10

Physical methods of determining the three-dimensional structure of proteins

Introduction

Advances in biochemistry in the 20th century included the development of methods leading to protein structure determination. Atomic level resolution requires that the positions of, for example, carbon, nitrogen, and oxygen atoms are known with precision and certainly with respect to each other. By knowing the positions of most, if not all, atoms sophisticated 'pictures' of proteins were established as highlighted by many of the diagrams shown in previous chapters.

The impact of structural methods on descriptions of protein function includes understanding the mechanism of oxygen binding and allosteric activity in haemoglobin as well as comprehending the catalytic activity of simple enzymes such as lysozyme. It would extend further to understanding large enzymes such as cyclo-oxygenase, type II restriction endonucleases, and amino acyl tRNA synthetases. Structural techniques are also applied to membrane proteins such as photosynthetic reaction centres, together with complexes of respiratory chains. More recently, the role of macromolecular systems such as the proteasome or ribosome were elucidated by the generation of superbly refined atomic structures. All of these advances stem from understanding the arrangement of atoms within proteins and how these topologies are uniquely suited to their individual biological roles.

At the beginning of 2004 over 22 000 protein structures (22 348) were deposited in the Protein Data-Bank. Over 86 percent of all experimentally derived structures (∼19 400) were the result of crystallographic studies, with most of the remaining structures solved using nuclear magnetic resonance (NMR) spectroscopy. Slowly a third technique, cryoelectron microscopy (cryo-EM) is gaining ground on the established techniques and is proving particularly suitable for asymmetric macromolecular systems. Although this chapter will focus principally on the experimental basis behind the application of X-ray crystallography and NMR spectroscopy to the determination of protein structure the increasing impact of electron crystallography warrants a discussion of this technique. This approach was originally used in determining the structure of bacteriorhodopsin and is likely to expand in use in the forthcoming years as structural methods are applied to increasingly complex structures.

Proteins: Structure and Function by David Whitford

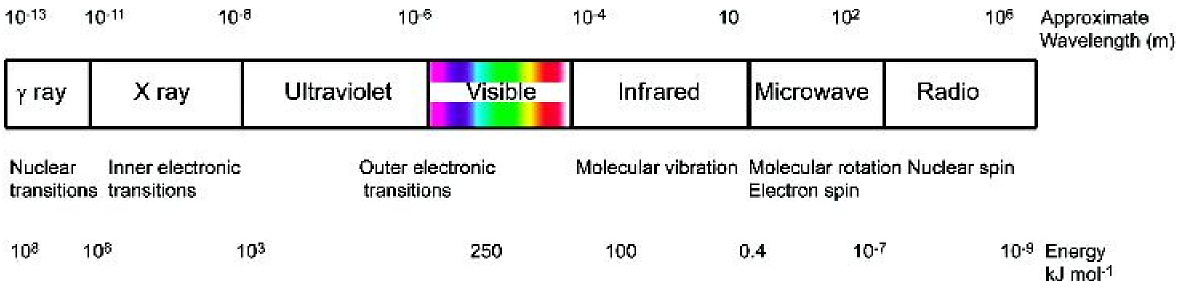


Figure 10.1 Bar showing the distribution of wavelengths associated different regions of the electromagnetic spectrum used in the study of protein structure

These methods are the only experimental techniques yielding 'structures' but other biophysical methods provide information on specific regions of a protein. Spectrophotometric methods such as circular dichroism (CD) provide details on the helical content of proteins or the asymmetric environment of aromatic residues. UV-visible absorbance spectrophotometry assists in identifying metal ions, aromatic groups or co-factors attached to proteins whilst fluorescence methods indicate local environment for tryptophan side chains.

The use of electromagnetic radiation

It is useful to appreciate the different regions of the electromagnetic spectrum involved in each of the experimental methods described in this chapter. The electromagnetic spectrum extends over a wide range of frequencies (or wavelengths) and includes radio waves, microwaves, the infrared region, the familiar ultraviolet and visible regions of the spectrum, eventually reaching very short wavelength or high frequency X-rays. The energy (E) associated with radiation is defined by Planck's law

$$E = h\nu \qquad \text{where } \nu = c/\lambda \tag{10.1}$$

where c is the velocity of light, and h, Planck's constant, has a magnitude of 6.6×10^{-34} Js, and ν is the frequency of the radiation. The product of frequency times wavelength (λ) yields the velocity – a constant of $\sim 3 \times 10^{8}$ m s^{-1}. From the relationship $c = \nu\lambda$ it is apparent that X-rays have very short wavelengths of approximately 0.15×10^{-9} m or less. At the opposite end of the frequency spectrum radio waves have longer wavelengths, often in excess of 10 m (Figure 10.1). The frequency scale is significant because each domain is exploited today in biochemistry to examine different atomic properties or motions present in proteins (Table 10.1).

