the air between branches, the extending of the legs lowers the feet and raises the trunk a little but leaves the path of the center of gravity unaltered. And if the center of gravity's path is unchanged, the body's energy is unchanges as well. Bending the legs increases the body's energy and extending them leaves it

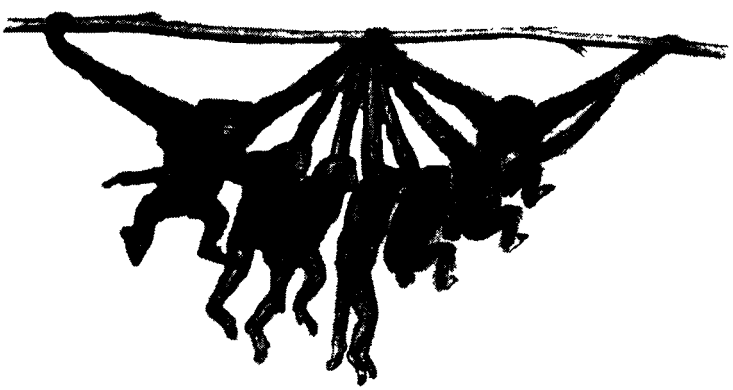


Figure 8.7 A branching siamang (*Hylobates syndactylus*, a species of gibbon) builds up speed like a child "pumping" a swing.

unchanged, so the effect of the whole cycle of pumping movements is to increase the energy, making the animal travel faster.

A child pumping a swing stand up as it passes through the vertical and bends the knees at the top of the swing. This stand-and-bend sequence seems the reverse of the gibbon's action if you look only at the legs: the child is straightening its legs when the rope is vertical, but the gibbon is bending its legs when its arm is vertical. However, the effect on the center of mass is the same in both cases: it is being raised at the stage when rope or arm is vertical.

8.2.2 Gripping a Smooth Surface

Gibbons spend their lives in the trees, seldom or never descending. In contrast, squirrels and many monkeys often come down to the ground, where they find much of their food. To get back up a tree, these animals need to climb a vertical trunk. In accomplishing their climb, not only must they be correctly balanced to counteract the force of gravity, but they need a way to hang on to the surface. Indeed, for some animals, the difficulty of maintaining a good grip can make getting down trickier than getting up. A adroit as climbing mammals are at maneuvring up and down trees, however, the true masters are some insects and lizards that are able to hold on to smooth surfaces, even when upside down.

To climb a vertical trunk, an animal must pull with its fore limbs and push with the hind. The force exerted by the limbs must balance the animal's weight, so you might think that only vertical forces would be needed. But if the only forces on the animal were its weight acting downward and upward forces on its paws, it would not be in equilibrium: that combination of forces would make it rotate head over heels, for equilibrium, the forces on the feet must slope. The force on the fore feet may be more vertical and force on the hind more horizontal, or vice versa. The (downward) weight is then balanced by the (upward) components of the forces on the paws; the (forward) component of force on the fore paws is balanced by the (backward) component on the hind paws; and forces that would rotate the animal clockwise are balanced by others that would rotate it counter clockwise.

Squirrels have claws that dig into bark and help them climb. The hind claws are pushed into the bark, so the animal does not have to rely on friction to prevent it from sliding down the tree. The fore claws are hooked into the bark, enabling it to pull on

them. Claws on the hind feet are not essential for their kind of climbing unless the surface is slippery, but claws on the fore feet are essential. The squirrel could not pull on its fore feet if its claws were not dug in.

The opposite is true for coming down a tree trunk headfirst: the hind claws must be hooked in. Not only does the animal need hind claws, but it also must be able to turn its feet back to front so that the claws can pull in the required direction. Squirrels can do this, and so can kinkajous and various other mammals that live in trees. Domestic cats cannot reverse their hind feet, though the South American tree-living cats called margays can. A domestic cat that climbs up a tree easily may have great difficulty coming down.

Reversing the hind foot as squirrels and kinkajous can do requires a very mobile ankle. To understand what is required, sit on the floor with the soles of your feet flat on the floor in front of you. You will be able to turn your feet inward so that the soles of the feet are vertical, each facing the other. To do as squirrels and kinkajous do, you would have to be able to turn them through a further 90 degrees until the soles were horizontal, facing upward.

Coming down tree trunks is relatively easy for birds because their gripping feet have one or more backward-pointing toes, so there are claws curving in the right direction to hook into the bark whether the bird is running up or down the tree. Creepers (known as tree creepers in Britain) work their way up tree trunks, searching for insects in the crevices in the bark, then fly to get down again, but nuthatches run down tree trunks as well as up. Creepers and woodpeckers have stiff tail feathers, which they press against the tree trunk, using them to supply the force that climbing squirrels get from their hind feet.

Only one group of small monkeys, the marmosets, have claws, and even they lack a claw on the big toe. Other monkeys have finger nails and toenails, like humans, and rely on friction for their grip. To be able to pull with their arms (which they must do to climb vertical trunks), they must be able to reach well around toward the back of the tree. They need long arms to be able to climb even moderately large tree trunks, and they cannot climb very thick ones.

The accomplishments of monkeys pale beside those of creatures that seem able to climb the smoothest surfaces and even defy gravity by walking upside down. Flies and other insects can walk up vertical walls and even windowpanes that give no chance for claws to hook

in. They can also walk upside down on the ceiling. Small lizards called geckoes are often seen hanging upside down on the ceilings of houses in the tropics, and tree frogs can cling to vertical surfaces by means of the soft pads on their feet. All these animals depend on adhesive feet.

Aphids (greenfly and similar bugs) suck out the juices from plants through mouthparts that resemble a hypodermic needle. They have to walk on smooth plant surfaces, including vertical stems and the undersides of leaves, and they have to adhere strongly enough not to be blown off or shaken off as the leaf waves in the wind. The force of adhesion has been measured by putting an aphid on a clean piece of glass on the pan of a sensitive scientific balance. The balance is adjusted to read zero with the glass and aphid in place, then a thread that has been dipped in glue is dangled over the aphid so as to stick to it. The investigator pulls gently upward on the thread, making the balance register a negative load. The balance reading at the instant when the aphid detaches is about –6 milligrams (thousandths of a gram) for aphids weighing 0.3 milligram. These aphids can hold on to glass with a force 20 times their own weight.

That experiment measured the force at right angles to the glass needed to detach the aphid, but aphids can resist forces at least as large when they act parallel to the glass surface, as a gust of wind might. This has been demonstrated by letting aphids crawl up the sides of glass centrifuge tubes and then spinning them at various speeds and observing whether they were thrown to the bottom of the tube.

Aphids have tiny claws on their feet, but you would not expect these to be any use for attaching to smooth glass surfaces. Thus it is not surprising that aphids can still adhere after the claws have been cut off with a fine scalpel. The attachment organ is a spongy pad on the foot. Various experiments have been performed to find out how it works.

One possibility is that the pads attach by suction. If there were muscles that could lift the center of the pad, the pressure under the pad would be reduced and it would hold on like a sucker. However, no likely muscle has been found, and the force of adhesion is affected very little by anesthetizing the aphid with carbon dioxide, which makes muscles relax. Furthermore, and this is the conclusive evidence, aphids remain firmly attached in a vaccum. An important clue comes from the observation that aphids lose the ability to adhere

to smooth surfaces after walking for a while on silica gel, which absorbs water strongly. Aphids that had walked on silica gel for 15 minutes took about 30 minutes to recover their power of adhesion—unless they were placed on water-soaked filter paper. Allowed to walk on this moist surface, they recovered within 2 minutes. Adhesion seems to depend on water. There is a thin film of water between the foot and the surface to which it is attached. Pulling the foot away from the surface, widening the gap, would draw the water surface in around the edges of the foot. But because this inward pull is resisted by surface tension, a large negative pressure can develop in the water. The negative pressure holds the aphid in place.

The feet of beetles and some other wall-climbing insects are very different from those of aphids. Instead of having spongy pads, their feet are covered by a dense pole of very fine, flexible setae (bristles) with wider ends. Such setae can make exceedingly close contact with and solid surface, whether it be smooth or rough, and it has been suggested that they attach by van der Walls forces, which are forces of attraction between molecules. These forces are very weak unless the adhering surfaces are very close together; the force is inversely proportional to the cube of the distance, so if you move twice as close you get eight times the force. For insect adhesion to work this way, adhering surfaces would have to be no more than about 10 nanometers, or about the diameter of a large protein molecule, apart. (A nanometer is one millionth of a millimeter.) There seems to be no direct evidence that van der Walls forces are involved, but all the other obvious possibilities have been fairly convincingly eliminated: no liquid or adhesive has been observed, beetles remain attached in a vaccum, and they are not detached by an antistatic gun, which would release them if they depended on electrostatic forces.

Geckoes have setae on their feet like those of beetles, only finer. Even the ends, where they widen, are only about 200 nanometers across, and at this width the setae are too narrow to be visible by light microscopy. Electron microscopy is convenient for looking at the setae on beetle feet and is essential for observing the setae of geckoes.

8.2.3 Traveling Waves

Smooth gliding of snakes, without legs, may seem mysterious until we realize what is happening. Seeing a snake for the first time, you might almost imagine that tiny wheel must be hidden

under its belly, but the actual mechanisms need neither legs nor wheels. In the most usual method of crawling used by snakes, the body forms an S, or a more complicated wavy shape, and then slides foreward along the wavy path of its own curves. The head slides forward forming new waves, and the tail slides forward losing waves from the rear. It is as if a wavy line were drawn on the ground and then the animal slid forward along it, every part of the body following all the curves of the line.

The mechanism is simple. The snake forms its body into waves, winding between stones, tussocks of grass, or even slight irregularities on the ground. It then makes the waves travel backward along its body, from head to tail. The body can slide past stones or other obstacles more easily than it can rise over them, so the waves remain stationary on the ground and the snake moves forward. This method of crawling works well on rough ground but not on smooth surfaces that have no obstacles for the snake to push against. A snake on a slippery floor gets nowhere.

This same mechanism is also effective for burrowing through loose sand: less force is needed to push the snout forward into the sand than to push all the waves of the body backward through the sand. The movements are like those of a swimming eel, and this method of burrowing is sometimes described as swimming through sand.

Only a few snakes burrow, but it seems likely that snakes evolved form burrowing lizards. Many lizards live in sandy deserts, where they are in a danger of overheating in the sun. They can escape this danger by burrowing in the heat of the day and coming to the surface only when it is cooler. Even quite shallow burrowing is enough to make a big difference in temperature. For example, 2.5 centimeters below the surface of bare soil in a hot part of Australia, the temperature fluctuates at midsummer between a daily minimum of about 25°C and a maximum of 53°C, but at a depth of 30 centimeters in the same soil the fluctuation is only between 35 and 38°C.

A lizard with normal-sized legs could fold them against its sides and swim through sand, but they would be rather in the way. Burrowing is easier without such appendages, and probably for this reason many desert lizards have reduced legs or no legs at all. The legless ones are easily mistaken for snakes, but there is more to being a snake than just having no legs. The easiest way to tell a

lizard from a snake is to look at the scales along the underside. A lizard has irregular scales there, but a snake has a single row of large rectangular scales.

The wavy motion of snakes and legless lizards may remind you of the side-to-side bending movements that legged lizards make as they run, but there is a fundamental difference. The legless animals form traveling waves: each wave crest starts at the head and travels back along the whole length of the animal. The legged animals, however, form standing waves: crests form at one or two particular points on the body and do not move. Standing waves are fine for running but would be ineffective for burrowing.

The travelling-wave style of crawling that has been described so far is the most usual gait of snakes and is known appropriately as serpentine crawling. It works well on firm, rough ground, but less well on loose sand, which offers no fixed points for pushing on, and this method would be no use at all for climbing vertical tree trunks. For travel on these surfaces, two other snake gaits have evolved, both also formed by traveling waves. Sidewinding is the gait that rattlesnakes use to travel over loose sand. Bends travel backward along the body as in serpentine crawling, but there is no sliding of the belly over the ground. Instead, each part of the body is stationary while on the ground but forms new curves in the air as it is lifted periodically to a new position. Marks left in the sand behind the snake show where the body has lain.

A quite different technique (again using bends) enables snakes to climb up grooves in the trunks of trees or fissures in rock. Parts of the body are folded up like the pleats of an accordion to wedge them tightly in the groove or fissure, an arrangement that has inspired the name "concertina locomotion" for this style of travel. At the front of a wedged region the folds are opening out, pushing the snakes's head forward. At the back of a wedged region new folds are forming, drawing the tail forward. Thus groups of folds seem to travel backward along the length of the snake.

More energy is needed to drag a box along a road than to pull a wheeled cart of the same weight, because the sliding of the box is resisted by friction. Using similar reasoning, you might assume that the sliding motion of a snake would require more energy than walking at the same speed, but you would be wrong. Michael Walton Jayne, and Al Bennett, have measured the rates of oxygen consumption of black racer snakes crawling on a moving belt. When the snakes

traveled by sepenatine crawling, the team of scientists found that the energy cost per unit distance was about the same for the snakes as for lizards and mammals of equal mass (100 grams). The snakes were slow (they could manage short bursts at 1.4 meters per second, but the highest speed they could sustain was only one tenth of that), but they were not uneconomical. Concertina locomotion was much slower and also used more energy per unit distance.

8.2.4 Stretching and Squeezing

All the movements that have discusses so far have been powered by muscles that pull on jointed skeletons. Our legs and those of other mammals and of insects have skeletons of stiff rods or tubes, jointed together. Even backbones (including those of snakes) consist of rigid vertebrae connected by movable joints. Without our skeletons we would be flabby and ineffective, but worms, and molluses such as slugs, move very effectively without any such stiffening.

Earthworm burrowing works on the same principle as the concertina locomotion of snakes: some parts of the body are jammed tightly in the borrow while others move forward. The earthworm, however, is jammed into the available space by making the body swell, not by throwing it into folds. An earthworm's body consists of a line of about 150 segments, which appear as rings on the outside of the body. Inside, the fluid filled body cavity is divided into more or less water tight compartments, one for each segment. The body wall includes two layers of muscle, one of longitudinal fibers running lengthwise along the body and one of circular fibers running circumferentially. When the longitudinal muscle of a segment

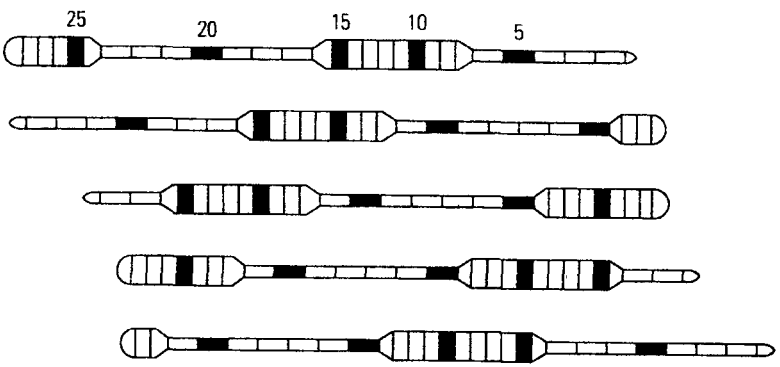


Figure 8.8 Earthworms crawl by passing waves of muscular contraction backward along the body.

contracts, that segment gets shorter, but because its fluid contents cannot escape it also gets fatter. When the circular muscle contracts, the segment gets thinner but also longer. The segment becomes short and fat or long and thin as the two sets of muscle contract in turn, but its volume remains constant.

In the diagram, segments 7 to 16 and 24 to 29 are short and fat, jammed in the burrow. Segment 7 is about to lengthen and push segments 1 to 6 forward, driving the worm's head onward through the soil. Segment 17 is about to shorten and jam itself tightly in the burrow, and as it shortens it will pull the segments behind it forward. Thus the segments that are fat at any particular instant are stationary and those that are thin are moving forward. Each segment moves forward intermittently.

For this method of burrowing to work as described, the segments at the front of a fat region must always be getting thinner, becoming part of the thin region in front, and the segments at the front of a thin region must be getting fatter, joining the fat region on front. Thus waves of thickening must travel backward along the body.

Earthworms spend most of their time underground, but movements that serve for burrowing also enable them to crawl on the surface. The reason is partly that the thick parts of the body rest on the ground and are held in place by friction, while the thin parts are raised or rest more lightly on the ground. In addition, the worm is prevented from slipping backward by bristles that protrude slightly from the underside of the body, tilted at an angle that lets the worm slide more easily forward than backward. You can feel these bristles if you stoke the underside of a worm with your fingertip. The skin feels relatively smooth as your finger moves backward, but rougher as it moves forward, catching on the bristles. These bristles work like the scales on cross-country skis, which allow the skis to slide forward freely but prevent backward sliding as you climb a slope. The same principle is applied by ratchets that enable machinery to rotate in one direction but not the other.

The crawling of slugs and snails looks even more mysterious than that of snakes and worms, for the animals glide forward without obvious movement of the body. However, if you allow them to crawl on a sheet of glass and you watch them from below, you will see something happening on the undersurface of their bodies, or the sole of the foot. A pattern of darker and lighter bands moves along the sole as the animal crawls. Other gastropod molluscs such

as limpets, whelks, and periwinkles also crawl on their feet and show similar patterns of bands, with some variations. In limpets the bands are split into left and right halves: bands having a dark side on the left and a light side on the right alternate with bands having the sides reversed. As the animal crawls forward, these bands travel backward along the foot. In slugs each band is light or dark across the whole width of the foot and moves forward. The bands move over the sole of the foot at a rate considerably faster than the molluse travels. For example, the bands on the foot of a slug crawling at a typical speed of 2 millimeters per second move forward over the foot at 7 millimeters per second. Other molluscs show slightly different patterns of bands, but whether the bands move backward or forward, the effect is always to move the animal forward.

The moving bands suggest the possibility that limpets may crawl essentially as earthworms do. The bands would be waves of muscular contraction formed by alternate bands shortening and elongating. The shortening bands might swell, lifting the elongated ones off the ground. The lengthening of a band would push on the elongated band ahead of it and, unattached to the ground, the elongated band would move forward easily. A shortening band would pull the elongated band behind it forward. The same reasoning tells us that if the contractions travel backward along the foot, the limpet will move forward.