Radio waves and microwaves cause changes to the magnetic properties of atoms that are detected by several techniques including NMR and electron spin resonance (ESR). In the microwave and infrared regions of the spectrum irradiation causes bond movements and in particular motions about bonds such as 'stretching', 'bending' or rotation. The ultraviolet (UV) and visible regions of the electromagnetic spectrum are of higher energy and probe changes in electronic structure through transitions occurring to electrons in the outer shells of atoms. Fluorescence and absorbance methods are widely used in protein biochemistry and are based on these transitions. Finally X-rays are used to probe changes to the inner electron shells of atoms. These techniques require high energies to 'knock' inner electrons from their shells and this is reflected in the frequency of such transitions ($\sim 10^{18}$ Hz).

All branches of spectroscopy involve either absorption or emission of radiation and are governed by a fundamental equation

$$\Delta E = E_2 - E_1 = h\nu \tag{10.2}$$

where E_2 and E_1 are the energies of the two quantized states involved in the transition. Most branches of spectroscopy involve the absorption of radiation with the elevation of the atom or molecule from a ground state to one or more excited states. All spectral

Table 10.1 The frequency range and atomic parameters central to physical techniques used to study protein structure

Technique	Frequency range (Hz)	Measurement
NMR	$\sim 0.6 - 60 \times 10^{7}$	Nucleus' magnetic field
ESR	$\sim 1\text{–}30 \times 10^{9}$	Electron's magnetic field
Microwave	$\sim 0.1\text{–}60 \times 10^{10}$	Molecular rotation
Infrared	$\sim 0.6\text{–}400 \times 10^{12}$	Bond vibrations and bending
Ultraviolet/visible	$\sim 7.5\text{–}300 \times 10^{14}$	Outer core electron transitions
Mossbauer	$\sim 3\text{–}300 \times 10^{16}$	Inner core electron transitions
X-ray	$\sim 1.5\text{–}15 \times 10^{18}$	Inner core electron transitions

lines have a non-zero width usually defined by a bandwidth measured at half-maximum amplitude. If transitions occur between two discrete and well-defined energy levels then one would expect a line of infinite intensity and of zero width. This is never observed, and Heisenberg's uncertainty principle states that

$$\Delta E \Delta t \approx \mathrm{h}/2\pi \qquad (10.3)$$

where ΔE and Δt are the uncertainties associated with the energy and lifetimes of the transition. Transitions do not occur between two absolutely defined energy levels but involve a series of sub-states. Thus if the lifetime is short (Δt is small) it leads to a correspondingly large value of ΔE, where ΔE defines the width of the absorption line (Figure 10.2).

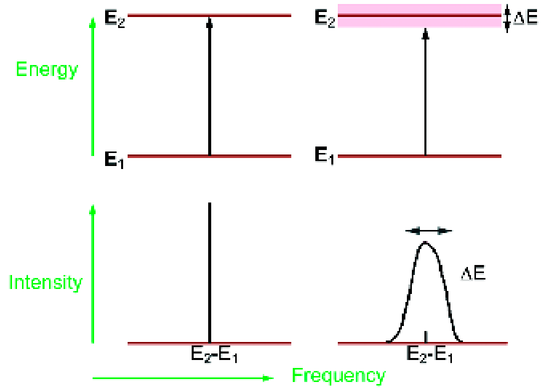


Figure 10.2 Theoretical absorption line of zero width and a line of finite width (ΔE)

X-ray crystallography

X-ray crystallography is the pre-eminent technique in the determination of protein structure progressing from the first low-resolution structures of myoglobin to highly refined structures of macromolecular complexes. X-rays, discovered by Willem Röentgen, were shown to be diffracted by crystals in 1912 by Max von Laue. Of perhaps greater significance was the research of Lawrence Bragg, working with his father William Bragg, who interpreted the patterns of spots obtained on photographic plates located close to crystals exposed to X-rays. Bragg realized 'focusing effects' arise if X-rays are reflected by series of atomic planes and he formulated a direct relationship between the crystal structure and its diffraction pattern that is now called Bragg's law. All crystallography since Bragg has centred around a basic arrangement of an X-ray source incident on a crystal located close to a detector (Figure 10.3). Historically, detection involved sensitive photographic films but today's detection methods include charged coupled devices (CCD) and are enhanced by synchrotron radiation, an intense source of X-rays.

Bragg recognized that sets of parallel lattice planes would 'select' from incident radiation those wavelengths corresponding to integral multiples of this wavelength. Peaks of intensity for the scattered X-rays are observed when the angle of incidence is equal to the angle of scattering and the path length difference is equal to an integer number of wavelengths. From diagrams such as Figure 10.4 it is relatively

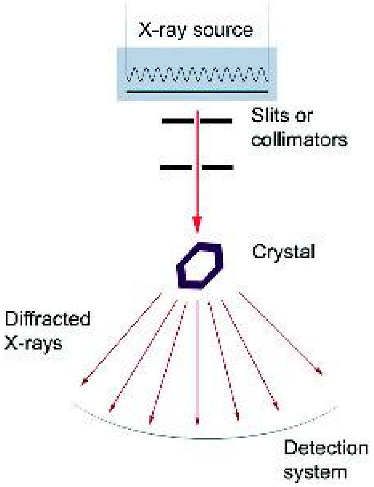


Figure 10.3 The basic crystallography 'set-up' used in X-ray diffraction. Monochromatic X-rays of wavelength 1.5418 Å, sometimes called Cu Kα X-rays, denote the dislodging of an electron from the K shell and the movement of an electron from the next electronic shell (L). After passing through filters to remove Kβ radiation the X-rays strike the crystal

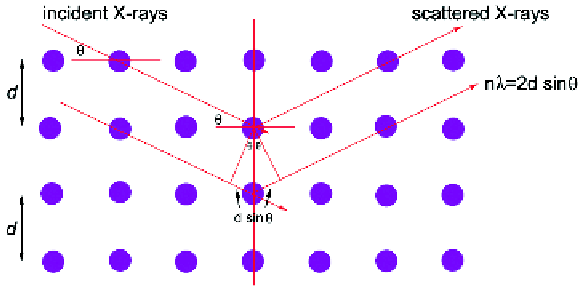


Figure 10.4 X-rays scattered by a crystal lattice

straightforward to establish the path difference ($n\lambda$) using geometric principles as

$$n\lambda = 2d\sin\theta \tag{10.4}$$

This equality, a quantitative statement of Bragg's law, allows information about the crystal structure to be determined since the wavelength of X-rays is closely controlled. From the Bragg equation diffraction maxima are observed when the path length difference for the scattered X-rays is a whole number of wavelengths. The arrangement of atoms can be redrawn to make this equality more clear. The path difference is equivalent to

$$d\cos\theta_{\mathrm{i}} - d\cos\theta_{\mathrm{r}} = n\lambda \qquad (\text{where } n = 1, 2, 3, \ldots) \tag{10.5}$$

This formulation is readily extended into three dimensions and is called the Laue set of equations (Figure 10.5). The Laue equations must be satisfied for diffraction to occur but are more cumbersome to deal with and lack the elegant simplicity of the Bragg equation. For each dimension Laue equations are written as

$$a(\cos\alpha_{\mathrm{i}} - \cos\alpha_{\mathrm{r}}) = h\lambda \qquad (\text{where } h = 1, 2, 3, \ldots) \tag{10.6}$$

$$b(\cos\beta_{\mathrm{i}} - \cos\beta_{\mathrm{r}}) = k\lambda \qquad (\text{where } k = 1, 2, 3, \ldots) \tag{10.7}$$

$$c(\cos\gamma_{\mathrm{i}} - \cos\gamma_{\mathrm{r}}) = l\lambda \qquad (\text{where } l = 1, 2, 3, \ldots) \tag{10.8}$$

where a, b and c refers to the spacing in each of the three dimensions (shown by d in Figure 10.5).

Within any crystal the basic repeating pattern is the unit cell (Figure 10.6) and in some crystals more than one unit cell is recognized. In these instances the simplest unit cell is chosen governed by a series of selection rules. The unit cell can be translated (moved sideways but not rotated) in any direction within the crystal to yield an identical arrangement (Figure 10.7).

The unit cell, the basic building block of a crystal, is repeated infinitely in three dimensions but is characterized by three vectors (a, b, c) that form the edges of a paralleliped. The unit cell is also defined by three angles between these vectors (α, the angle between b and c; β, the angle between a

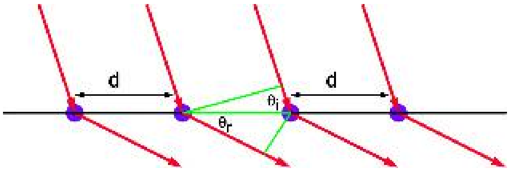


Figure 10.5 The Laue equations

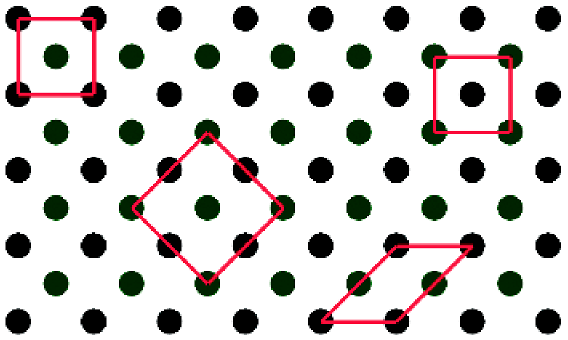


Figure 10.6 Possible unit cells in a two-dimensional lattice

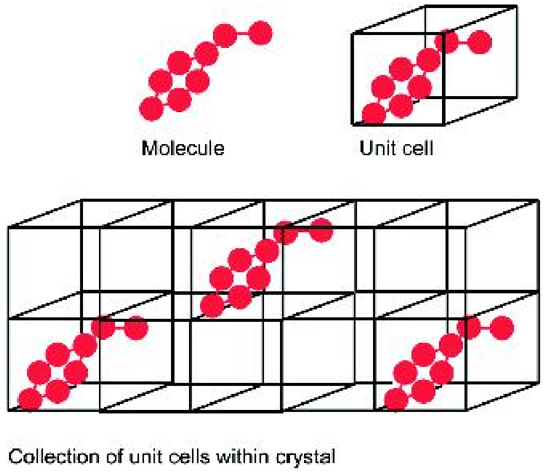


Figure 10.7 A unit cell for a simple molecule

and c; γ, the angle between a and b, Figure 10.8). The recognition of different arrangements within a unit cell was recognized by Auguste Bravais during the 19th century.[1] In two dimensions there are five distinct Bravais lattices, whilst in three dimensions the number extends to 14, usually classified into crystal types. Any crystal will belong to one of seven possible designs and the symmetry of these systems introduces