A simple experiment seemed to confirm that limpets do move by muscle contractions in the same manner as earthworms. Marks were made on the feet of limpets, and the limpets were then filmed from below as they crawled on glass. Each mark remained stationary while its part of the foot was short and presumably resting on the ground, and moved forward while its part of the foot was elongated. Similarly, in crawling earthworms the contracted (fat) segments are anchored and the elongated (thin) ones move forward.

The theory seemed good for limpets, but how about for slugs, which crawl by waves that move *forward* along the foot? Films of slugs and snails with marks on their feet showed that the marks remained stationary while their part of the foot was elongated and moved forward while it was contracted—just the opposite of what happens in limpets.

Perhaps some of the puzzle could be resolved by taking a closer look at the sole of the foot through a microscope. To catch the sole in midmotion, the slug would have to be frozen in the act of

crawling. A good specimen was obtained by picking up a crawling slug and dropping it into a container of liquid nitrogen, where it froze immediately. It was allowed to thaw in a solution of a chemical fixative that coagulated its proteins, preserving the shape of the soft tissues. Part of the slug was then sliced into thin sections, and the sections were examined under a microscope. They showed that the surface of the foot had ridges running across it, as suggested in the theory of limpet crawling. However, the tissue of the projecting rides that presumably rested on the ground was elongated and that of the furrows contracted; again, this pattern is the opposite of the supposed pattern in limpets.

A little thought shows that the crawling of slugs could be explained by a mechanism that is only a little different from that of limpets. Suppose that the extended bands of the foot are anchored by resting on the ground. Suppose also that the rear edge of each anchored band is shortening and that the rear edge of each raised part is lengthening. The raised parts will be moved forward, but the waves will also travel forward because it is at its rear edge that each band is turning into the other kind.

This explanation of molluse crawling seemed to be further confirmed by the arrangement of muscle fibers within the foot. There are no distinct layers of fibers running in different directions, as in earthworms. Instead, the muscular foot consists of interwoven fibers running in various directions. In limpets, some of the fibers run vertically through the thickness of the foot and others transversely across its width. If the vertical fibers contract in part of the foot, squeezing it thinner, that part must grow longer or wider or both in order to retain constant volume. If at the same time the transverse fibers prevent that part of the foot from widening, it has no other alternative but to become longer. Thus we can expect contraction of the vertical fibers to lengthen part of the foot, and because the fibers pull upward they will lift it off the ground. The muscle fibers of slugs have a different arrangement; instead of running vertically and horizontally, they slope, some pulling obliquely up and forward and others up and back. When these muscles contract, they *shorten* their part of the foot and lift it off the ground. Each molluse thus has the arrangement of muscles that seems to be needed for its style of crawling.

The observations of the animal's movements, of the waviness of the feet when they were quickly frozen, and of the arrangement

of muscles seemed to add up to a clear, coherent story that told how molluses crawl. The explanation seemed complete until the work of an American graduate student forced some drastic rethinking.

While still an undergraduate at Duke University, Mark Denny had made his name as a research worker through a beautiful study of the properties of spider silk. He went on to work for his doctorate in Vancouver, and his work on slugs there, published in the early 1980s, gave us a new understanding of molluse crawling.

Denny saw two problems with the theory. First, it failed to explain the trail of slime (muscus) on which slugs and snails crawl. That slime consists of water and dissolved salts, with a small but significant proportion (3 to 4 percent) of glycoprotein, a compound of protein with sugar molecules. (The mucus that runs from your nose when you have a bad cold has a similar composition.) The mucus presumably did something useful, since the animal would save water and protein by not producing it, but the theory did not explain its function.

The second problem that Denny saw with the theory was that huge forces would be needed to lift the foot from the surface when ever a muscular wave passed. If you lay damp sheets of glass on top of each other, they are very hard to separate (except by sliding) because a thin layer of water forms between the sheets that glues them together very effectively. If the sheets were smeared with mucus instead of water, separating them would be even more difficult, because mucus is viscous like molasses. The flexible feet of molluses fit very closely onto rocks and other hard surfaces, and the short distance between the two surfaces ensures that the molluses are firmly glued down. Some limpets of 3 centimeters diameter can hold on to rocks against forces of more than 200 newtons (45 pound's force).

Denny performed a very simple experiment to check whether the foot really is lifted as the muscular waves move along it. He persuaded a slug to crawl on metal foil and then froze it suddenly with out detaching the foil. Only after the specimen had been chemically fixed and thawed was the foil peeled off. When the sole of the foot was sectioned and examined microscopically after this treatment, it was found to be flat. There were lengthened and shortened bands running across it, but they were not raised into ridges. Denny concluded that there are no ridges on the feet of crawling molluscs and that the ridges seen previously had formed

only after the animal was pulled off the surface on which it had been crawling.

If the whole foot lies flat on the ground, why do some parts move forward while others remain anchored? There is no obvious ratchet like the bristles of earthworms. Denny looked for an answer by investigating the mechanical properties of mucus.

A layer of rubber is sandwiched between two steel plates and glued firmly to both. After fixing the lower plate rigidly to a table top, you push the upper one horizontally, distorting the rubber layer. The farther you push it, the bigger the force you feel on your finger. If you hold the plate in position the force remains, and if you release it the rubber springs back to its original shape, returning the steel plate to its original position. This is elastic behaviour.

Now imagine the same experiment with a layer of molasses (treacle, to British readers) between the plates instead of rubber. A small force will move the upper plate slowly and a large force will move it faster, but if it is kept moving at constant speed the force remains constant. If you stop moving the plate, in any position, the force disappears, and the plate does not spring back to its original position when released. This is viscour behaviour. The force does not depend on how far the plate has moved, but on how fast the plate is moving.

Denny collected a supply of slug mucus and performed essentially the same experiment on it. He put the mucus between two metal plates, one flat and one slightly conical, that formal part of an instrument known as a cone and plate viscometer. The flat plate was rotated, and the torque needed to prevent the cone from rotating was measured. In rotating the plate, Denny wanted to simulate the conditions under the foot of a crawling slug, where each point on the foot moves at one stage of each muscular wave and remains stationary at another. Therefore, he rotated the plate at constant speed for one second, then held it stationary for a second, then rotated it for an other second, and so on.

If the experiment had been performed with an elastic solid such as rubber instead of mucus, the torque would have increased when ever the cone was rotating and would have remained constant while it was stationary. If the experiment had been performed on a viscous liquid such as molasses, there would have been a constant torque whenever the cone was turning, and that torque would have fallen to zero whenever the cone stopped. The actual result resembled

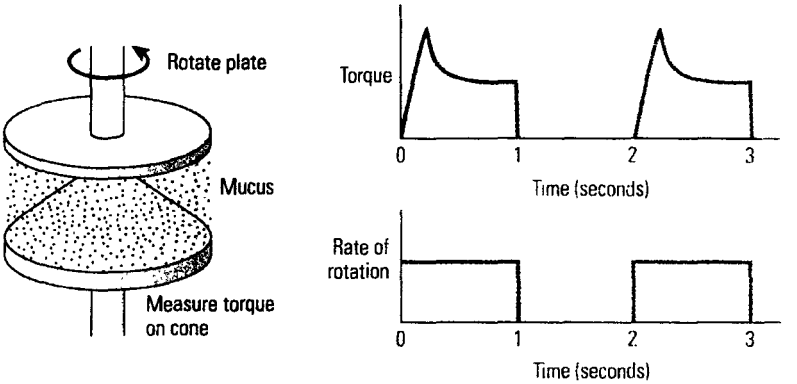


Figure 8.9 Denny studied the properties of mucus by testing it in a viscometer.

neither of these possibilities. In the early stages of each rotation, the torque rose progressively as if the mucus were an elastic solid. After a while, however, it fell a bit and then remained constant for as long as rotation continued, as if the mucus were a viscous liquid. The period of rest between rotations was apparently enough for the mucus to revert to its original elastic state.

Denny suggested that the state of the mucus under a slug's foot changes with the passage of each muscular wave. Under the extended parts of the foot, it behaves like an elastic solid, but under the contracted, moving parts it behaves like a viscous liquid. This switch in properties might be expected to happen because the total area of extended parts, at any instant, is greater than the area of the contracted parts. The force needed to move the contracted parts forward must be balanced by the anchoring force on the extended parts (otherwise the extended parts would slide back). The same force must act on both, and so if the contracted parts have less area, more stress (force per unit area) will act on them. The stress under the contracted parts will build up to the level needed to liquefy the mucus, while the mucus under the extended parts remains solid. As the waves travel along the foot, there will always be solid mucus under the extended parts, anchoring them, and liquid mucus under the contracted parts, allowing them to slide forward. The changing properties of the mucus have the same effect as would a ratchet under the foot.

The same mechanism can also work for the backward-moving waves of limpets, if the total area of the parts of the foot that are extended at any instant is less than the area of the parts that are

contracted. In that case, the contracted parts will be anchored and the animal will move in the direction opposite to the waves.

The mollusc's manner of crawling seems to have severe disadvantages. A snail's pace is proverbially slow. 2.5 millimeters per second (one tenth of an inch per second) or less, a hundred times slower than some bettles. Snails are inevitably slow, because if they were moving fast very large forces would be needed to overcome the viscosity of the mucus under the moving parts of the foot. A second disadvantage is that the method of crawling is expensive of energy and materials. Measurements of oxygen consumption show that the metabolic energy cost per unit distance is about 12 times higher for a slug crawling than for a mouse of the same mass running. For every 10 meters that it travels, a 15-gram slug loses about 1 gram of water in its mucus and 30 milligrams (0.03 gram) of glycoprotein.

Costly though it may be, slug and snail crawling has one vary clear advantage. The animal is quite firmly struck to the surface it is crawling on, at all times. Limpets are not easily dislodged by waves, and a snail can climb up a plant stem without much danger of falling off.

9

Gliding Activities

Flying squirrels, which lack the refined as design of birds, use gliding as a fast and saving means of travel through the forest *Glaucomys volans*. Due to flying adaptations, animals become beautifully modified, due to which, they are offered the least possible resistance by air to the attainment of speed. Volant adaptations among the invertebrates occur in insects. Among the vertebrates, the volant adaptations are found in fishes, amphibians, reptiles, birds and mammals.

Volant adaptations are concerned with the flight. This flight may be : (i) *passive* or *gliding type* characterized by leaping (jumping) from a high point and held up by certain sustaining organs, then glides to lower level. There is no locomotive force other than the gravity (ii) *True flight* is the aerial flight caused by the action of wings. It is found in insects, pterodactyles, birds and bats. In them the nature of development and structure of wings are quite different which suggest that the flight has evolved independently in different

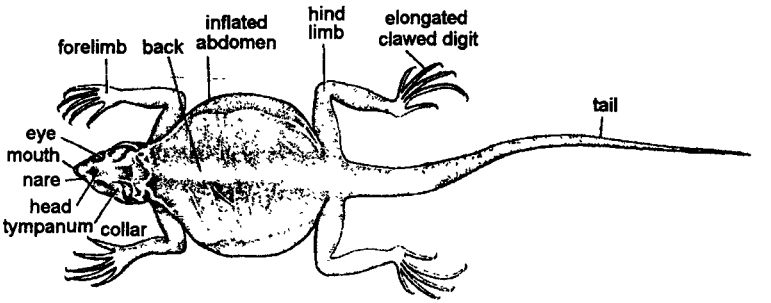


Figure 9.1 *Crotaphytus.*

groups. In such flights the power is implied and the movement in air is sustained.

9.1 TYPES OF FLIGHT

Flight is of two sorts :

9.1.1 Passive or Gliding Flight

In this kind of flight, the animals only take an initial leap from a high point and are held up by certain sustaining-organs and impelled by gravity, they glide to a lower level in a horizontal plane. Sometimes, the distance covered by gliding may be many yards. This kind of flight can be compared to a gliding aeroplane without engine-power. The gliding flights are performed by various lizards (flying dragon—*Draco volans*); fishes (*e.g. Exocoetus* etc.); birds (ostriches etc.); mammals (lemurs) and amphibians (*Rhacophorus*). The following are noticable adaptation met with in such forms:

9.1.1.1 Development of patagia

The sustaining surface for the gliding is a fold or series of folds of the skin known as *patagium*. This may be supported by ribs as in *Draco* (flying dragon). It lies between forelimbs and hindlimbs. This patagium can be folded like a fan against the sides of the body when not in use. Another example among reptiles is *Ptychozoon*. "The flying or fringed gecko", in which lateral expansion of skin (patagium) extends along the side of neck, body, tail and limbs and between toes. Flying snakes (*Chrysopelea*) also leap by the concave ventral side.

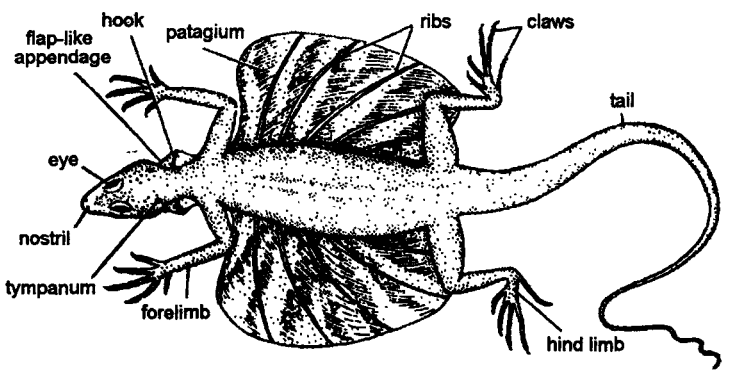


Figure 9.2 *Draco maculatus.*

The *pterodactyls* of Mesozoie era were volant creatures akin to birds. They possessed true flight patagia. Patagia were extensions

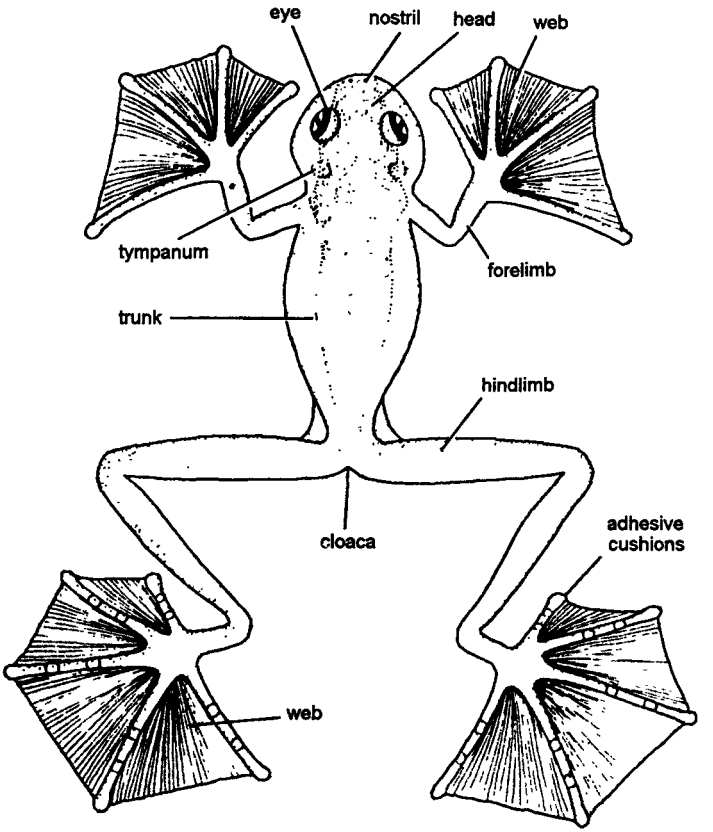


Figure 9.3 *Rhacophorus.*

between limbs supported by ribs. In flying lemur (*Galeopithecus volans*), which is a soaring mammal, patagium extends from side of the neck to the tip of tail even including digits, which are webbed as for aquatic life. In the bats, the patagium is supported mainly by the elongated forelimbs and the second, third, fourth and fifth digits. The first digit is free. Traces of patagia are also found in front of and behind arm is birds which have adequate supporting function before the feathers usurped their place.

9.1.1.2 Enlargement and high insertion of pectoral fins

Among fishes (*Exocoetus* etc.) pectoral fins become enlarged in the form of parachutes and are highly inserted on the body. The lower lobe of tail is also invariably longer, helping in leaping. The pectoral fins do vibrate. It flies upto 200-300 meters. Other genera

are *Dectylopterus, Pantodon* (African flying fish) and *Pegasus*, a little fish found along the coasts of Japan, China, India and Australia etc. They skim along the surface of the water for 40 feet or more. Fossil flying fish are *Chirothrix, Gigantopterus.*

9.1.1.3 Webbing of feet

In flying frog (*Rhacophorus paradailis*), the feet are webbed which sustain the prolonged leaps to which it is addicted. The digits terminate in adhesive pads which help in adhesion to trees. There are also rudiments of patagia in front and behind the arms.

9.1.2 Active or True Flight

In active flight, there is a sustained movement through the air. It covers long distances and involves locomotive force. True flight has evolved thrice among vertebrates: in the reptilian Pterodactyles, the birds and the bats. Whether flying fishes should be included is a much controversial question. However, the animals having true flight, move their wings with varying degrees of rapiduty or after having gained their high altitude by flapping their wings, they may also soar or sail on motionless wings for several hours continuously.

9.2 MODIFICATIONS

Body contour in volant animals has been emphasized and is second only to that of the purely aquatic forms in its degree of perfection for the lessening of resistance.

9.2.1 Sustaining Surface

The sustaining surface is primitively, except in the fishes, a fold or series of folds of the skin known as the *patagium* (Lat. *patagium*, an edge or border). This may be supported in various ways, but with one exception, in the little lizards (*Draco spp.*) which inhabit the Indo-Malayan region, the limbs form the chief supporting agents. In the "flying dragons," *Draco*, just mentioned, the body is depressed and the sides extend outward into a pair of large, winglike membrances, supported by five or six elongated ribs. The entire device can be folded like a fan against the sides of the body when not in use. The soaring powers are not very great but when resting among these luxuriant foliage of their habitat the animals are said to resembles butterflies in their habit of opening and closing the wings.