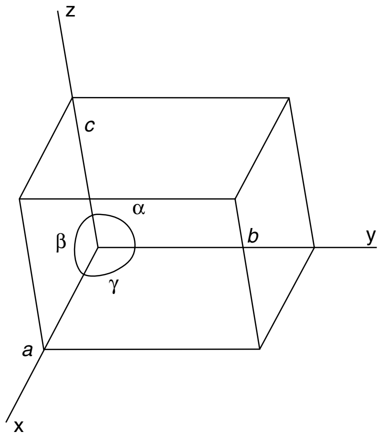


Figure 10.8 The angles and vectors defining any unit cell

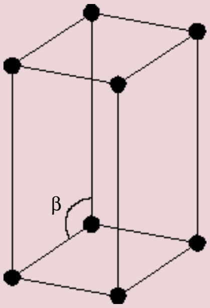
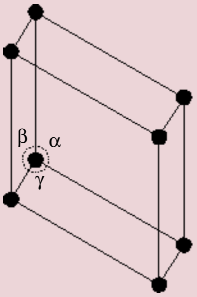
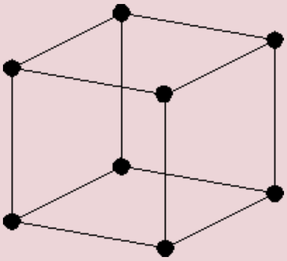
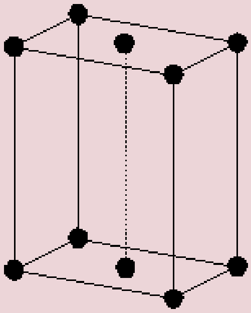
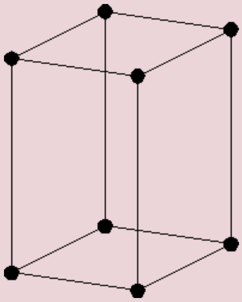
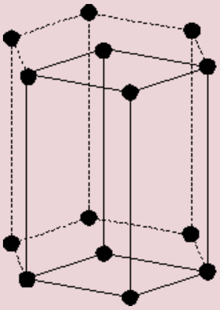
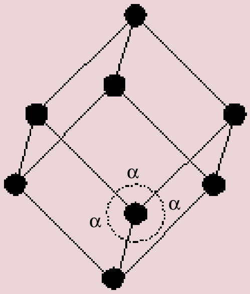
constraints on the possible values of the unit cell parameters. The seven crystal systems are triclinic, monoclinic, orthorhombic, tetragonal, rhombohedral, hexagonal and cubic (Table 10.2). The description of unit cells and lattice types owes much to the origins of crystallography within the field of mineralogy.

In biological systems the unit cell may possess internal symmetry containing more than one protein molecule related to others via axes or planes of symmetry. A series of symmetry operations allows the generation of coordinates for atoms in the next unit cell and includes operations such as translations in a plane, rotations around an axis, reflection (as in a mirror), and simultaneous rotation and inversion. Collection of symmetry operations that define a particular crystallographic arrangement are known as space groups, and with 230 recognized space groups published by the International Union of Crystallography crystals have been found in most, but not all, arrangements.

Scattering depends on the properties of the crystal lattice and is the result of interactions between the incident X-rays and the electrons of atoms within the crystal. As a result metal atoms such as iron or copper and atoms such as sulfur are very effective at scattering X-rays whilst smaller atoms such as the proton are ineffective. The end result of X-ray diffraction experiment is not a picture of atoms, but rather a map of the distribution of electrons in the

[1] First described by Frankenheim in 1835 who incorrectly assigned 15 different structures as opposed to the correct number of 14, recognized by Bravais which to this day carries only his name.

Table 10.2 The parameters (a, b, c and α, β, γ) governing the different crystal lattices together with some of their simpler arrangements

Crystal system	Representation	Crystal system	Representation
Number of Bravais lattices. Vector and angle properties. Monoclinic 2 $\alpha = \gamma = 90^\circ$		Number of Bravais lattices. Vector and angle properties Triclinic 1 No restrictions	
Cubic 3 $a = b = c$; $\alpha = \beta = \gamma = 90^\circ$		Orthorhombic 4 $\alpha = \beta = \gamma = 90^\circ$	
Tetragonal 2 $a = b$; $\alpha = \beta = \gamma = 90^\circ$		Hexagonal 1 $a = b$; $\alpha = \beta = 90^\circ$, $\gamma = 120^\circ$	
Rhombohedral 1 $a = b = c$; $\alpha = \beta = \gamma$			

Several permutations are possible within groups and leads to descriptions of primitive unit cells, face-centred containing an additional point in the center of each face, body-centred contains an additional point in the centre of the cell and centred, with an additional point in the centre of each end.

molecule -it is an electron density map. However, since electrons are tightly localized around the nucleus of atoms the electron density map is a good approximation of atomic positions within a molecule.

Diffraction of X-rays from a single molecule perhaps containing only one or a few electron-dense centres would be very difficult to measure and to distinguish from ambient noise. One advantage of a crystal is that there are huge numbers of molecules orientated in an identical direction. The effect of ordering is to enhance the intensity of the scattered signals (reflections). The conditions for in phase scattering can be viewed as the reflection of X-rays from (or off) planes passing through the collections of atoms in a crystal. A consideration of Bragg's law ($n\lambda = 2d \sin\theta$), i.e. the relationship between scattering angle (θ) and the interplanar spacing (d) shows that if the wavelength (λ) is increased the total diffracted intensity becomes less sensitive to the spacing or to changes in angle. The resulting diffraction pattern will be less sensitive and fine detail will be obscured. At a fixed wavelength (the normal or traditional condition) a decrease in planar spacing will require higher angles of diffraction to observe the first peak in the diffracted intensity. This inverse relationship between spacing within the object and the angle of diffraction leads to the diffraction space being called 'reciprocal space'.

Initial applications of X-ray crystallography were limited to small biological molecules or repeating units found in fibrous proteins like collagen. In 1934 J. D. Bernal and Dorothy Hodgkin showed that pepsin diffracted X-rays; observations consistent with the presence of organized and repeating structure. Despite this success structure determination of large proteins seemed a long way away. In 1947 Dorothy Hodgkin solved the structure of vitamin B_{12}, at that time a major experimental achievement, but further application of these methods to proteins proved very difficult. The partial success for smaller molecules relied on a 'trial and error' approach.

Thus at the beginning of the 1950s it seemed that X-ray crystallography was going to falter as a structural technique until Max Perutz demolished this barrier by introducing the method of isomorphous heavy atom replacement. In a typical diffraction pattern the irradiation of a protein crystal with monochromatic X-rays results in the detection of thousands of spots or reflections. These 'spots' are the raw data of crystallography and arise from all atoms within the unit cell. A complete analysis of the diffraction pattern (Figure 10.9) will allow the electron density map, and by implication the position of atoms to be deciphered.

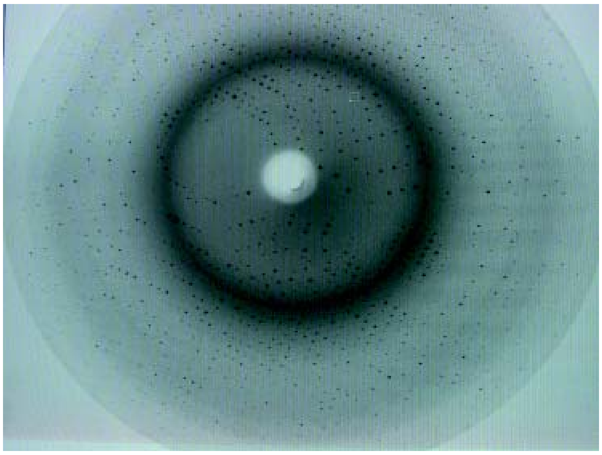


Figure 10.9 Protein diffraction patterns

Since all of the atoms within a unit cell contribute to the observed diffraction pattern it is instructive to consider how the properties of a wave lead to the location of atoms within proteins in the unit cell of crystals. A wave consists of two components – an amplitude f and a phase angle ψ (Figure 10.10). The wave can be further described as a vector ($\mathbf{f}$) of magnitude f and phase angle ψ, and using the relationship between vector algebra and complex numbers the vector becomes the product of real and imaginary components in a complex number plane.

From Figure 10.11

$$\mathbf{f} = f \cos\psi + \mathrm{i} f \sin\psi \qquad (10.9)$$

leads to

$$\mathbf{f} = f(\cos\psi + \mathrm{i}\sin\psi) = f\mathrm{e}^{\mathrm{i}\psi} \qquad (10.10)$$

This minor bit of algebra combined with trigonometry becomes important when one considers that all atoms contribute a scattered wave to the diffraction pattern.

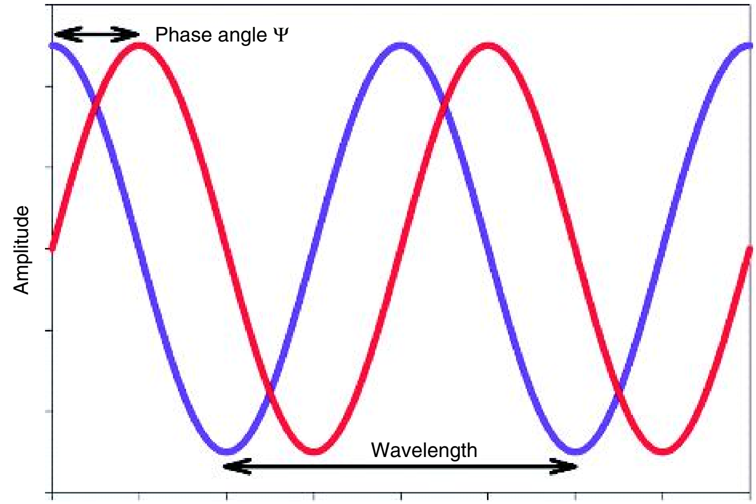


Figure 10.10 A wave described by its amplitude and phase angle. A cosine and sine wave are simply related by a phase shift of 90° or π/2 radians