Most soaring mammals have the patagium supported between the fore and hind limbs and sometimes the skin-fold extends in

front of the fore limb to the neck and again between the hind limbs and the tail. Perhaps the extreme of development may be seen in *Galeopithecus*, the so-called "flying lemur," for here the patagium extends from the sides of the neck to the tip of the tail, even including the digits, which are webbed as though for aquatic life.

Where the patagium is supported mainly by the elongated fore limbs and their digits, true flight ensues as in the pterosaurs, wherein the enormously elongated outer finger sustains over half the membrane, and in the bats, whose second, third, fourth, and fifth digits perform a like function, the first along being free. In both groups the membrane extends from the arm to the sides of the body and also to the front of the hind limb. An interfemoral membrane, which, however, may have existed, has not been demonstrated in the pterosaurs but is variable present in bats.

9.2.2 Gliding Vertebrates

9.2.2.1 Fishes

There are enumerated several genera of flying fishes, each of which represents a separate volant adaptation. Of these the first to be mentioned are the several species of the genus. *Exocaetus*, allied to the skippers and garfish, which live in all tropical and subtropical seas where the fly in shoals in their efforts to escape the relentless tunny and albicore. These flying fishes are trim-built creatures with large pectoral fins, which are the main organs of flight, and variably developed but much smaller pelvics. The lower lobe of the tail is invariably the longer and aids in giving the final impetus to the fish as it leaves the water and also in accelerating its speed if in the course of its flight it comes near enough to the surface of the sea. The length of flight is said to vary up to 200 or 300 yards and it is sometimes sufficiently high to strand the creature on the deck of an

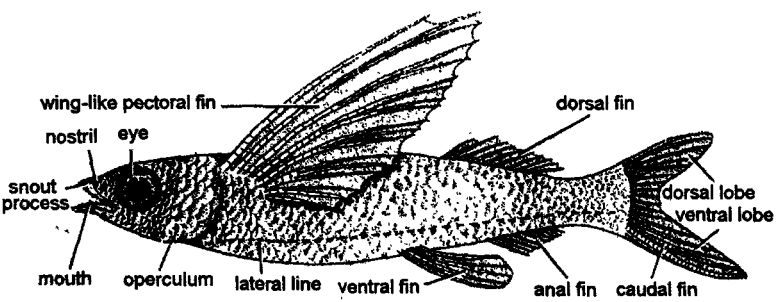


Figure 9.4 *Exocoetus.*

ocean-going craft. Whether the flight of *Exocaetus* is true flight or merely a soar is a much disputed question and the evidence, briefly summarized, is as follows:

The wing expanse is hardly sufficient for such extended soaring. The wings (pectoral fins) do vibrate, but whether due to muscular effect or to friction, as a flag is flapped in the wind, is not clear. The muscular development seems insufficient for true flight, but on the other hand it is more highly developed than in allied non-flying fishes. It may well be that while true flight as such does not exist among fishes, rapid wing vibration insufficient in itself to support or drive the animal may aid in maintaining or prolonging a soaring flight of which the main propulsive effort is acquired by the tail before having the water. At all events, their flight is remarkable and the creatures are one of the most interesting features of the storied tropical seas.

The various species of *Dactylopterus* are known as the flying gurnets and while the flying is by no means as sustained as in *Exocaetus,* of the former fishes Moseley writes: "I have distinctly seen species of flying gurnets move their wings rapidly during their flight.....especially in the case of a small species of *Dactyolpterus* with beautifully colored wings, which inhabits the Sargasso Sea." Moseley likens the flight of the gurnets to that of grasshoppers.

There is an African flying fish found in the Congo and Niger rivers—*Pantodon,* a form but three or four inches long—which leaps out of water and flutters through the air for some distance. Still another flying type is *Gastropelecus,* a small, compressed fish with long and curved but not particularly large pectoral fins, which occurs in the rivers of British Guiana. It skims along the surface of the water for 40 feet or more, beating the water with its pectoral fins. Then it leaves the water for a distance of five to tea feet and when exhausted falls sideways into the water again *Pegasus volitans,* a little fish found along the coasts of Japan, China, India, and Australia, skims a short distance above the surface of the water by means of its broad pectoral fins.

Several extinct forms have been described as "flying fishes." These are: *Dollopterus* from the Middle Trias of Jena, *Thoracopterus* and *Gigantopterus* from the Upper Trias of Austria, *Exocaetoides* and *Chirothrix* of the upper Creataceous of Mt. Lebanon, Syria. The last mentioned is of particular interest on account of the huge size of the pectoral fins, which seem to imply powers of flight fully

equal to those of the living *Exocaetus* and *Dactylopterus*. We have therefore among fishes no fewer than ten separate adaptations to aerial conditions, one or two, possibly three of which approach very near to, if they have not attained, true flight.

9.2.2.2 Amphibia

The only volant adaptation among amphibia is that of the tree-frog, *Rhacophorus* whose webbed feet sustain it in the prolonged leaps to which it is addicted. This genus includes a large number of species in the Oriental realm, especially in Borneo. The digits terminate in adhesive pads, in common with those of other tree-frogs, and are connected by web-like expansions of the skin. There are also rudiments of patagia in front of and behind the arms. In *Rhacophorus pardalis* the total alar expanse is about three square inches, which would imply rather feeble gliding powers.

9.2.2.3 Reptilia

Lizards includes at least two genera and several species of gliding forms, of which the most remarkable is the flying dragon, *Draco,* already referred to in which the patagium is supported by a number of extended ribs. They occur principally in the Malay Peninsula and Archipelago and average some eight to ten inches in length.

Ptychozoon is the flying or fringed gecko of the Malay countries, which is bedecked with lateral expansions of skin along the sides of the neck, body, tail, and limbs, and between the toes. While these may aid in breaking the creature's fall, they may also, coupled with the colour, serve a cryptic function and render the animal less conspicuous against the back of the tree upon which it rests.

Several so-called flying snakes are recorded, such as *Chrysopelea,* the flying snake of Borneo, which descends obliquely through the air, its body right, and the ventral side concave to sustain the creature in its fall.

The pterodactyls or flying dragons of the Mesozoic were a very remarkable group of reptiles whose first recorded appearance is in rocks of the Rhaetic or uppermost Triassic period. They range through the Jurassic and on into the Upper Cretaceous, when they become extinct through racial death. They were undoubtedly skin to the birds, but that simply means, in all probability, derivation from a common, possibly Permian ancestry; nevertheless the two groups show a number of highly comparable, homoplastic characters, some of which have already been referred to. The remarkable thing

is that, like the turtles, they first appear fully developed and characteristic of their order, with no record thus far discovered of their antecedent evolution, and the subsequent changes are mainly increases in size, perfection of the shoulder girdle articulation and loss of tail and of teeth. In size they range from that of a sparrow to the mightiest of nature's airplanes, for the replica of the late Cretaceous *Pteranodon* mounted at Yale measures 13 feet 6 inches in alar expanse and Eaton is authority for the statement that at least one individual, judging from the relative proportions of the bones which have been preserved, had an estimated breadth of 26 feet 9 inches from tip to tip. The pterodactyls possessed true flight, which in those from the Kansas chalk must have been sustained, as their remains are found in association with marine reptiles, fishes, and invertebrates, apparently far from the ancient shore. Three of the principal horizons whence these pterodactyls come, the Lias of Lyme Regis of England, the lithographic limestone of Bavaria, and the Kansas chalk, are all marine in origin, which is also true of the source of the Mesozoic birds. It is highly probable therefore that in each instance we have not as yet knowledge of the great bulk of the group, but only of a few of aberrant habits and adaptation.

9.2.2.4 Birds

The birds in all probability include but a single evolution to aerial life, although certain excellent authorities "believe more or less firmly that possibly birds had not one, but two points or origin, and feet that if we could follow back their lines of descent we should find that the ostriches came from one and the birds of flight from another". If this be true, the ratite or ostrich group as a whole, with a single exception, have degenerated and lost the power of flight, although an examination of the skull and skeleton shows them to have been descended from flying normal birds; whereas in the carinate or flying birds loss of flight, while it has occurred (flightless rail, penguin, dodo, etc.), is relatively extremely rare. Flightless pterosaurs and bats, on the other hand, are inconceivable, as their flight mechanism involves the hind limbs which have, as a consequence, largely lost their terrestrial locomotion function; whereas birds, being double adapted, can lose their flying powers and still progress easily on the ground or in the water, as their legs are not thus involved.

Birds first appear in time in the Upper Jurassic (Solenhofen limestone) long after the initial record of the pterodactyls. These

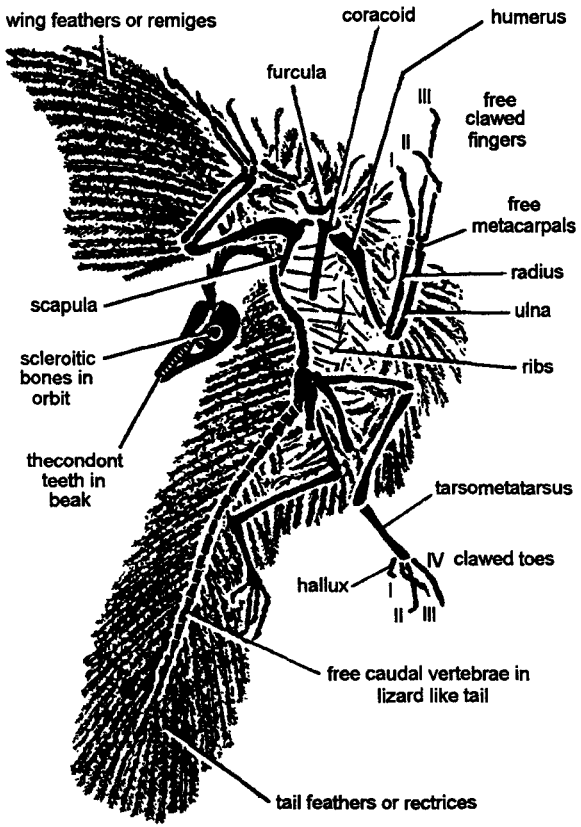


Figure 9.5 Skeleton of *Archaeopteryx.*

first birds, of which but two or three specimens have been recovered, are known as *Archaeopteryx* and *Archaeornis* and are so reptile-like that were it not for the preserved feathers it is doubtful whether they could be surely proved to have been birds. The reptilian traits are teeth, free clawed fingers in the hand, feeble breast-bone, abdominal ribs, etc.

9.2.2.5 Mammals

Among the mammals there are upward of thirteen separate volant adaptations, one of which, that of the bats, attained the power of true flight. Among the flying forms the first to be mentioned are the *marsupials,* of which the flying phalangers include several unrelated species: *Petaurus spp., Petauroides volans,* and *Acrobates pygmaeus.* These are characterized alike by having a well developed

skin fold along the sides of the body between fore and hind limbs, and a feebly developed one in front of the fore leg. In each genus the flying form is especially related to a separate type of non-flying phalanger.

In *Petauroides* the flying membrane extends from wrist to ankle, but is very narrow along the distal segment of each limb. The tail is very bushy except for its prehensile tip, which is naked on the under side. The tail, together with the long fur of the body, must supplement to a considerable extent the buoyancy of the patagium. This genus, with its single species, includes the so-called Taguan flying phalanger found in Australia from Queensland to Victoria.

Petaurus has a much broader patagium and there is a naked prepatagium as well. The tail is very large and bushy, but lacks the naked prehensile tip of the preceding form. There are three species of this genus, ranging over New Guinea and part of Australia.

The genus *Acrobates* includes two small species of flying phalangers which have narrow patagia extending from the elbow to the knee along the flank. The long fringing hairs borne by the patagium, together with those on either side of the tail, aid materially in flight. *Acrobates pygmaeus* is found in New South Wales, Queensland and Victoria, while a second species, *A. pulchellus*, is a native of Papua.

Two milies of *rodents* contain flying forms : the Anomaluridae, including the genus *Anomaluridae*, and the Sciuridae, of which three genera, *Pteromys, Sciuropterus,* and *Eupetaurus,* are volant. *Anomalurus* has a well developed patagium extending from wrist to ankle but narrowing in front of the leg from the knee down. As a compensation, however, there is an interfemoral membrane from the heel to slightly beyond the base of the tail. The genus includes six flying species, all found in Africa.

Ptermys has a highly developed patagium extending as far a the digits. There are also prepatgia and an interfemoral membrane, and the tail is large. The genus is found in the wooded districts of tropical southeastern Asia, Japan, and some of the Malasian Islands, and is said to soar through a distance of nearly 80 yards.

Sciuropterus has no interfeormal membrane, but has a much better developed tail than *Ptermoys* as a compensation. This wide hairy tail is further supplemented by the hairy fringe on the patagium and along the rear of the thighs, and they give collectively a broad supporting area. While members of this genus are smaller than those

of *Pteromys*, their geographical range is much greater, as it includes the northern part of the North American and Eurasian contents, and India.

*Eupetaurus*n is of especial interest in that it is not arboreal but rock-and precipice-climbing. It inhabits the high elevations of northwestern Kashmir.

Among the *Insectivora, Galeopithecus,* the sole representative of the suborder Dermoptera, stands alone in its adaptation, for it exhibits the highest degree of aviation of any of the Mammalia except the bats, and while not of course ancestral to the latter, it is evidently derived from a common stock and gives a very clear idea of the manner in which the evolution of the bats was accomplished. In this genus the patagium reaches its highest development, as it extends from the rear of the head along the front of the arm (prepatagium), between the fingers to the base of the claws, between the fore and hind limb, webbing the toes as well as the fingers, and between the hind limbs and the tail (interfemoral membrane), including the entire length of the latter organs as in the insectivorous bats (Microcheiroptera). The musculature and innervation of the patagium resemble those of the bats and differ decidedly from those of all other volant mammals. The hand is much larger than the foot, but the finger show no trace of elongation. If they did the entire creature would be still more batlike. As it is, the brain is midway in its development between that of a typical insectivore and that of a bat, and the alimentary canal is also, batlike except for an elongated colon or large intestine, which in the bats and birds is very short.

Galeopithecus is nocturnal, as are most volant mammals, resting suspended, head down, from a branch. Its soaring powers are very great, for Wallace tells us of a record of 70 yards with a descent of not more than 35 or 40 feet, or less than one in five. The genus includes two species : *Galeopithecus volans,* from the Peninsula, Sumatra, and Borneo, and *G. philippinensis,* which inhabits the Philippine Islands.

9.3 GLIDING SKILL OF BIRDS

Much of our understanding of the gliding flight of birds comes from the work of Colin Pennycuick, an English zooligist who is now a professor in Miami. He set out to discover how well birds can glide by comparing their gliding abilities with those of human-made gliders and model aircraft. He started by observing seabirds from a cliff top but soon decided to attempt more accurate

measurements in his laboratory, using domestic pigeons. There was not much flying space his laboratory, so instead of having his birds glide through still air he decided to attempt the flying equivalent of running on a moving belt : he would have a stationary bird gliding in moving air. Instead of the bird moving forward and downward through still air, the air would blow backward and upward past a stationary bird.

Pennycuick built a wind tunnel with a powerful electric fan that drove a jet of air through a nozzle of 1 meter diameter. The nozzle was tapered and grids arranged within to ensure that the air flowed out very smoothly, at almost exactly the same speed everywhere across the width of the nozzle. Now all that Pennycuick needed to perform his experiment was to train a bird to fly in the jet of air, at just the right speed relative to the air to keep it stationary relative to the laboratory. It may seem hard to imagine how a bird could be trained to do that, but Pennycuick achieved it with the help of a teaspoon bowl soldered to the end of a metal tube. He held the teaspoon in the jet of air where he wanted the bird's has to be and rolled chick peas down the tube so that they landed in the bowl. The only way the bird could get the food was by flying where Pennycuick wanted it to fly.

The wind tunnel was quite large and could blow air very fast, at speeds of up to 20 meters per second (45 miles per hour). Unable to find a suitable room for a tunnel of the size and power, Pennycuick hung it in the stairwell of a Victorian building at the University of Bristol, where he was working at the time. When he first switched it on it blew out a few windows, but the damage was soon repaired and the experiment worked well.

Pennycuick wanted to test birds ads they glided, not as they flapped their wings. If the wind tunnel was tilted up at a sufficiently large angle, which it had been constructed to do easily, and if it was blowing air fast enough, the bird did not need to flap its wings: it could simply glide into the wind (still remaining stationary relative to the laboratory). By varying the speed of the jet and the angle of tilt, Pennycuick was able to find the shallowest angle at which the bird could glide, at each speed. From that he calculated the rate at which it would have lost height if it had been gliding in still air.

The bird could not glide slower than its stalling speed of 8 meters per second (18 miles per hour). At high speeds it lost height rapidly. Like artificial gliders it had a minimum sink speed at which

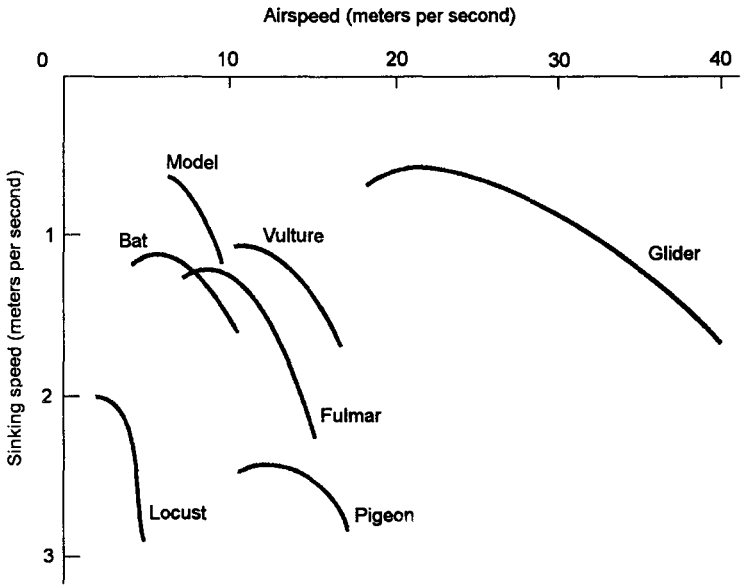


Figure 9.6 The sinking speeds of gliders, birds, and a fruit bat, gliding at various speeds.

it lost height least quickly, but it performance was poor by engineering standards. Good gliders, both full-size ones and models, lose only about 0.5 meter of height per second at the minimum sink speed, but the pigeon apparently could lose no less than 2.5 meter per second. It is not clear whether pigeons really are very poor gliders or there was something about the experiment that prevented the pigeon from gliding well. Since that pioneering study, several similar experiments have been performed on other flying animals. Pennycuick himself moved his wind tunnel to Kenya and repeated the experiment on fruit bats (of about 0.55 meter, or 22 inches, wing span). Vance Tucker of Duke University repeated it with a small species of vulture and with a buzzard. All these, and fulmar petrels that Pennycuick had observed from a cliff top, did much better than the pigeon but slightly less well than artificial gliders : all lost height at minimum rate of around 1 meter per second. Pennycuick has pointed out that comparisons with gliders are really rather unfair: a gliding bird is not a proper glider, but is move like a powered aircraft with its engine switched off.

Pennycuick is a keen pilot, which is probably why he is so interested in bird flight. He was able to turn his piloting abilities to

good advantage during several years he spent in Kenya, where he had a motor glider (an aircraft designed to glide, but with an engine that could be used to get it airbone or to get out of difficulties). Aloft in his aircraft, he glided with the vultures, storks, and pelicans over the East African plains, observing their behaviour in the air. By this means he was able to study the ability of these birds to glide for great distances in weather conditions that are especially favourable for that means of travel.

9.3.1 Thermal Soaring

Although vultures fly for most of the day, they flap their wings very little, keeping themselves airbone by soaring in thermals. These are columns or patches of rising air, formed over ground that has been heated by the sun: the hot ground heats the air, which expands, becomes less dense, and so rises. Thermals are often reasonably easy to find because they form in predictable places (for example, over bare rock). Their tops are often marked by the presence of cumulus clouds (the ones that look like cotton wool). Strong thermals do not appear until the sun is fairly high in the sky, and they disappear in the evening, so thermal soaring is possible only between those times. Vultures spend the night roosting on cliffs or trees.

Vultures and other birds circle upward in thermals to gain the height they need to start a straight glide. They circle by banking, tilting so that one wing tip is higher than the other. The tilt gives the lift a horizontal component that serves as centripetal force. If the birds tried circling in still air, they would lose height—indeed, they would lose height faster than if they were gliding in a straight path, because circling requires more lift (and therefore more induced drag) that does straight gliding. A bird circling in a thermal is able to rise if the air in the thermal is rising faster than the bird would sink in still air. For example, if a bird in a typical thermal is sinking relative to the air at 1 meter per second, and the air in the thermal is rising at 4 meters per second, the bird will gain height at a rate of 3 meters per second. It is common for soaring birds to gain height at such rates.

Vultures travel by gliding from thermal to thermal, sometimes over long distances. For example, Ruppel's griffon vulture nests on cliffs at the edge of the Serengeti and, in the breeding season, commutes daily between its nest and the herds of wildebeest and other mammal from which it takes its food. The vultures soar over the herds, watching for deaths so that they can feed on the careasses.

Pennycuick used his glider to follow vultures on their daily journeys. On one occasion he stayed with a Ruppel's griffon vulture for 96 minutes while it travelled back to its nest from the herds. In that time it covered 75 kilometers (47 miles) entirely by soaring, stopping to circle in only six thermals and rising to heights of up to 1500 meters (5000 feet) above the ground.

Even more impressive are the much longer journeys of white storks. These birds glide from the thermal to thermal for much of their annual migrations between Europe and southern Africa. A straight route would take them over the Mediterranean, where there are no thermals, but they manage to get help from thermals almost all the way by taking a detour. The western European populations travel over the Straits of Gibraltar and the eastern ones over suez.

9.3.2 Slope Soaring

Slope soaring is an alternative to thermal soaring, much used by seabirds. Gulls can often be seen soaring over a hillside or cliff, gliding backward and forward without beating their wings. They are held aloft by the upward air movements formed as the slope of the ground deflects wind upward. If the air is rising as fast as they would sink in still air, they can maintain their height.

The same technique works well over the open ocean, on the windy side of large waves. Albatrosses soar over the Antarctic Ocean, watching for fish or squid swimming near the surface. When they spot one, they often land on the sea to feed. Although they depend largely on slope soaring, they also use another soaring technique that involves swooping up and down between the strong wind high above the sea and the slower wind near its surface.

Pennycuick has made many observations of slope soaring using an instrument that he invented, called the ornithodolite. It is a combination of theodolite and rangefinder that will measure the compass bearing of a flying bird, its angle of elevation above the horizon, and its distance from the observer: with this information the bird's position can be fixed in three-dimensional space. Pennycuick has only to keep the instrument focused on the bird and press a button whenever he wants to record the bird's current position and the time in his portable computer. In that way he gets the information needed to calculate the speeds of soaring birds and their rates of change of height. He also measures the speed and angle of the wind, using an anemometer (wind gauge) attached to a pole as near as possible to where the birds are flying.

Pennycuick usually encountered the largest of albatrosses, the so-called wandering albatross, soaring between 2 and 12 above the waves. The wing span of this species measures more than 3 meters, and these impressive wings hold aloft an adult bird weighing from 8 to 10 kilograms. The bird would typically glide in still air at speeds of about 12 meters per second (27 miles per hour), but when it was slope soaring, gliding into the wind, it gravelled faster than 12 meters per second relative to the air and slower than that speed relative the sea. For example, an albatross gliding into a 10-meter-persecond headwind would typically travel at 16 meter per second relative to the air, gaining around at 16 − 10 = 6 meters per second. In very calm conditions, slope soaring could not keep these birds airborne and they had flap their wings occasionally, but unless it was very calm, they flapped their wings very seldom.

To stay in the upwardly deflected wind over a wave, the bird must soar along the wave—which may not be the direction in which it wants to travel. It can travel in other directions by taking a zigzag course, soaring for a while along a wave and gaining height or speed, then gliding for a while at an angle to the waves before joining another wave and soaring about it. Pennycuick observed albatrosses following a ship in this way and found that the total length of their zigzag paths averaged 1.5 times the straight-line distance that they travelled. On one occasion, for example, his ship was sailing directly into a 6-meter-persecond wind when a wandering albatross overtook it from astern. The bird changed direction every 10 to 15 seconds to follow a zigzag path and achieved a straight-line speed relative to the sea of 6 meters per second, entirely without flapping its wings.

The energy cost of gliding is much less than that of flapping flight, but even when the wings are not being flapped, metabolic energy is needed to maintain tension in their muscles. The main muscle involved is the pectoralis muscle, which pulls the wings down in the downstroke of flapping flight and holds them in position against the lift forces that act on them in gliding. Vultures, storks, and albatrosses have the pectoralis muscle divided into two distinct parts: a large superficial part that is dark red and a much smaller deep one that is paler. It is believed that the superficial part consists of fibers capable of shortening rapidly to beat the wing and the deep part of slowly contracting fibers that can maintain tension at little energy cost during gliding. Only in soaring birds is the muscle

divided into these distinct parts. Albatrosses differ from the others in having what seems to be an even more effective device for saving energy while soaring. A fan-shaped ligament locks the shoulder joint when the wings are horizontal and fully spread. Consequently, the wings cannot be lifted above the horizontal until they are moved slightly back from the fully spread position. If the shoulder lock holds an albatross's wings horizontal the bird may be able to soar with very little tension in its muscles.

Whereas albatrosses travel long distances by slope soaring, small falcons known as kestrels have a different use for the technique. They hang stationary in the air, ready to pounce on any mice or voles that they see moving on the ground below. To hold themselves in place over level ground they fly into the wind, matching their speed exactly to that of the blowing air. Often, however, they slope soar over sloping ground, keeping themselves stationary without having to flap their wings.

John Videler of Groningen University in the Netherlands has made many studies of Kestrel flight. He showed how precisely they stay in position by filming them with a camera on a tripod, locking the camera rigidly to the tripod as soon as it was focused in the bird. His films show that during 20 or 30 seconds the bird's head often moves less than a centimeter from its initial position relative to the ground. Those observations were of birds flapping their wings against the wind, but Videler has also made careful observations of slope soaring over a sea dike, against onshore winds avderaging 9 meters per second. The windward side of the dike sloped at 14 degrees to the horizontal, and at the height where the birds flew (on average, 6.5 meters above the ground), the wind was angled upward upward at about 7 degrees to the horizontal. When the bird remained stationary in this situation, its movements *relative to the air* and the forces on it were exactly the same as if it had been gliding in still air at 9 meters per second and at an angle of 7 degrees to the horizontal.

Some soaring birds seem to walk on water. These are the storm petrels, small seabirds of 30 to 40-grams. They hand suspended just above the sea surface with their wings spread but not flapping, dipping their feet into the water from time to time and looking out for the small fish and squid on which they feed. They are using their own peculiar soaring technique, seaanchor soaring, which works by the same principle as the flight of a kite. Their weight is supported

by lift on their wings, while the drag exerted on them by the air is balanced by the forces that the water exerts on their feet. The bird is blown slowly backward over the water, so the drag on the feet is directed forward.

A bird that is good at slope soaring will not be particularly good at thermal soaring, and vice versa, because these two soaring techniques work best at different speeds. Many thermals are only a few tens of meters in diameter, and therefore thermal soaring requires the ability to turn in small circles. If thermal soarers circled at high speed, they would have to increase the lift on their wings greatly to obtain the necessary centripetal force. As a result, induced drag would increase and the birds would sink faster relative to the air. Accordingly, the best thermal soarers are birds that can circle very slowly. In contrast, good slope soarers can glide fast. To remain stationary as kestrels do, they must be able to glide as fast as the wind. To make headway against the wind as albatrosses do, they must be able to glide faster than the wind. Thermal soarers, then, should be good at slow gliding and slope soarers at fast gliding.

The two different soaring strategies are reflected in the areas of the wings. Both the minimum sink speed and the maximum range speed are proportional to the square root of wing loading, so thermal soarers should have low wing loadings (large wing areas for their weights) and slope soarers should have higher wing loadings. As expected, vultures have much bigger wing areas than albatrosses of the same weight; for example, Ruppel's griffon vultures averaging 7.6 kilograms mass (74 newtons weight) had an average wing area of 0.83 square meter, making the wing loading 90 newtons per square meter. Wandering albatrosses averaging 8.7 kilograms mass had 0.61 square-meter wings, giving a wing loading of 140 newtons per square meter.

Comparisons of this kind must be made between birds of similar size, because different-sized birds of similar habits have different wing loadings. To understand this, imagine two geometrically similar birds, one twice as long as the other. The longer bird will also be twice as wide and twice as high, and so it will be eight times as heavy. Its wings are twice as long and twice as broad, so have four times the area. Four times the wing area has to support eight times the weight; as a consequence, wing loading is twice as high for the longer bird as for the shorter. A smaller marine slope soarer like a 1-kilogram white-chinned petrel might be expected to have half the

wing loading of an 8-kilogram albatross, which indeed it has. Similarly, small vultures have lower wing loadings than large ones. For any particular body mass, albatrosses and other slope soarers have smaller wing areas (and so large wing loadings) than vultures and other thermal soarers.

Wing span is the distance from one wing tip to the other, with the wings spread, and the chord is the distance from the front to the rear edge of the wing. Although vulture's wings have much larger areas than those of albatrosses of equal mass, their spans are a little less. The 7.6-kilogram vultures that we have been discussing had wing spans averaging 2.4-meters, and geometrically similar 8.7-kilogram vultures would have had 2.5-meter spans. In contrast, the 8.7-kilogram albatross had a span of 3.0 meters. Thus vultures have short broad wings and albatrosses have long narrow ones. The difference in shape can be described be calculating the aspect ratio, which is the span divided by the mean chord. The aspect ratio is 7 for the vulture that we have been discussing and 15 for the albatross.

As a general rule, the higher the aspect ratio the better the aerodynamic performance of a wing. The reason is that the bigger the span, the more air is driven downward each second. If two wings of different spans traveling at the same speed have to produce the same lift, the one with the longer span is able to do it by driving a larger mass of air downward, with a smaller downward velocity. It gives less kinetic energy to the air than the wing of smaller span, which gives a larger downward velocity to a smaller mass of air. For this reason, a bird with a longer wing span suffers less induced drag. A glider with a longer wing span will lose height less fast than a similar glider with a shorten span, especially at the low speeds at which induced drag is larger than profile drag.

Simple theory suggests that vultures would fly better if they had longer, possibly narrower wings. However, if, say an 8.7-kilogram vulture were to be given an albatross like aspect ratio without reducing its wing area, it would need wings with a 3.7-meter span, 22 percent more than the span of an albatross of the same mass. Wings that long might be difficult to manage when a vulture leaving a kill was taking off from the ground.

There is one group of birds that combines low, vulture-like wing loadings with high, albatross like aspect ratio, but the largest of them have a mass of only about 1.5 kilograms and a wing span of no more than about 2.3 meters, less than some vultures. These

are the frigate birds, tropical seabirds with exceptional soaring habits. They live in the latitudes where the trade winds carry cool air from temperate regions over the warm tropical sea. Here, unlike at other latitudes, there are thermals over the sea, created as air is warmed by the sea surface and rises upward. Frigate birds soar in these thermals, apparently remaining airborne day and night, for they travel far out over the oceans but apparently never land on the sea surface. They feed on flying fishes and so squids that leap out of the water, and they also chase other birds and steal food from them. They need low wing loading for thermal soaring, and long wings are probably less of a problem than they would be for vultures, because frigate birds have no need to take off from the ground. They nest in treetops and can gain speed, when they take off, by diving out of the tree.

Not only do different kinds of birds have different wing loadings to suit their flying habits, but by partially folding its wings, an individual bird can increase its wing loading when it want to go fast. The pigeons that flew in Pennycuick's wind tunnel kept their wings spread to their full 65-centimeter span when gliding at 9 meters per second, but folded their wings progressively as speed increased, until at 22 meters per second the span measured only 25 centimeters. The folding of the wings reduced the wing area from 600 square centimeters with the wings spread at 9 meters per second to 400 square centimeters at 22 meters per second.

9.3.3 Making a Landing

However fast a bird glides, it must slow down for a safe landing. Birds use their feet as air brakes, holding them close to the body in fast flight but lowering them to lose speed. They cannot reduce their speed too much, however, or they will stall and be unable to produce enough lift to support their weight. There are two features of wing design that seem to enable birds to glide more slowly without stalling than would otherwise be possible. One of these is the alula, a tuft of feathers on the front edge of the wing supported by the bone of a rudimentary index finger. It lies flat against the rest of the wing in fast flight, but at low speeds is lifted and may help to keep the air flowing smoothly over the wing at angles of attack at which stalling would otherwise occur. The same principle is used in aircraft, in the device called a leading-edge slot. The stalling speed is probably also reduced by separating the large feathers at the wing tip, giving a multislotted effect. The feathers at the wing tips

of crows, vultures, and many other birds separate in slow flight. Even birds that have such devices seem to stall deliberately when landing: increasing the angle of attack until the wings stall is a very effective way of increasing drag. Stalling can be detected in photographs of landing birds because the irregular airflow over the upper surface of the wing ruffles the feathers.

There are other ways of slowing down, besides using the feet as air brakes or stalling the wings. Ducks land at high speed on ponds and slow themselves by holding their feet in the water. Guillemots landing on the cliffs where they nest approach at a lower level than the nest and veer upward, slowing themselves by converting their kinetic energy (energy of speed) into potential energy (energy of height) and stalling just before landing.

The birds that we have discussed so far all glide well: they can glide at shallow angles, losing height only slowly. There are other animals that glide much less well but are particularly interesting because they give us hints suggesting how birds and bats may have evolved the power of flight.

9.4 FLYING SQUIRRELS

These less good but still very effective gliders include flying squirrels which look much like ordinary squirrels but have flaps of skin between their fore and hind legs. By spreading their limbs they can open out this skin and stretch it tight, making a surprisingly effective airfoil. Much of what we know of the habits of flying squirrels come: from the observations of Keith Scholey, who while a Bristol University graduate student studied a species that live in Borneo. His base was a house in a forest clearing, from which he would watch the flying squirrels moving through the surrounding trees. They spend the day in their nests, in holes in dead trees, but came out at dusk to feed on leaves throughout the night. To find young leaves, which they prefer, they often had to travel some distance through the forest gliding from one tree to the next.

As dusk approached each evening, Scholey settled down on chair on the sun deck, waiting for the flying squirrels to emerge from their holes. With him he had a panoramic photograph of the forest edge and a stopwatch. When the squirrels came out and began to glide, he would mark, on the photograph, the positions on the trees a which each glide start and ended, and he would time the glide: with the stopwatch. He watched the squirrels until they disappeared into the forest or it became too dark to seem them.

On the following day, using surveying instruments, he measured the length of each glide and the height lost.

The flying squirrels travelled by climbing to the top of a tree trunk, gliding to a lower point on a nearby tree, climbing to the top of its trunk, gliding again and so on. They glided distances of 30 to 130 meters, diving steeply at first to gather speed and the continuing a shallower angle, about 12 degrees to the horizontal (losing, that means, about a meter of height for every 5 meters that they traveled). They glided fast, at about 15 meter per second, and so lost height at about 3 meters per second. This rate of loss of height is poor in comparison with that of birds, since birds may lose only 1 meter per second when gliding at their minimum sink speed.

The glide has to be fast because wing loading is high. A specimen of this particular species, the giant red flying squirrel, had a body mass of 1.3 kilograms (weight 13 newtons) and a wing area of 0.11 square meter. The wing loading was thus 120 newtons per square meter, high in comparison with birds.