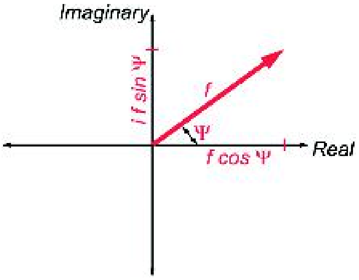


Figure 10.11 Vector representation of a wave and its expression as a complex number

Diffracted waves of intensity (amplitude, f) and phase (ψ) are summed together and described by the vector $\mathbf{F}_{\text{hkl}}$ known as the structure factor. Equation 10.10 becomes

$$\mathbf{F}_{\text{hkl}} = \sum f \cos\psi + \sum \mathrm{i} f \sin\psi \tag{10.11}$$

and leads to

$$\mathbf{F}_{\text{hkl}} = F_{\text{hkl}}(\cos\varphi_{\text{hkl}} + \mathrm{i}\sin\varphi_{\text{hkl}}) = F_{\text{hkl}}e^{\mathrm{i}\varphi_{\text{hkl}}} \tag{10.12}$$

F_{hkl} is the square root of the intensity of the observed (measured) diffraction spot often called I_{hkl}, whilst the φ_{hkl} term represents the summation of all phase terms contributing to this spot.

The next problem is to relate the structure factor $\mathbf{F}_{\text{hkl}}$ to the three-dimensional distribution of electrons in the crystal – the distribution of atoms. The structure factor is the Fourier transform of the electron density and is a vector defined by intensity F_{hkl} *and* phase φ_{hkl}. The value of the electron density at a real-space lattice point (x, y, z) denoted by $\rho\ (x, y, z)$ is equivalent to

$$\rho(x, y, z) = 1/V \sum_{\text{hkl}=-\infty}^{+\infty} \mathbf{F}_{\text{hkl}} e^{-2\pi \mathrm{i}(hx+ky+lz)} \tag{10.13}$$

where ρ is the value of the electron density at the real-space lattice point (x, y, z) and V is the total volume of the unit cell. This is rearranged using Equation 10.12 to give

$$\rho(x, y, z) = 1/V \sum_{\text{hkl}=-\infty}^{+\infty} F_{\text{hkl}} e^{\mathrm{i}\varphi_{\text{hkl}}} e^{-2\pi \mathrm{i}(hx+ky+lz)} \tag{10.14}$$

where φ_{hkl} is the phase information.

To obtain '3D pictures' of molecules the crystal is rotated while a computer-controlled detector produces two-dimensional electron density maps for each angle of rotation and establishes a third dimension. A rotating Cu target is the source of X-rays and this generator is normally cooled to avoid excessive heating. X-rays pass via a series of slits (monochromators) to the crystal mounted in a goniometer. The crystal is held within a loop by surface tension or attached by glue to a narrow fibre and can be rotated in any direction. Nowadays crystals are flash cooled to ~77 K (liquid nitrogen temperatures) with benefits arising from reduced thermal vibrations leading to lower conformational disorder and better signal/noise ratios. Cryocooling also limits radiation damage of the crystal and it is possible to collect complete data sets from a single specimen. Finally, a detector records the diffraction pattern where each spot represents a reflection and has parameters of position, intensity and phase. For a typical crystal there may be 40 000 reflections to analyse, and although this remains a time-consuming task computers facilitate the process. In general all of the diffraction apparatus is controlled via computer interfaces. The data collected is a series of frames containing crystal diffraction patterns as it rotates in the X-ray beam. From these frames data analysis yields a list of reflections (positions) and their intensities (amplitudes) but what remains unknown are the relative phases of the scattered X-rays.

A useful scheme with which to understand the occurrence of reflections is the Ewald construction. If a sphere of radius 1/λ centred around the crystal is drawn then the origin of the reciprocal lattice lies where the transmitted X-ray beam passes straight through the crystal and is described as at the edge of the Ewald sphere (Figure 10.12). Diffraction spots occur only when the Laue or Bragg equations are satisfied and this condition occurs when the reciprocal lattice point lies exactly on the Ewald sphere. At any one instant the likelihood of observing diffraction is small unless the crystal is rotated to bring more points in the reciprocal lattice to lie on the Ewald sphere by rotation of the crystal (Figure 10.13).