Scholey estimated that the squirrels lost 6 meters of height in the initial steep dive. A body that starts at rest and falls freely through 6 meters has a final speed of only 11 meters per second. Thus even a dive of this depth is not quite enough to accelerate the animal to 15 meters per second, the speed of most of the glide: the animal must have continued to accelerate in the shallower part of the glide. Because the squirrel needs an initial steep drop to accelerate, it would hardly have been worth gliding if there had not been many trees over 20 meters tall. The flying squirrel's method of travel is much more effective between the tall trunks of a tropical forest that it would be in most North American or European woods.

Near the end of a glide the squirrel is approaching a tree trunk at 15 meters per second and must slow down to avoid a damaging impact. It brakes its onward rush by the same means guillemots use when landing on cliffs, veering upward to convert some of its kinetic energy back into potential energy. The wing stalls as the squirrel slows down, but that does not matter if the animals has timed its action skillfully so that it does not stall until immediately before landing. In principle, the rise at the end of the glide could recover most of the height lost in the initial steep dive, but some height is lost in the fraction of a second before landing, because of the stall.

The squirrels need good control of their flight to reach their intended landing points, steering around any obstacles on the way, and land safely. They display their skill in the breeding season, when they come out of their holes by day and make spectacular glides, apparently as a courtship display. Scholey once saw a squirrel make a midair turn through 180 degrees and land again on the tree from which it had taken off.

Flying squirrels travel reasonably fast. To travel 60 meters, for example, at the gliding speed of 15 meters per second would take only 4 seconds, but a 60-meter glide actually takes 6 seconds because of time lost in the initial dive and in landing. The height lost in a 60-meter glide is about 18 meters (6 meters for the initial dive plus 1 meter for every 5 meters travelled). To recover the lost height the squirrels climb 18 meters up the trunk after landing, which at about 0.7 meter per second takes 26 seconds. Thus the total time needed to travel 60 meters, gliding and then climbing, is about 32 seconds, giving a mean horizontal speed of 1.9 meters per second. It is doubtful whether the squirrel could travel this fast if it did not glide. When an ordinary, nonflying squirrel was trapped and then released and timed as it ran away, it ran at 2.2 meters per second, but that was probably a sprinting speed that could not have been sustained for long. The same species running and jumping from branch to branch through the forest canopy was estimated to travel at only 1.0 meter per second.

9.5 OTHER GLIDERS

There are flying lizards in the tropics as well as flying squirrels, *Draco* has simple wings on the sides of its body, consisting of flaps of skin stiffened by extensions of the ribs. Where these ribs emerge from the body, there are joints that enable the animal to spread its wings or to fold them against its sides. Like the flying squirrels, it glides only moderately well, losing 7 meters of height in a 20-meter glide. Because it feeds on common insects, it does not have to commute through the forest to suitable feeding sites, as flying squirrels do. It climbs up a tree, searching the trunk for insects, then glides to the base of the next tree and searches its trunk for food. Gliding enables the lizard to move to the bottom of the next tree quickly, at little energy cost.

Birds and bats must have evolved from ancestors that flew less well than they do, possibly gliding from tree to tree like flying squirrels and flying lizards. It is easy to imagine bats evolving from an ancestor

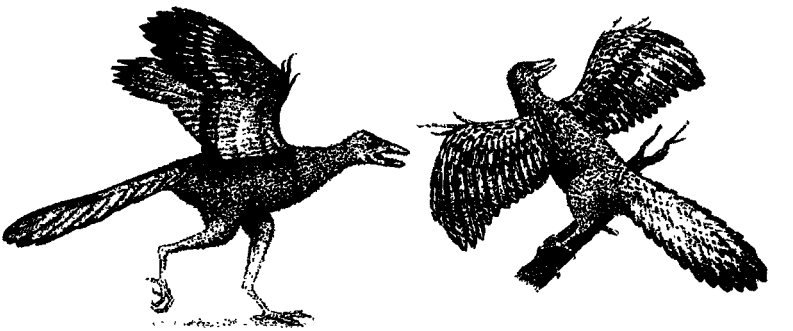


Figure 9.7 *Archaeopteryx*, a bird from the time of the dinosaurs, as it may have looked in life.

that looked like a flying squirrel and that used its gliding ability either as flying squirrels do, to commute through the forest to trees bearing young leaves or fruit, or as *Draco* does, to get from tree to tree while searching for insects. There are two distinct groups of bats, and they are believed by some zoologists to have evolved separately. The largest number of species (including all the North American and European ones) are microbats. Most of these small bats feed on insects. Microbats usually catch their prey in the air, but they may have evolved from an ancestor with feeding habits like *Draco's*. The megabats of Africa and Asia are generally larger than microbats, as their name suggests: the biggest have a mass of about 1.4 kilograms and a wing span of 1.2 meters. They eat only plant food (largely fruit) and probably evolved from an ancestor that behaved like a flying squirrel, as well as looked like one.

There are two theories about the evolution of birds. One is that ancestral birds lived in trees, moving from tree to tree by gliding. The other is that they were fast-running animals that lived on the ground. This theory suggests that these ground-dwelling birds leapt from time to time as they ran, spreading their wings and gliding to extend the leap. The problem with that suggestion is that wings were probably small in the early stages of evolution, so the wing loading would be high and the animals would have to travel very fast indeed to be able to glide. For example, it can be calculated from the wing loading of the giant red flying squirrel that it would probably stall at 13 meters per second. No flying bird or small mammals can run at such a speed, which is quite fast even for antelopes. This chapter has been almost entirely about gliding vertebrates, because few insects make much use of gliding. The

reason seems to be that small flying animals have low wing loading, and therefore their minimum sink and maximum range speeds are also low. They glide too slowly to be slope soarers, for they cannot make headway against any but the slowest winds. Although a slow speed is suitable for gliding up thermals, thermal soarers must be able to glide reasonably fast from one thermal to the next. It seems impossible to design a glider that will sink more slowly than 0.5 meter per second in still air, so a slow glider cannot travel very far before hitting the ground and might find it difficult to reach the next thermal. Vultures, storks, and albatrosses are large, and few soaring birds are small.

Despite these impediments, there are a few large butterflies that soar. Monarch butterflies (wing span 11 centimeters, 4.3 inches) soar while migrating between Canada and Maxico, although their maximum range speed is less than 3 meters per second and they lose a meter of height for every 4 meters gliding in still air. If the wind is against them they use flapping flight, keeping close to the ground where the wind speed is low, or else they do not fly at all: if they tried to soar against the wind, they would be blown in the opposite direction. On good days when the wind is behind them, however, they may spend as much as 80 percent of the time soaring, circling in thermals and slope soaring over building, letting the wind carry them along. This chapter has shown how much animals can achieve, simply by gliding and soaring. Vultures soar in the thermals over the African plains, and storks migrate by thermal soaring between Europe and Africa. Slope soaring supports kestrels as they watch for prey and enables albatrosses to remain airborne day and night over the Antarctic Ocean. Even the lesser gliding skills of flying squirrels enable them to travel faster through the forest than if they had to run. Though so much can be done by gliding alone, even more is possible for animals that are capable of powered flight.

10

Physical Movement in Water

The resistance to the movement of a macroscopic animal through water is determined by its size and velocity, and the density and viscosity of the water. The density of seawater increases steadily with increase in pressure whereas viscosity changes in a biphasic manner, reaching its lowest value at about 500 atm. Neither the viscosity nor density of seawater changes more than 4 per cent as a result of naturally occurring pressures so deep water offers virtually the same resistance to macroscopic animals as does shallow water.

In the case of the movement of microscopic animals and moving parts such as cilia, it is conceivable that pressure may affect the behaviour of water at their surfaces. The small size and the complex shape of microscopic animals, especially small planktonic Crustacea, further complicates the issue. For example, in the swimming of the shallow water copepod *Labidocera trispinosa* Vlymen (1970) has argued that in the 'hop and sink' progression typical of copepods, the bursts of rapid acceleration are not an energetically expensive form of rapid acceleration are not an energetically expensive form of progression. In fact, it is argued that the energy cost of movement through the water is negligible in the animal's total energy budget. Vlymen points out how the animal derives benfit from rapid movement in escaping predators without paying a high cost. It is difficult to imagine such a state of affairs in macroscopic animals where rapid acceleration involves a mass of muscle which undoubtedly requires significant energy from the animal's metabolism. Some bathypelagic fish appear to have relinquished powerful muscle and live a quiescent and economic life.

To a first approximation then, the oceanic does not present mid–water animals with special mechanical problems of locomotion. This is also true of those animals which crawl over, or burrow in, oceanic sediments where the nature of the sediment determines the problems of locomotion. This some what unexciting observation is worth making if only because the floor of the oceanic, covering twice the area of the terrestial environment, is populated by such creatures. Nevertheless, these same benthic organisms may well be interesting in respect of their buoyancy.

Animals tend to sink because protein-based tissues and skeletal meterials are denser than seawater. Lipids, certain body fluids, and gases in bulk are lighter than seawater and are deployed in ways which vary in degree of refinement to bring animals close to neutral buoyancy. At depth, buoyancy devices have to generate expansion against significant hydrostatic pressure, and it will be argued here that uplift is obtained in one of two ways. The increase in partial molar volume of a solute (or its displacement of water) either involves a phase change or a low density arrangement of water around solute particles. Because the problem of oceanic buoyancy forces us to think of molecular events, it is worth noting the molecular forces which act to make tissues denser than seawater in the first place. The case of skeletal salts is simple; these are dense because of their atomic constituents and close packed arrangement.

10.1 PROTEIN

In the dry crytalline state the specific volume (reciprocal of density) is 0.802 which diminishes to 0.751 in aqueous solution. Thus the solid state protein is less dense than the dissolved substance. Part of the change in density is attributed to the internal pressure of water and part to the electro–striction of the water by charged groups on the protein. β–lactoglobulin is probably typical of many proteins. Ovalbumin in the dry state has a density 1.2655, casein 1.318, horse serum globulins 1.279-1.312, and horse serum allumin 1.27-1.28. When the last three proteins enter aqueous solution at room temperature their density was found to increase by 5-8 per cent, doubtless due to the forces already mentioned.

So far as is known, buoyancy devices in marine animals appear to compensate for heavy tissues as if protein-based tissues have a density of 1.3 compared to the density of normal seawater which is 1.026. Buoyant proteins are an intriguing possibility however.

Neutral buoyancy is likely to be of advantage to an animal for two reasons. It can save energy and it can allow the animal to hover inconspicuously. The energetice of neutral buoyancy have been quantified in a most illuminating way but the value of being able to hang silently in the water is less easily quantified.

10.1.1 Oceanic Life

The sea embraces about one half of the entire globe and it forms a special biotype by its unique ecological features in the biosphere.

10.1.2 Bionomic Features

Briefly speaking, there are four factors such as cold, quiescence, darkness and the total absence of living green plants beyond the light zone and this is the summary of conditions in the oceanic.

Oceanic fauna consists of water breathing organisms only and so the higher Arthropods and all the vertebrates above fishes are debarred from such a habitat. Among invertebrates, sponges, corals, echinoderms (Brittle stars, stalked crenoid, sea-cucumbers and starfish, etc.). Bryzoans, Brachiopods, tube dwelling annelids, barnacels, ostracods and pelecypods, decapods (among Mollusea) are chief deepsea fauna.

Vertabrate fauna include sharks, rays, chimeroids (among Elasmobranchs) and many teleosts. The oceanic fishes live under great pressure below the water at great depths. When released from this terrific pressure, *i.e.,* taken out from the sea in which they are adapted, they are spoken of as weak and flabby of flesh. The jaws are relatively great. There great jaws and the armament of cruel mouth are characteristic features of oceanic fishes.

Another characteristic feature of oceanic fishes is their very small size but they can swallow a fish bigger than themselves.

10.1.3 Oceanic Adaptation

The Oceanic animals are frail or weak because of the absence of water-currents at these depths. They have small amount of carthy matters in their bones. The flesh is generally thin and flabby. They have simplified colours red colour dominates over the others. Some oceanic fishes are blind, some have telescopic eyes *e.g. Gigantura,* while other have eyes like concave mirrors. Either of the two latter types of eyes, helps in absorbing the greatest possible volume of light-rays. To compensate for the loss of vision, tactile organs are well developed. Almost all oceanic animals are luminiscent. Many

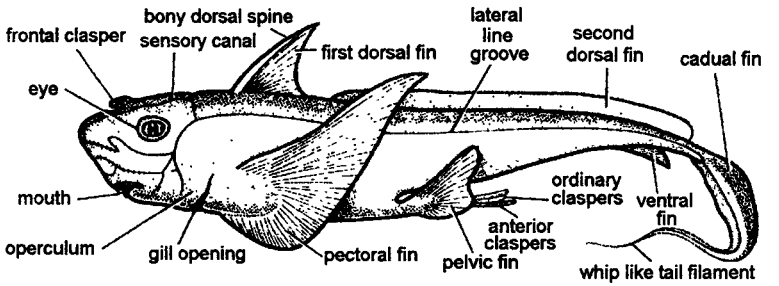


Figure 10.1 *Chimaera* (Rat fish).

oceanic fishes have lost their mastication powers as they live on decaying matters. Others are predatory with powerful jaws.

Perhaps the most striking features of both oceanic and of the cave–dwellers also is their changelessness, *i.e.*, after they have adapted in the peculiar environment, their evolution gets practically ceased as they diminish their individual activity and consequent metabolism.

Oceanic modifications, met in the animals, are the following:

10.1.3.1 Frail and weak body

The animals are frail and weak having thin and flabby flesh and with very little earthy matter in their bones, but they possess relatively great jaws and cruel mouth armanent *e.g.*, silver sharks, *Gastrostomus*. In invertebrates, long legged crabs, delicate long stalked crinoids, fragile siliceous glass sponges and hexactinellids attached to soft oozo through glassy rope like root spicules are the oceanic forms.

10.1.3.2 Simple body colour

Diversified colour of animals of the light zone are lacking from the oceanic forms. They are red or brown, pearly grey or black,

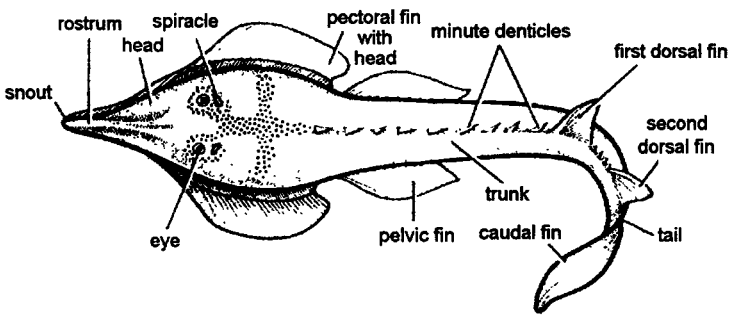


Figure 10.2 *Rhinobatus* (Guitar fish).

although some fishes have scarlet fins. But *Macrurus filicauda* has black belly and is silvery on the top.

10.1.3.3 Eyes

Some oceanic fishes are *blind (Ipnops)* belonging to the family Amblyopsidae; *Oncirodes* (anglers), others have *telescopic eyes,* yet others have *eyes like concave mirrors* to absorb the greatest possible volume of light rays.

In carbs eyes are not found and in pecten these are reduced.

10.1.3.4 Presence of tactile organs

As a compensation for the loss of vision, oceanic forms possess long recrers and slender attenuations of the fins. In *Bathypterois,* one fin ray of the pectoral fins is long and acts as a sensory filament. In *Stylophorus paradoxus,*caudal fin a produced into a long filament, which is sensory.

Crustaceans also possess long antennae. These may be 8 to 10 times the length of the body. A crustacean isopod (*Munnipsis longicornis*) has antennae 8 times longer then the body.

Lateral line system is very well developed in oceanic fishes.

10.1.3.5 Presence or luminescent organs

Luminescent organs are found either over the entire body or on the belly or localized on highly modified organs. These are large oval bodies or irregularly elliptical shape, siuated on the head near the mouth or smaller round globular bodies placed symmetrically in series along the side of the body and tail. Teeth and mouth may also be luminous. It is due to the dark habitat, *e.g., Ipnops, Linophryne, Oneirodes*—deepsea fishes..Luminescence is also found

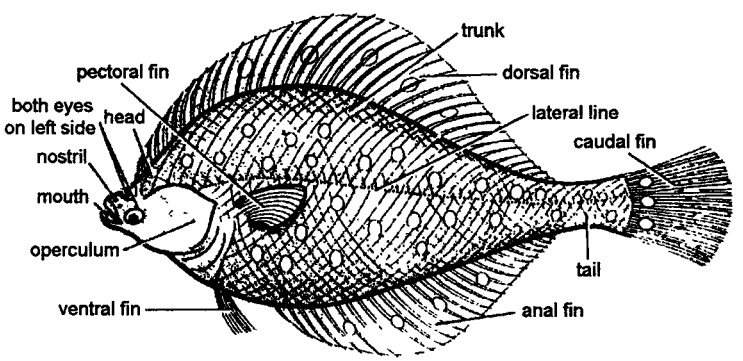


Figure 10.3 *Pleuronectes.*

in crustaceans, cophalopods, starfishes, many coelenterates and annelids. Luminescence is useful for the recognition of sea and for attracting the prey.

10.1.3.6 Loss of power of mastication

Many oceanic fishes loose the masticatory power since these feed on decaying ooze, while others have powerful rapacious jaws, *e.g., Cetomimus.* Most of the oceanic fishes have very large mouths, and long sharp teeth and enormous stomach, capable of devouring large animals, *e.g., Saccopharynx and Eurypharynx.*

10.1.3.7 Sexual dimorphism

In the abyssal dark none, it is very difficult to search the mate. In angler fish (*Photocorynus spiniceps*) the males are very small and remain attached to a process on the head of the female.

10.1.3.8 Perental care

Many oceanic fishes take care of their youngs, while other produces youngs in large numbers.