A detection device perpendicular to the incoming beam records diffraction on an arbitrary scale with the most obvious attributes being position and intensity of the 'spots'. The intensity is proportional to the square of the structure factor magnitude according to the relationship

$$I_{\text{hkl}} = k|F_{\text{obs}}|^2 \tag{10.15}$$

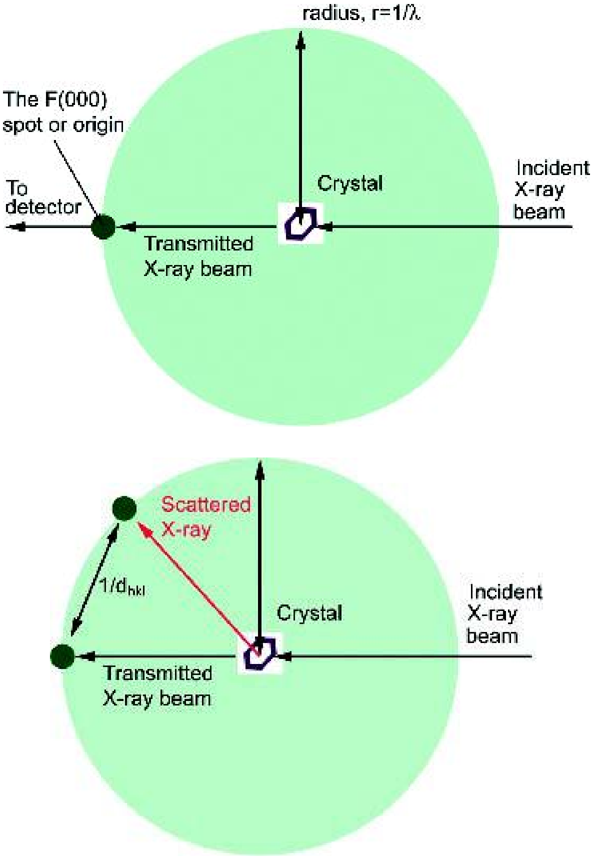


Figure 10.12 Scattering of X-rays by atoms within a crystal

where k is a constant that depends on several factors such as the X-ray beam energy, the crystal volume, the volume of the unit cell, the angular velocity of the crystal rotation, etc.

From Equation 10.14 the relationship between intensity, phase and the fractional coordinates of each atom (x, y and z) is clearly emphasized and from this equation it is clear that if we know the structure we can generate $\mathbf{F}_{\text{hkl}}$. However, in crystallography the aim is to determine the position of atoms, i.e. x, y and z from $\mathbf{F}_{\text{hkl}}$ – the inverse problem. A complete description of each reflection – its position, intensity *and* phase – is represented by the structure factor, $\mathbf{F}_{\text{hkl}}$ and if this

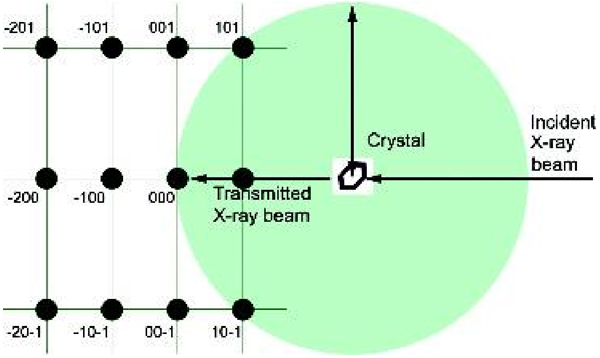


Figure 10.13 Rotation of the crystal brings more planes (collections of atoms) into the Ewald plane

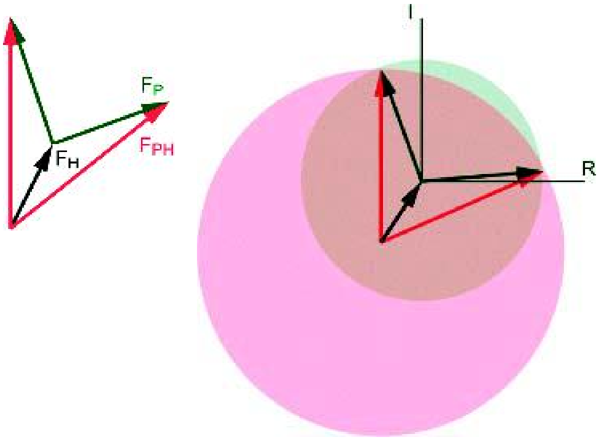


Figure 10.14 A vector diagram showing intensity and phases attributed to protein, heavy atom and derivative. The Harker construction emphasizes two possible phases for F_P. Diffraction data from a second derivative identify a unique solution. The diagram is constructed by drawing a circle with a radius equal to the amplitude of F_P and centred at the origin (shown by brown/green shading). The circle indicates a vector obtained for all of the possible phase angles for F_P. A second circle with radius F_{PH} centred at a point defined by F_H (pink in the figure above). Where the two circles intersect represents possible values for F_P (magnitude and phase) that satisfy the equation $F_{PH} = F_H + F_P$ while agreeing with the measured amplitude F_{PH}

parameter is determined accurately enough then sufficient information exists to generate an atomic structure for proteins.