Vertical distribution of animals. Life exists at all depths, *i.e.,*from pelagic to abyssal zone, but the numbers of individuals decrease as they descend to greater and greater depths. The abyss is comparatively rich in species, but poor in number of individuals. Accourding to Sir John Murray, the number of bottom–dwelling forms or various depths are as follows:

Pelagic region	Down to	200	meters,	about	4,200 species.
"	2000	"	"	600	"
"	4000	"	"	400	"
over	5000	"	"	150	"

10.1.4 Oceanic Fauna

Among *invertebrates* are found the sponges, corals, hydroids and their allies (coelenterates); brittle–stars stalked crinoids, holothurians, and starfishes descend upto 2000 fathoms or more; sea urchins upto 2000 to 3000 fathoms (*Echinodermata*); tube dwelling annelide upto 4000 fathoms deep; 2900 fathoms; amongst Mollusca, the Pelecypoda (*Mytilus phaseolinus*) ranges upto 3000 fathoms having delicate shells; of the arthropods, the barnacles dredged from 3000 fathoms; ostracods from 2000; crabs, shrimps and lobsters ranging down to 2500 fathoms.

Among *vertebrates* are found the following fishes; *Spinax niger* (luminous shark) from 500 to 1500 fathoms; silver sharks (*Chimaera*

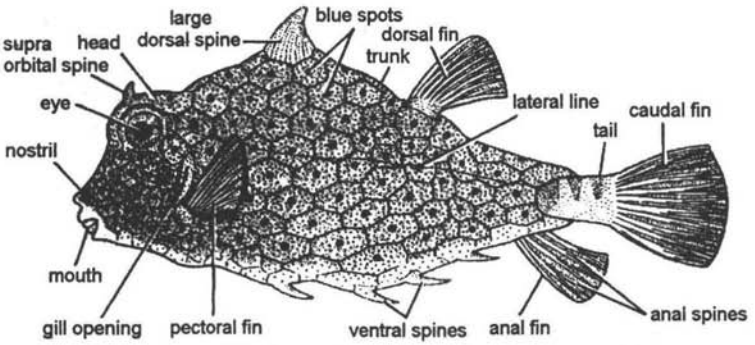


Figure 10.4 *Ostracion.*

affinis, Collorhynchus, Harriott and Chimaera monstrosa) with huge eyes and long attenuated body and tail are abyssal; Esociformes group of teleost fishes includes (i) *Cetomimus* having huge mouth, small teeth, no scales and very small eyes, (ii) *Ipnops* having no eyes and bear two large, luminescent organs on the head to allure prey; Clupeiformes includes oceanic family Stomiatidae which is characterized by delicate scales or none and by well–develope luminescent organs; Anguilliformes (ece–like) inculdes the larva of true eel, known as *Leptocephalus,* another *Gastrostomus* having very long, slender body and enormous jaws are abyssal types; Gadiformes (cod–like) also includes oceanic fishes having mouth and dentition reduces in size, eyes very large, trunk shortened and tail tapering to filament, although they are not abyssal, they are down migrants; angler include a remarkable group of fishes, having horrible jaws and paired fine, have become adapted for bottom crawling and the anterior fin rays of the dorsal fin are tipped with fuminous organs to attract their prey *e.g., Linophryne, Oneirodes* is blind but with luminescent organs.

10.1.5 Swimbladders

The physiology of the swimbladder of deep dea fish embodies such a range of physiological and physical phenomena that it could well be the centre–piece of a book on oceanic physiology. For two good reasons the topic is only dealt with in summary here; first, because the author has no practical experience in this field and second, because the subject is well dealt with in the literature, having attracted the attention of some brilliant experimenters and lucid writers. Accordingly, for the sake of completeness there follows only a brief description of the structrure and function of the swim–bladder

of shallow water fish and a few comments on the special problems of its function at great depth.

In some fish swimbladder connects by a duct to the alimentary canal, but in all oceanic fish it is a closed sac. The gas gland is perfused by capillaries which form a long coonter current rete which is normally buried in the swimbladder wall. The gas gland is the point from which gas leaves solution as a gas phase. The oval is another perfused patch which permits dissolved gas to be transported away from the bladder. Note the absence of a rete there. Gas exchange between oval and bladder is determined by muscular control over the extent to which the oval is exposed to the gas contents. The wall of the swimbladder is highly impermeable to gases and compliant. Its volume in relation to the body it buoys is fairly constant, for example in shallow water fish it lies in the range of 3-6 per cent of the body volume. The relative constancy is, of course, due to the approximately uniform denisities of fish and seawater. In fish which lay down a lot of lipid during their life span, the swimbladder may be relatively small.

Marshall's (1972, 1960) thorough studies and reviews of oceanic swimbladders have yielded a number of important conclusions. About one third of all mesopelagic species of fish possess a swimbladder of a size suited for buoyancy purposes. Others, insluding such common animals as *Cyclothne* spp. and *Stoimias* spp. do not possess functional swimbladder in the adullt state.

Bathypelagic fish living deeper than 2000 m do not possess gas filled swimbladders.

Fish living near the oceanic floor (benthopelagic) possess a gas filled swimbladder irrespective of depth. Marshall states that of the 16 species known to live in this way at depths greater than 3500 m, 11 possess an apparently functional swimbladder.

Benthic fish generally do not have a functional swimbladder, irrespective of depth. In this context the celebrated case of *Basogigas profundissimus* is often quoted. A single specimen was retrived from a trawl which collected sedentary animals from the floor of the Sunda trench (7160 m,1.5°C). On retrieval it was found to have an undamaged swim–bladder, which is difficult to reconcile with the suggestion that it contained sufficient gas at a pressure of 700 atm to give it useful buoyancy. Further, there is no evidence that the trawl was deployed and finally, no evidence to show that the animal was even alive when it was caught.

One might expect at least two pieces of evidence to establish the maximum depth at which swimbladders function as true buoyancy organs. It is essential to know that the fish in question actually possesses a fully developed swimbladder and not much lipid, and secondly it is important to observe the animal floating at depth. The present evidence suggests that useful upthrust might be obtained from a swimbladder in fish at depths to about 5000 m. The rat tail *Coryphaenoides* has been filmed 'floating or swimming slowly', presumably at a depth of 2000 m.

Pressure is not the most important limiting factor in the distribution of gas filled swimbladders; an adequate food supply appears an equally important prerequisite to their inclusion in the repertoire of fish physiology.

The density of atmospheric gases at oceanic pressures and temperatures is less than that of the liquid hydrocarbons or lipides. Oxygen, the major constituent of the gas mixtures in the swimbladders found at septh, has a density of approximately 0.45 cm^{-3} at 400 atm. The upthrust obtained from a given volume of swimbladder gas measured at environmental pressure declines markedly in the bathypelagic zone. Obviously the region where change in depth affects a given gas volume most is the top few hundred metres.

10.1.5.1 Functioning of the gas filled swimbladder at depth

The solubility of atmospheric gases physically dissolved in the blood and the association of oxygen with haemoglobin decreases as the blood enters the effetent capillaries where a higher lactate level prevails. The partial pressure rises in accordance with Henry's Law, causing gas to diffuse across to the afferent rete and, by a continuous process, builds up a high partial pressure of gas in the blood at the swimbladder end of the afferent rete. This high partial pressure sustains that of the gas in the swimbladder.

Taking the rate of blood flow and the trans–rete diffusion of gas into account, an equation defining the partial pressure of gas in the blood at the gas gland end of the afferent capillaries has been derived by Enns:

$$P_L = P_0 \exp\left(\frac{K}{F} L \frac{\Delta\alpha}{\alpha}\right) \qquad ...(1)$$

where P_L = partial pressure of stated gas at the gas gland end of afferent capillaries, P_0 = partial pressure of stated gas at the entrance

to afferent capillaries, K = trans–rete diffusion of gas (Q (ml min^{-1} $length^{-1}$ pressure $difference^{-1}$), divided by α the solubility coefficient of the stated gas in distilled water), F = blood flow rate (ml min^{-1}), L = length of rete, and

$$\frac{\Delta\alpha}{\alpha} = \frac{\text{gas solubility in distilled water} - \text{solubility in solution}}{\text{solubility in distilled water}}$$

The important point to notre is that K is a very large figure and it dominates the exponential term; $\Delta\alpha/\alpha$, or the salting out effect, is a small figure.

Taking $\Delta\alpha/\alpha$ as 0.00153 (a value based on the afferent–efferent difference in lactate in shallow water fish), F as 1 ml min^{-1}, and L as 1 cm, then P_L equals 112 atm when P_o is 0.78 atm as it is in the case of nitrogen. Thus the generation of 100 atm of nitrogen is broadly accounted for, provided leakage through the swimbladder wall is small, which it is in shallow water conditions. It therefore appears that a shallow water swimbladder only requires a long rete if it is called upon to buoy a fish at considerable depth. To reinforce that over–simplifying assertion we can consider the permeability of the swimbladder wall of shallow water fish to very steep partial pressure gradients. Table shows the permeability of the wall of two swimbladders from shallow water fish. The low permeability of the conger eel bladder is attributes to the presence of guanine crystals and is unlikely to be upset by a high ambient pressure. Kutchai and Steen argue that oxygen will leak through the wall of the conger swimbladder at a rate equal to a plausible rate of secretion when the partial pressure of the gas equals 250–fold that of the ambient water, i.e., at 250 atm or 2500 m.

The equilibrium pressure of dissolved gas rises with increase in hydrostatic pressure in an exponential manner. This some what surprising state of affairs has already been discussed in Chapter 4. The experimental verification of the theromodynamic arguments which yield this conclusion is limited, but in the absence of data to the contrary we will assume, with Enns *et al.* (1967) that dissolved gas activity rises 14% per 100 atm all the way to the greastest depths. These workers make the point that if we regard the sea as equilibrated with nitrogen at normal atmospheric pressure and consider a swimbladder being inflates with nitrogen, then at both 3000 m depth and a hypothetical depth of 13300 m, the ratio of P_{N_2} in the bladder to P_{N_2} in solution in the sea will be 246. The

situation with oxygen is more complecated but it is clear that the exponential ries in gas equilibrium pressure with hydrostatic pressure helps rather than hinders the maintenance of gas in a swimbladder at great depth.

10.1.5.2 Special features of oceanic swimbladders—the rete

L, in equation (1) is increased in oceanic fish. Table shows data selected from Marshall (1972) to illustrate this point and the original papers should be consulted to verify the generalisation. The rete in Table are uncorrected fot body size or swimbladder size but the correlation with living depth is real and clearly connected with the working of the countercurrent multiplier.

Table 10.1 Length of swimbladder rete in oceanic fish.

Species	Length of rete (mm)	No. of rete	Range of depths where most commonly found (m)
Bassogigas profundissimus	15	1	
Bassozetus taenia	25	2	4570—5610
Nemotonurus armatus	25	5	2600—3600
Lionus carapinus	25	6	> 2000
Coryphaenoides guntheri	20	4	1000—2000
Synaphobranchus Kaupi	10	2	800—2000
Hymenocephalus italicus	6	2	300—800
Malacocephalus laevis	4	2	150—600
Vinciguerria attenuata	0.8	—	150—600
Myctophum punctatum	2	—	150—600

The energy required to maintain the high partial pressure of gas within the swimbladder has been estimated by Alexander (1972). If the partial pressure of gas (oxygen or nitrogen) in the ambient seawater is P_W and the partial pressure in the swimbladder is P_S, then the work of compressing the gas us RT In (P_S/P_W); R and T are the gas constant and absolute temperature repectively. Diffusional losses must be matched by the secretion of gas to maintain a steady state. Let D represent the number of moles of gas diffusing out of the swimbladder, a quantity which can be compuuted from permeability data and measurements of a particular swimbladder. Thus the steady state power requires to maintain a swimbladder volume constant at depth is DRT In (P_S/P_W). Using permeability data from the conger eel swimbladder and arbitrarily assuming that

gas secretion proceeds with an efficiency of 5 per cent, Alexander calculates a 1 g fish would expend energy equivalent to respiring 35 cm^3 O_2 kg^{-1} $hour^{-1}$ at 100 atm pressure to maintain its swimbladder full of oxygen. This is 17 times the amount of energy required at 10 atm pressure, and would appear to be rather a high of expenditure to apply to a real fish. The efficiency of gas secretion in even shallow water fish may be much higher than 5 per cent. Alexander's argument shows that swimbladders may become metabolically very demanding at great depth and we might reasonably anticipate adaptations in oceanic swimbladders to improve their efficiency.

10.1.6 Composition of the Gas in Swimbladders

Oxygen is the major gas in oceanic swimbladders, but nitrogen and argon are usually present at partial pressures in excess of that prevailing in the ambient seawater. These results are consistent with a gas secreting machanism which works in a physical rather than a selective chemical way. High concentrations of oxygen were also found in the swimbladders of hatchet fish retrieved from the oxygen minimum layer at a depth of about 800 m. Under these conditions the P_{O_2} of the gas was 10000 times the ambient P_{O_2} which may still be accounted for by the 'salting out' mechanism of gas secretion. Note also the volume of gas present in these animals.

10.1.7 Deleterious Effects of High Partial Pressures of Gases

At partial pressures in excess of about 1 ATA oxygen is toxic to intact animals and impairs the function of many *in vitro* preparations.

Nitrogen at partial pressures of about 4 ATA exerts a depressive action on the mental performance of men, and at slightly higher pressures narcotises a number of physiological systems. The gas gland of the swimbladder works at elevated partial pressures of both these gases, and in particular, the gas gland of oceanic fish functions at a P_{O_2} of 100 or more atmospheres. D'Aoust (1969) has brought this into sharp focus by subjecting the epipelagic fish *Sebastodes miniatus* to a partial pressure of oxygen which only its gas gland would normally encounter. It is not yet clear if a gas gland which tolerates a few atmospheres partial pressure of oxygen would remain unaffected by a P_{O_2} of 100 atm. Intuitively, one might suppose further adaptive changes would be necessary. In this connexion pressure itself may play a role in modifying the potency of the gas.

The sensitivity of gas glands to high pressures of nitrogen has not been studied. In general, there is at present little evidence to

show that they are exposed to partial pressures much greater than 20 atm, but such pressures are, of course, mildly narcotic in some systems. Hydrostatic pressure ca n enhance or antagonise the narcotic action of nitrogen in a number of preparations.

In a shallow water species, gas bubbles have been seen in a live gas gland. Lipids rich in cholesterol and phospholipids have been found within the swimbladder of *Coryphaenoides acrolepis* and *Antimora rostrata* and it is conceivable that these substandes are connected with the secretory process. The role and metabolism of lipids both inside and outside swimbladders of oceanic fish is obscrure.

10.1.8 $\Delta\alpha/\alpha$ at Great Depth

A generalised salting out of dissolved gases could give rise to a multiplying effect in the rete sufficient to generate gas pressures in excess of those needed in the oceanic. In addition, a shift in the equilibrium $HbO_2 = Hb + O_2$ could also cause oxygen to leave the blood and enter the swimbladder. From solubility measurements carried out by Enns *et al.*(1967) the ratiofor argon and nitrogen can be computed for different molarities. It would be interesting to know how hydrostatic pressure affects these data.

If lactate released into the efferent capillary blood decreases the solubility of nitrogen in a manner comparable to NaCl then quite high partial pressures of nitrogen may be slowly generated. Lactate also shifts blood pH which might affect the equilibrium shown above. Indeed Scholander and Van Dam (1954) have shown that acidified blood will dissociate off oxygen at quite high pressures (Root effect), but probably not at a P_{O_2} of more than 40 atm. Perhaps other substances can releade oxygen from oxyhaemoglobin against higher partial pressures. Oxygen may be driven from the aqueous and haemoglobin components of blood by two different machanisms. A pH effect on the oxygen binding capacity of haemoglobin dissociates the oxygen off into solution. A salting out effect drives it from solution to diffuse into the swimbladder. The first machanism is limited to a P_{O_2} of 40 atm, the second is not limited by a naturally occurring pressure.

The problem of the rate of gas secretion is important. For example, Enns *et al.* (1967) suggest that a hypothetical 100 g fish at 200 atm might take 174 days to fill its swimbladder with oxygen. In short, in swimbladders working at great depth the equilibrium situation is broadly explicable. The rate at which buoyancy equilibrium is achieved in oceanic fish is not known but extrapolation from

those systems which have been studied suggests that it may be achieved only very slowly unless supplementary mechanisms exist.

10.1.9 Buoyancy by a Gas Filled, Rigid Shell

Certain cephalopods achieve neutral buoyancy means of a gas filled shell. The pressure within the shell is atmospheric or less and the shell with–stands the ambient pressure. *Nautilus* and *Spirula* are mesopelagic animals living at pressure of less than 100 atm. Their coiled and chambered shells fail catastrophically at pressures of 60 and 170 atm respectively. The mechanism by which the shell, which is initially filled with water, is evacuated during the growth of these animals, is not fully understood; nor is it clear how the siphuncle, the perfusing stem running the length of the shell, withstands the ambient hydrostatic pressure and remains waterproof. An osmotic pressure gradient maintained by the active transport of salts appears one of the forces which holds back the water pressure, but such a mechanism in its simple version cannot account for the way in which *Spirula* withstands hydrostatic pressures greater than the osmotic pressure of its blood (about 20 atm).

10.2 LIPIDS

Some inactive myctophids (lantern fish) show an interesting reduction in their swimbaldders and an accumulation of lipid material during their life. Capen (1967) has studied example of this progress in some detail and found both a 'cottony' tissue growing inside and an oily tissue outside the swimbladders of, for example, *Lampancytus mexicanus*. The adults of a number of mid–water fish of the deep scattering layer may be slightly negatively buoyant in sea–water and it may be that gas filled swimbladders are too slow to adjust during vertical migrations. Other myctophids including the genus *Lumpanyctus* sp. contain interesting lipids. Certain species contain large amounts of wax esters which probably provide significant upthrust (e.g. *Stenobrachius*); others, e.g. *Diaphus,* contain negligible amounts. Presumably,. the proportion of these substances present in an animal is related to its buoyancy and nutritional economy.

An outstanding example of a liquid buoyancy system is seen in a number of oceanic sharks whose livers contain a large quantity of the hydrocarbon squalene. Its specific gravity is 0.86 and when present in sufficient amounts it is capable of giving appreciable lift. Heller, Heller, Springer and Clarke (1957) found the liver oils in *Dalatias licha* and *Centrophorus uyato* to contain at least 70 per cent squalene, whereas in many other species, *Squalus* sp., *Carcharhinus*

sp. and *Scyliorhinus* sp. squalene constituted less than 1 per cent of the liver oils. Corner, Denton and Forster (1969) observed that sharks such as *Dalatias licha* floated in surface seawater and subsequently demonstrates that the squalene was responsible for the animals buoyancy. This was first seen qualitatively in dead animals which were caused to sink by the removal of their livers. A more precise demonstration invovled corrections for the differences in salinity, temperature and pressure between the oceanic water where the animals live and surface seawater in which they were weighed. Table shows the example of *Centroscymnus coelolopic* which, in Plymouth seawater, reqyured the addition of 18 g to achieve neutral buoyancy. At the animals normal depth (about 1500 m) the salinity was greater than that of Plymouth water and a corresponding correction was made by Corner *et al.* in the following way. If the volume of the animals is *V,* and the surface seawater density in which the animal floated with an upthrust of 18 g is d_a, then the change in that upthrust will be $V(d_a - d_b)$ g when the animal is transferred to a different salinity seawater of density d_b. The effects of temperature and pressure were allowed for separately. The coolor oceanic water caused a change in the upthrust prosuced by the squalene which was related to the remperature coefficient of expansion of oil and seawater. The upthrust in grams of oil floating in water changes with temperature according to αW in which W is the weight of oil and α is $d_B/d_2 - d_A/d_1$ where d_A and d_B are seawater densities at high and low temperatures and d_1 and d_2 are densities of oil at the respective temperatures. Oils are caused to contract a lot more than seawater by a temperture decrease and in the buoyancy balance sheet it is seen that this is the largest single adjustment.

The negative buoyancy of the animal lacking its liver is affected by the cooler oceanic water to the extent. Corrections for the effect of pressure on the buoyancy of squalene and the negative buoyancy of the animal lacking a liver are both small. The compres-sibility of squalene has not been directly determined but it probably lies in the range typical of oils generally which is slightly in excess of that of seawater. It would be interesting to know the melting points of squalene and lipids at high pressure and low temperatures. The buoyancy correction is given by 10^{-5} × $(P_B - P_A)$ × (density of seawater) × (volume of oil) where P_A and P_B are 1 atm and the normal ambient pressure of the animal respectively. The difference between the bulk compres-sibilities of protein and seawater is even smaller and Corner *et al.* quite reasonably discounted it in the

buoyancy balance sheet. How much swimming effort does a shark have to make in order to move its huge squalene–filled liver through the water? Indirectly this question in answered by Alexander's theoretical studies (1972) which show that, to a first approximation and in small fish, the energy cost of moving a mass of lipid through the water is much less than the cost of providing an equivalent amount of hydrodynamic lift in the absence of the lipid. In many sharks the pectoral fins generate uplift and Corner *et al.* (1969) showed that in neutrally buoyant sharks the pectoral fins are significantly smaller than the fins of similar animals which are negatively buoyant.

The metabolic cost of synthesising squalene has been estimated at 0.7 cal g^{-1} compare to 0.5 cal g^{-1} for oleic acid, but more upthrust is obtained from squalene. The control of the synthesis of squalene is presumably connected with any control the animals might have over their buoyancy. Figures from Corner *et al.* (1969) suggest that the mass of squalene would have to be controlled to within 1 per cent in order to keep the animal within 0.1 per cent of neutral buoyancy. Malins and Barone (1970) have demonstrated a possible basis for the regulation of the upthrust provided by the liver of deep water sharks. They used *Squalus acanthias* whose liver does not contain large amounts of squalene. Although the whole animal is not neutrally buoyant, the liver does contain significant amounts of diacyl glyceryl ethers and triglycerides. 3 kg animals which were loaded with 100 g weights showed a change in the mixture of liver lipids over 2 days, with diacyl glyceryl ethers present in higher proportions. The specific gravity of the diacyl glycerol ethers is lower than that of the triglycerdes so the changed ratio would to dome extent, offset the effect of the weights.

The naturally occurring lipid which might provide most buoyancy in animals, pristane, has a specific gravity of 0.78 but is not found in sufficient quantities to exert much effect. Blumer, Mullin and Thomas (1963) have shown that starving *Calanus* do not utilise pristane, which only constitutes 1-3 per cent of the animal's lipid. The energy cost of negative buoyancy in copepods may well be negligible in the overall economy of the animal anyway.

10.2.1 No Buyancy Device

A conditin of near neutral buoyabcy is reached in some bathypelagic fishes by a reduction of the skeleton and muscles. The buoyancy balance sheet of *Gonostoma elongatum* for example reads

as follows, The weight of protein in water is 1.1 g per 100 g of fish, and the weight if skeletal material is also 1.1 g per cent. Upthrust is provided by the dilute body fluids, 1.2 g per 100 m fish, and from lipid to the extent of 0.5 g per cent. Thus without bulk lipid or gas buoyancy devices this fish weight only 2.2–1.7 g or 0.5 g per 100 g in seawater, which approximates a fifth of the weight in water of a normal shallow water fish with its swimbladder removed. This ingenious economy is probably closely related to the animal's nutritional economy and feeding habits.

10.2.3 Aqueous Solutions

Certain marine animals, representatives of which may live in the oceanic, reduce their weight in water through the regulation of their ionic contents. The coleomic fluid in a variety of squid and a few Crustacea is isosmotic but lighter than seawater due to the presence of ammonium ions. *Gigantocypris mulleri* is one of the oceanic crustaceans which buoys itself in this way. Other invertebrates appear to exclude SO_4^{2-} ions, thereby reducing weight. The partial molar volume of SO_4^{2-} in aqueous solution is large and negative; it is an intensely hydrated ion and notably heavy. The prevalence of these buoyancy devices in the oceanic is not known. Only a limited uplift is obtained from a large bulk of 'light' fluid.

10.2.4 Some Conclusions About Buoyancy

The act of generating upthrust either involves motive power or molecular expansion. The oceanic has little effect on the resistance of water to muscular propulsion but exerts a variety of effects on buoyancy mechanisms. Some of the buoyancy devices employed in shallow water are adaptable to great depths. The buoyancy chamber of the cephalopods *Spirula* and *Nautilus* appears severely limited, and a theoretical study of its depth limitations would be an interesting sequel to the full elucidation of the buoyancy mechanism in these animals. Other buoyancy systems work less well at increasing depths but are not subject to such an abrupot limitation by pressure as is the case in *Spirula.* Bulk phase bouyancy utilising gas or lipid molecules, is achieved by a phase change and the accompanying expansion of the system is due, in the case of lipids, to the formation of hydrophobic bonds. The energy required to generate the gas pressure required in oceanic swimbladders is not particularly demanding, but the work required to sustain adequate buoyabcy against diffusional losses, or to vary buoyancy daily, seems high.

Aqueous systems are different as no phase change is seen. Uplift is obtained by the selective maintenance of ions in which the process of molecular expansion has already taken place. Heavy and heavily hydrated ions are excluded. Although little upthrust is obtained by light aqueous solutions, oceanic animals may well be found to make extensive use of this type of buoyancy device.

The energy cost of generating upthrust at depth has yet to be investigated. Alexander's theoretical approach points the way. Making plausible assumptions, Alexander (1972) calculates that in fairly shallow water upthrust by muscular propulsion at constant depth uses about 12 times the energy which upthrust from a swimbladder would cost. Upthrust obtained from buoyant lipids is intermediate in its energy cost. At the 1000 m (100 atm) depth level the energy saving advantage of a buoyant swimbladder seems greatly reduced, but the widespread occurrence of swimbladders at depths greater than this suggests otherwise.

As we descend into the depths Nature has increasing difficulty in providing neutral buoyancy to save energy which, in turn, is becoming increasingly scarce. A more satisfactory description these interacting forces and solutions to them can only come from physiological study of specific cases. Barham (1971) has suggested how two types of mesopelagic myctophids may be distinguished. There is an active type, exemplified by *Myctophum,* which have muscles, and in those cases which possess a swimbladder it is well developed and perhaps capable of rapid gas exchange. It is these active myctophids which migratre all the way to the surface at night and are menbers of the deep scattrering layer in daytime. Inactive myctophids like *Lampanyctus mexicanus,* with a reduced swimbladder, high lipid content and weak body musculature, migrate over a lesser range and many have seen to hang vertically and immobile in the water during daylight hours. Barham has also made the interesting observation that these suspended animals may obtain upthrust from their opercular pump mechanism.

The biochemistry, physiology and behaviour of mesopelagic fish have a lot to tell us about how different design–solutioins to a common physical environment are arrived at. The oceanic elasmobranchs, far less numerous than muctophids but rather more robust as experimental animals, may prove good bathypelagic animals to study.

11

Radioimmunoassays

Of the numerous quantitative immunoassays, the RIA is perhaps the most well established. The method was developed for the analysis of insulin in blood. Medically, the RIA is now used not only to measure the levels of critical circulating hormones, but also to quantitate the levels of numerous drugs in serum, cerebrospinal fluid, or urine. In basic research, the RIA has become extremely valuable for studying the regulation and distribution of various endogenous substances, including peptide hormones, steroids, and neurotransmitters. The RIA will continue to be a valuable technique in neurotransmitters. The RIA will continue to be a valuable technique in neurobiology research because the assay is highly sensitive and less subject to artifacts than are many other quantitative immunoassays. Furthermore, it can readily be used in conjunction with various chromatographic procedures.

In this chapter we focus on RIAs for peptides and proteins, although many of the techniques are also applicable to other classes of compounds.

11.1 PRINCIPLES OF THE RADIOIMMUNOASSAY

The RIA is based on the displacement of a radiolabeled antigen, or tracer (Ag*), by an unlabeled antigen (Ag) from a limited number of binding sites on antibodies (Ab). The relationships of the components of the assay can be appreciated by studying the following reaction:

$$2Ab + Ag + Ag^* = AbAg + AbAg^*$$

In the concentrations of tracer and antibody in the final reaction mixture are held constant, the ratio of the amount of antibody-

bound tracer to the amount of unbound tracer (AbAg*/Ag*) will depend on the amount of unlabeled antigen present. As the amount of unlabeled antigen increases, the amount of bound tracer decreases and the amount of unbound tracer increases, thus decreasing the ratio of bound to unbound tracer.

To perform an RIA, the following are required :

1. Antibodies specific for the antigen of interest
2. Unlabeled antigen of a known concentration for constructing the standard curve.
3. A method of obtaining a purified radiolabeled antigen (i.e., the tracer) that retains the ability to bind to the antibody.
4. A method of separating the antibody-bound antigen (AbAg and AbAg*) from the unbound antigen (Ag and Ag*)
5. A sample containing the antigen of interest at an unknown concentration (this concentration will be determined in the RIA).

A crucial element in the RIA is the use of an unlabeled antigen of a known concentration to prepare a standard curve. The standard curve represents the amount of tracer displaced by various known concentrations of this unlabeled antigen standard. The concentration of the antigen of interest in a sample is determined by comparing the amount of tracer displaced by the sample to the amounts displaced by various concentrations of the unlabeled antigen standard.

11.2 ANTIBODIES FOR THE RADIOIMMNOASSAY

Antibodies with a high affinity for the antigen are preferable for RIAs. A high-affinity monoclonal antibody can be excellent antibody for developing an RIA, but such antibodies may be difficult to generate. Polyclonal antisera are therefore used almost universally for RIAs. It should be noted that antisera that work well for ELISAs or immunocytochemistry may not necessarily work well in RIAs and vice versa. The criteria for a good antibody differ for each of these procedures.

11.3 GENERATING RADIOLABELED ANTIGENS FOR USE AS TRACERS

Generating the tracer for an RIA generally requires a source of antigen of high purity that can be radiolabeled or method of purifying the antigen after is labeled. Many isotopes are available for labeling peptides and proteins, but ^{125}I is the isotope of choice for peptide and protein RIAs.

11.3.1 Iodination

The isotope ^{125}I emits primarily low-energy γ particles and has a half-life of approximately 60 days. Many methods are used routinely for incorporating this isotope into proteins. The majority of these methods use reagents such as Chloramine-T, IODO-GEN, IODO-BEADS, and lactoperoxidase to oxidize Na ^{125}I. These reagents oxidize I^- to I_3^- which covalently modifies the phenolic ring of tyrosine residues, generating mono and di-iodo derivatives. In less favorable reactions, it covalently modifies the heterocyclic ring of histidine residues and the sulfhydryl groups of cysteine residues. The sulfhydryl derivatives are usually unstable and break down, but some are stabilized by the local environment in the peptide or protein.

Here, we provide iodination procedures that use either chloramines-T or IODO-GENE as the oxidizing reagent. Although chloramines-T is a somewhat harsh oxidizing reagent, it can be used in a reliable and very simple iodination procedure. Because the oxidization proceeds very rapidly and releases large amount of volatile $^{125}I_2$, it is important to e well prepared for each step of the reaction so that the danger of contamination from volatile radiation can be minimized. For iodination using IODO-GEN, glass tubes are coated with IODO-GEN (1,2,3,6-tetrachloro-3α,6α -diphenylglycoluril), which is an oxidizing agent that is insoluble in aqueous buffers. A peptide or protein in a solution containing Na^{125}I can be covalently derivatized with Na ^{125}I by placing the solution in a tube coated with IODO-GEN. The reaction is stopped by simply removing the solution from the IODO-GEN-coated tube. The iodination method using IODO-GEN is more gentle and requires fewer manipulations than the method using chloramines-T; the timing is also not as critical.

If the peptide or protein is unstable to oxidizing reagents, does not contain tyrosine residues, or is to be used with an antibody that cannot tolerate an iodinated residue at the binding site, coupling a pre-iodinated reagent (e.g., Bolton-Hunter reagent) to the antigen may provide a useful alternative. The Bolton-Hunter reagent reacts readily with primary amines, such as those found on the side chain of lysine residues and free amino termini of peptides and proteins. Although iodination using the Bolton-Hunter reagent may provide a useful alternative to oxidative iodination, it is considerably more expensive than the methods above. Bolton-Hunter reagent is available commercially as the noniodinated reagent or as the mono-^{125}I or di-^{125}I-labeled derivative. The di-iodo derivative of the Bolton-Hunter reagent provides the highest specific activity and is probably the

most useful form of the reagent for iodinating peptides and proteins. It consists of a di-iodinated aromatic ring linked to *N*-hydroxy-succinimide. Although it is cheaper to use iodinated reagent synthesized by oxidative iodination of the non-iodinated compound, it is far less complicated to use the pre-iodinated derivatives. We would not recommend synthesizing the iodinated reagent because of the increased handling of radioactivity required and the inherent lability of the Bolton-Hunter reagent. Because the Bolton-Hunter reagent is labile in water, it is preferable to couple it to an antigen dissolved in an organic solvent such as DMF. For antigens that are not soluble in DMF, aqueous buffers can be used in the procedures on pages 472-474. Some proteins that are not iodinated well using IODO-GEN plus $Na^{125}I$ or the Bolton-Hunter reagent can be labeled to a higher specific activity using diazotized ^{125}I-labeled iodosulfanilic acid. Diazotized ^{125}I-labeled iodosulfanilic acid was originally used as a membrane-impermeable reagent for labeling membrane proteins. It reacts with tyrosine, histidine, and amino groups. Although a labeling kit is available for this type of iodination, it is far less expensive to prepare the reagents.

11.3.2 Labeling with Tritium

An alternative isotope often used in RIAs is ^{3}H. It is frequently used for RIAs of steroids and other alkaloids. ^{3}H has a considerably longer half-life than the ^{125}I isotope and is considered to be less of a health hazard. Because there are many reactions by which ^{3}H can be incorporated into the antigen (e.g., reductive methylation, which is often the method of choice), the integrity of the epitope can usually be maintained during radiolabeling. Many tritiated compounds are also commercially available for RIA protocols. The primary disadvantage of using ^{3}H is that it is a β-particle emitter. In general, tritiated compounds have lower specific activities than radioiodinated compounds. They also require scintillation counting, which is more labor-intensive than gamma counting and is sensitive to counting artifacts such as quenching.

11.4 PROTOCOLS

11.4.1 Iodination Using Chloramine-T

11.4.1.1 Materials

- 01.5 M Potassium phosphate buffer (pH 7.0)
- Chloramine-T: Protect from light and store at room temperature with a desiccant.

- Sodium metabisulfite

  $Na^{125}I$, carrier-free (~1000-2500 Ci/mmole, 100 mCi/ml; Amersham IMS.30, New England Nuclear NEZ-033A):

  If 200-100 μCi of $Na^{125}I$ is to be used in each iodination reaction, use the $Na^{125}I$ as supplied by the manufacturer. If 100-200 μCi of $Na^{125}I$ is to be used in each iodination reaction, dilute the $Na^{125}I$ with 0.01 N NaOH to a final $Na^{125}I$ concentration of 2 mCi/100 μl. $Na^{125}I$ is stable for weeks at room temperature in 0.01 N NaOH.

  Note: Commercially available $Na^{125}I$ is dissolved in various concentrations of NaOH. The buffer used during the iodination reaction should be able to maintain the pH at approximately 7.5. If a buffer or $Na^{125}I$ preparation different from those specified here is sued, make certain that the buffer capacity is sufficient to maintain the appropriate pH after the $Na^{125}I$ added.

- Peptide or Protein sample:

  The amount of peptide or protein used for iodination will depend on how the radiolabeled product is to be employed. The ratio of peptide or protein to $Na^{125}I$ must be optimized empirically. For RIAs, we recommend starting with approximately 1 nmole of trysoine (calculate this from the amino acid composition or perform amino acid analysis for each mCi of $Na^{125}I$ in the iodination reaction. For immunoprecipitations or other biochemical procedures, lower ratios of tyrosine to $Na^{125}I$ may be desirable (e.g., 0.01-0.1 nmole of tyrosine for each mCi of $Na^{125}I$). For a peptide or protein of unknown composition, the number of nanomoles of tyrosine in the sample can be estimated form the weight of the sample by assuming an average molecular weight of 100 each amino acid and the presence of one tyrosine for every 20 amino acid residues. The sample should be solubilized in 20-50 μl of 0.15 M postassium phosphate buffer (pH 7.0) in a 1.9 ml Eppendorf tube. Most samples are stable at –20°C, at least for short periods of time.

  Note: If the peptide or protein is insoluble in potassium phosphate buffer or a similar buffer, other conditions may be tried. Oxidative iodination can be performed under denaturing conditions (e.g., in buffers containing 1% SDS [SD is incompatible with potassium phosphate buffer], 8 M urea, or 4 M guanidine HCl), or in nondenaturing detergents (e.g., in buffers containing either 1% CHAPS or 1% n-octylglucoise). The detergent Triton X-100

cannot be used because it reacts with $^{125}I_3^-$. Proteins can also be labeled in whole cells, membrane fragments, or isolated organelles.

- *For gel-filtration chromatography*:
  2-ml Pasteur pipette plugged with glass wool
- Sephadex G-10 resin (Pharmacia)
  Note: Sephadex G-25 or G-50 can be used for large proteins.
- Elution buffer:
  The buffer should contain a protein, a detergent, or an organic acid to prevent nonspecific adsorption of the iodinated peptide or protein to the resin, glass or glass wool. We usually use 50% acetic acid in distilled water for chromatography of peptides. For chromatography of proteins, we usually use one of the RIA buffers containing 0.2% Tween 20 plus 0.1% gelatin.
- 12-mm × 75 mm Polypropylene tubes
- Hand-held gamma counter
- Glycerol
- Ascorbic acid
- Gelatin (Bloom 175)

Cautions: The $^{125}I_2$ formed during oxidation of $Na^{125}I$ is volatile. Work in an approved chemical fume hood with a charcoal filter when exposing the $Na^{125}I$ to oxidizing reagents such as chloramines-T, IODO-GEN, or acids. Because the oxidation proceeds very rapidly and release large amounts of volatile $^{125}I_2$ when chloramines-T is used, it is important to be well prepared for each step of the reaction so that the danger of contamination from volatile radiation can be minimized. All forms of the isotope should be shielded by lead. When handing the isotope, wear one or two pairs of gloves, depending on the amount of isotope being used and the difficulty of the manipulation required. This isotope accumulated in the thyroid and is a potential health hazard. Consult the local radiation safety office for further guidance in the appropriate use of radioactive materials. Glacial acetic acid is volatile and should be used in a chemical fume hood. Concentrated acids should be handled with great care; gloves and a face protector should be worn.

11.4.1.2 Procedure

1. Prepare a 1 mg/ml solution of chloramines-T in 0.15 M postassium phosphate buffer and a 2 mg/ml solution of sodium metabisulfite in 0.15 M potassium phosphate buffer.

2. Add 100-1000 μCi of $Na^{125}I$ to the 20-50-μl peptide or protein sample.
3. Add 10 μl of freshly prepared chloramines-T solution. Incubate for 15 seconds at room temperature with gentle shaking.
4. At the end of the incubation, add 20 μl of freshly prepared sodium metabisulfite solution to stop the oxidation.
5. Perform gel-filtration chromatography to separate the iodinated peptide or protein from the free $Na^{125}I$ and low-molecular-weight by-products of the reaction.
   (a) Prepare a Sephadex G-10 column by swelling the resin in elution buffer and packing it into a 2-ml Pasteur pipette plugged with glass wool until it is approximately 1 cm from the top of the pipette.
   (b) Place the pipette in a 12-mm × 75-mm polypropylene tube. Apply the entire iodination reaction mixture to the column and collect the eluate.

   Note: This type of column will not run dry.
   (c) Apply 100 μl of elution buffer to the column and collect the eluate.
   (d) Apply 100 μl of elution buffer to the column and collect the eluate.
   (e) Transfer the pipette to the first of five numbered 12-mm × 75 mm polypropylene tubes. Apply 200 μl of elution buffer to the column and collect the eluate. Repeat this step for each of the five tubes.

   Note: The movement of the iodinated peptide or protein down the column can be monitored with a hand-held gamma counter. The labeled peptide or protein usually elutes into the second or third numbered tube.
   (f) Store the iodinated peptide or protein in the elution buffer at –20°C. For proteins that are unstable or that precipitate when frozen, include glycerol at a final concentration of 50%. For iodinated peptide or protein solutions that prove to be unstable, include ascorbic acid and gelatin at final concentrations of 1% and 0.1% respectively, to prevent oxidation or damage by free radicals.

Note: Chromatographic procedures are the most convenient separation procedures. Some of these procedures (e.g., reserved-phase HPLC may separate iodinated peptide form noniodinated

peptide and thereby increase the overall specific activity of the iodinated peptide. (To date, there are no procedures for separating large iodinated proteins form noniodinated proteins). The gel-filtration chromatography descried above is a general purification that is routinely performed. It removes free iodide salts and yields a product that is sufficiently pure for most uses.

11.4.2 Iodination Using IODO-GEN

11.4.2.1 Materials

- IODO-GEN (Pierce)
- Methylene chloride or chloroform
- 12-mm × 75-mm Borosilicate glass tubes. Make certain that the tubes are dry and unused.
- SpeedVac Concentrator (Savant Instrument) (*optional*)
- 0.1 M Potassium phosphate buffer (pH 7.0) or an equivalent buffer.
- Peptide or protein sample:

  The amount of peptide or protein used for iodination will depend on how the radiolabeled product is to be employed. The ratio of peptide or protein to $Na^{125}I$ must be optimized empirically. For RIAs, we recommended starting with approximately 1 n mole of tyrosine for each mCi of $Na^{125}I$ in the iodination reaction. For immunoprecipitations or other biochemical procedures, lower ratios of tyrosine to $Na^{125}I$ may be desirable (e.g., 0.01–0.1 nmoles of tyrosine for each mCi of $Na^{125}I$). For a peptide or protein of unknown composition, the number of nanomoles of tyrosine in the sample can be estimated from the weight of the sample by assuming an average molecular weight of 100 for each amino acid and the presence of one tyrosine for every 20 amino acid residues. The sample should be solubilized in 50–100 μl of 0.1 M potassium phosphate buffer (pH 7.0) or an equivalent buffer in a 1.9-ml Eppendorf tube. Most samples are stale at –20°C, at least for short periods of time.

  Note: if the peptide or protein is insoluble in potassium phosphate buffer or a similar buffer, other conditions may be tried. Oxidative iodination can be performed under denaturating conditions (e.g., in buffers containing 1% SDS [SDS is incompatible with potassium phosphate buffer], 8 M urea, or 4 M guanidine HCl)

or in nondenaturing detergents (e.g., in buffers containing either 1% CHAPS or 1% n-octyglucoside). The detergent Triton X-100 cannot be used because it reacts with $^{125}I_3$. Proteins can also be labeled in whole cells, membrane fragments, or isolated organelles.

- $Na^{125}I$, carrier-free (~1000-2500 Ci/mmole, 100 mCi/ml; Amersham IMS.30, New England Nuclear NEZ-033A)
- Eppendorf tubes
- Materials for gel-filtration chromatography
- Materials for precipitating proteins with TCA (*optional*)

Cautions: Methylene chloride is toxic if inhaled, ingested, or absorbed through the skin. It is also an irritant and is suspected to be a carcinogen. Wear gloves and safety glasses when handling methylene chloride. Avoid breathing in the vapours.

Chloroform is irritating to the skin, eyes, mucous membranes, and upper respiratory tract. It should only be used in a chemical fume hood. Gloves and safety glass should be worn. Chloroform is a carcinogen and may damage the liver and kidneys.

The $^{125}I_2$ formed during oxidation of $Na^{125}I$ is volatile. Work in an approved chemical fume hood with a charcoal filter when exposing the $Na^{125}I$ to oxidizing reagents such as chloramines-T, IODO-GEN, or acids. All forms of the isotope should be shielded by lead. When handling the isotope, wear one or two pairs of gloves, depending on the amount of isotope being used and the difficulty of the manipulation required. This isotope accumulates in the thyroid and is a potential health hazard. Consult the local radiation safety office for further guidance in the appropriate use of radioactive materials.

11.4.2.2 Procedure

1. Dissolve the IODO-GEN in methylene chloride or chloroform to make a 10-40 μg/ml solution.
2. Add 100 μl of IODO-GEN solution to each borosilicate glass tube.
3. Evaporate the organic solvent using a stream of nitrogen gas or a SpeedVac Concentrator.
4. Either cap the dried tubes or cover them with Parafilm M to keep dust out of them. Store the tubes at 4°C with a desiccant (e.g., Drierite) until needed.
5. Just before starting the iodination, rinse the IODO-GEN-coated tubes twice with 0.1 M potassium phosphate buffer or an

equivalent buffer. Maker certain that the IODO-GEN coating does not peel off the tube.

6. Add 50-100 μl peptide or protein sample to each IODO-GEN-coated tube.
7. Add 100-1000 μCi of $Na^{125}I$ to each tube and incubate for 10-15 minutes at room temperature. Swirl the tubes every 5 minutes during the incubation.

   Note :The optimal incubation time can be determined empirically.
8. Transfer each reaction mixture into an Eppendorf tube that is not coated with IODO-GEN and incubate for 15 minutes at room temperature. This allows all of the unreacted $^{125}I_3^-$ that has been generated to reduce to $^{125}I^-$.
9. Apply the reaction mixture to a Sephadex G-10 or G-25 column and separate the iodinated peptide or protein from the free $Na^{125}I$ and by-products of the reaction.
10. (Optional) An accurate measure of the incorporation of the radioactive isotope into the protein can be obtained as follows:
    (a) Dilute a 1–2-μl aliquot of the iodinated protein to a total volume of 400 μl with the buffer used in the protein solution.
    (b) Precipitate the protein with TCA.
    (c) Measure the amount of radioactivity in the pellet in a gamma counter. A 100-μg sample of IgG may incorporate 10-20% of the total $Na^{125}I$ added in step 7.

11.4.3 Iodination Using Bolton-Hunter Reagent

11.4.3.1 Materials

- 15-ml Polypropylene tubes with Caps
- For antigens that are soluble in nonaqueous solutions:
  (a) AG 50W-X2 or AG 50W-X4 ion-exchange resin (Bio-Red)
  (b) DMF (sequanal grade)
  (c) LABQUAKE shaker (Labindustries)
  (d) Triethylamine (sequanal grade): If the triethylamine has been stored for a long period of time or if the presence of primary or secondary amines is suspected, perform the ninhydrin test to check for primary or secondary amines. If a blue color develops in the ninhydrin test, purchase fresh triethylamine.
  (e) Molecular sieve (4A) (Sigma)

- For antigens that are not soluble in nonaqueous solutions:
  Use an aqueous buffer that contains no primary or secondary amines or other nucleophiles. Acceptable buffers include 0.1-1 M phosphate, borate, or bicarbonate (~pH 8.0–8.5). Do not use buffers such as Tris or MOPS.
  Note: The pH of the reaction should be adjusted to obtain optimal labeling. The reaction is faster at pH 9.0 that at pH 6.5, but Bolton-Hunter reagent also hydrolyzes in water faster at the higher pH.
- Peptide or protein antigen
- Bolton-Hunter reagent, either the mono-^{125}I- or di-^{125}I-labeled derivative
  Note: We usually use the 250-μCi container from New England Nuclear (NEX-120 or NEX-120H). The reagent from Amersham is equally acceptable.
- Materials for gel-filtration chromatography
  Note: Sephadex G-25 or G-50 may be more effective than Sephadex G-10 in separating uncoupled Bolton-Hunter reagent and its by-products from large peptides or proteins.

Cautions: DMS is irritating to the eyes, skin, and mucous membranes. It can exert its toxic effects through inhalation, absorption, through the skin, or ingestion. Chronic inhalation can cause liver and kidney damage. Wear safety glasses and gloves when handling DMF.

Triethylamine is flammable. It is extremely corrosive to the mucous membranes, upper respiratory tract, eyes, and skin. It may be harmful if inhaled, ingested, or absorbed through the skin. It should be used in a chemical fume hood. Gloves and safety glasses should be worn.

Radioiodinated Bolton-Hunter reagent should be handled in chemical fume hood. Wear gloves when handling radioactive substances. The isotope ^{125}I-accumulates in the thyroid and is a potential health hazard. Consult the local radiation safety officer for further guidance in the appropriate use of radioactive materials.

Benzene is toxic and flammable. Avoid contact with it and avoid breathing in the fumes.

11.4.3.2 Procedure

1. For antigens that are soluble in nonaqueous solutions, prepare a mixture of DMF and triethylamine as follows:

(a) Add 0.5–1 g of ion-exchange resin to 10 ml of DMF in a 15-ml polypropylene tube. Cap the tube and invert it slowly on a LAB-QUAKE shaker overnight at 4°C to remove the primary amines from the DMF.

(b) Allow the resin to settle, and then decant the DMF into another 15-ml tube. Add triethylamine to a final concentration of 2%, and then add approximately 1 g of molecular sieve. Cap the tube and invert it slowly for approximately 8 hours at 4°C.

(c) The DMF/triethylamine may be stored at this point for up to 1 week at 4°C under nitrogen before it is used in step 2.

2. Dissolve the antigen in 25-50 μl of DMF/triethylamine (or if the antigen is not soluble in DMF, in 25-50 μl of aqueous buffer that contains no nucleophiles). Use the equivalent of approximately 1 nmole of primary amino groups on the antigen (calculate this from the amino acid composition or perform amino acid analysis for each mCi of isotope on the Bolton-Hunter reagent to be used in the iodination reaction. For antigens of unknown composition, use 0.2-10 μg of antigen for each mCi of isotope.
3. Dry the Bolton-Hunter reagent, which usually comes dissolved in benzene, using a stream of nitrogen gas or another inert gas. The supplied charcoal trap should be filtered properly to keep volatile radioactivity from escaping. For further information on preparing the reagent for use, refer to the manufacturer's instructions.
4. Add the dissolved antigen to the dried Bolton-Hunter reagent. Incubate for 2–12 hours at 4°C.
5. Separate the iodinated antigen from the by-products of the reaction by gel-filtration chromatography as described in step 5.

Note: Other chromatographic procedures, such as reversed-phase HPLC may be used. For small peptides, reversed-phase HPLC may provide a pure product.

11.4.4 Iodination Using Diazotized ^{125}-Labeled Iodosulfanilic Acid

11.4.4.1 Materials

- NA^{125}I carrier-free (~1000–2500 Ci/nmole, 100 mCi/ml; Amersham IMS 30, New England Nuclear NEZ-033A).

- Sulfanilic acid: Prepare a 10 ng/μl solution in distilled water. Store at –20°C,
- IODO-GEN-coated tubes.
- Eppendorf tubes
- 50 mM Sodium nitrite
- 0.1 N HCl
- 0.1 N NaOH
- Protein solution:

  Prepare a 0.2-1 μg/ml solution in 0.1 M HEPES (pH 7.4) or an equivalent buffer. Denaturing conditions (e.g., buffers containing 1% SDS, 8 M urea, or 4 M guanidine HCl) or nondenaturing detergents (e.g., buffers containing 1% CHAPS or 1$ *n*-octyl-glucoside) can be used for labeling. Triton X-100 and buffering agents containing primary amino groups or phenol groups should not be used because they may react with the diazotized ^{125}I-labeled iodosulfanilic acid.
- Materials for gel-filtration chromatography. Use Sephadex G-25 resin for this procedure.

Cautions : The $^{125}I_2$ formed during oxidation of Na^{125}I is volatile. Work in an approved chemical fume hood with a charcoal filter when exposing the Na^{125}I to oxidizing reagents such as chloramines-T, IODO-Gen, or acids. All forms of the isotope should be shielded by lead. When handling the isotope, wear one or two pairs of gloves, depending on the amount of isotope being used and the difficulty of the manipulation required. This isotope accumulates in the thyroid and is a potential health hazard. Consult the local radiation safety office for further guidance in the appropriate use of radioactive materials.

Concentrated acids and bases should be handled with great care; gloves and a face protector should be worn. Sodium nitrite is irritating to the eyes, mucous membranes, upper respiratory tract, and skin. It may be harmful if inhaled, ingested, or absorbed through the skin. It should only be used in a chemical fume hood. Gloves and safety glasses should be worn. Wear a mask when weighing SDS. Guanidine HCl is irritating to the mucous membranes, upper respiratory tract, skin, and eyes. Avoid breathing in the dust. Wear gloves and safety glasses when handling it.

11.4.4.2 Procedure

1. Mix 5 μl of Na^{125}I with 40 μl of distilled water, and then add 5 μl of sulfanilic acid solution.

2. Transfer the mixture to an IODO-GEN-coated tube. Incubate for 15 minutes at room temperature to iodinate the sulfanilic acid. Swirl the tube every 5 minutes during the incubation.
3. Transfer the ^{125}I-labeled iodosulfanilic acid to an Eppendorf tube. Allow it to sit for 15 minutes at room temperature, and then continue with step 4.

   Note: If necessary, the ^{125}I-labeled iodosulfanilic acid can be stored for up to a few hours at 4°C before it is used in step 4.
4. Mix 30 μl of the ^{125}I-labeled iodosulfanilic acid, 16 μl of sodium nitrite solution, and 16 μl of 0.1 N HCl in an Eppendorf tube. Incubate for 20 minutes at 4°C to make diazotized ^{125}I-labeled iodosulfanilic acid.
5. Add 16 μl of 0.1 N NaOH to stop the reaction.
6. Immediately add the entire mixture to 100 μl of protein solution. Mix well and incubate overnight at 4°C.
7. Apply the reaction mixture to a Sephadex G-25 column and separate the iodinated protein from the free Na^{125}I and by-products of the reaction.

"This page is Intentionally Left Blank"