The phase problem

When crystallographers worked on the structures of simple molecules it was possible to make 'guesses' about the conformation of a molecule. By calculating a diffraction pattern for the 'guess' the theoretical and experimentally determined profiles were compared. If the guess placed atoms in approximately the right position then the calculated phases were almost correct and a useful electron density map could be computed by combining the observed amplitudes with the calculated phases. In this empirical fashion it was possible to successively refine the model until a satisfactory structure was obtained. These direct methods work well for small molecule crystal structures but not proteins.

The solution to the phase problem determines the value of φ_{hkl} and overcomes a major bottleneck in the determination of protein structure by diffraction methods. The achievement of Max Perutz in solving the 'phase problem' in initial studies of myoglobin was significant because it pointed the way towards a generalized method for macromolecular structure determination. Perutz irradiated crystals of myoglobin soaked in the presence of different heavy metal ions. Isomorphous replacement required that metal ions were incorporated into a crystal without perturbing structure and is sometimes difficult to achieve. In 1954 Perutz and co-workers calculated a difference Patterson $(F_{PH} - F_P)^2$ using the amplitudes of a mercury 'labelled' haemoglobin crystal and the amplitudes of a native, but isomorphous, haemoglobin crystal. The scattering due to the 'light' atoms (from the protein) was mathematically removed leaving a low level of background noise with the residual peaks on the difference Patterson map showing the vectors existing between heavy atoms. These maps define the positions of the heavy atoms and allow structure factors to be calculated. This assumes that scattering from protein atoms is unchanged by complex formation with heavy atoms. With the proviso that the heavy atom does not alter the protein then the structure factor for the derivative crystal (F_{PH}) is equal to the sum of the protein structure factor (F_P) and the heavy atom structure factor (F_H), or

$$\mathbf{F}_{PH} = \mathbf{F}_P + \mathbf{F}_H \tag{10.16}$$

The structure factor is a vector and leads to a representation called the Harker construction (Figure 10.14). Since the length and orientation of one side ($\mathbf{F}_H$) is known along with the magnitude of $\mathbf{F}_{PH}$ and $\mathbf{F}_P$ there are two possible solutions for the phase of $\mathbf{F}_P$.

Heavy metal atom derivatives must give minimal perturbations of protein structure and retain the lattice structure of the unmodified crystal. The derivatives must also perturb reflection intensities sufficiently to allow calculation of phases. Protein crystals are usually prepared with heavy metal atoms such as uranium, platinum or mercury introduced at specific points within the crystal with thiol groups representing common high affinity sites. Amongst the compounds that have been used are potassium tetrachloroplatinate(II), *p*-chloromercuribenzoate, potassium tetranitroplatinate(II), uranyl acetate and *cis*-platinum (II) diamine dichloride.

Modern diffraction systems are highly specialized exploiting advances in computing, material science, semiconductor technology and quantum physics to allow the generation of accurate and precise protein structures. All of the originally tedious calculations are performed computationally whilst all equipment is controlled via microprocessors. A further enhancement of X-ray diffraction has been the use of highly intense or focused beams such as those obtained from synchrotron sources. This has allowed new solutions to the phase problem involving the use of synchrotron radiation at multiple wavelengths. Today one of the methods of choice in X-ray crystallography is called multiple anomalous dispersion (MAD). Just as the use of heavy metal derivatives allowed the extrapolation of phase information and permitted the determination of electron density from observed diffraction data the use of multiple wavelengths near the absorption edge of a heavy atom achieves a comparable effect. This is commonly performed by replacing methionine with selenomethionine during protein expression. Nowadays crystallography centres are based around the location of synchrotron sources and the speed and quality of structure generation has improved dramatically (Figure 10.15).

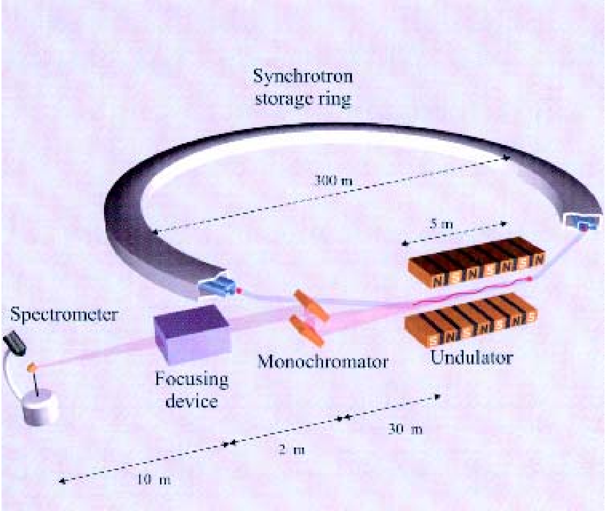


Figure 10.15 The arrangement of data collection centres about a synchrotron storage ring (reproduced with permission from Als-Nielsen, J. & McMorrow, D. *Elements of Modern X-ray Physics*. John Wiley & Sons)

Fitting, refinement and validation of crystal structures

The initial electron density map (Figure 10.16) does not resolve individual atoms and in the early stages several 'structures' are compatible with the data. Higher resolution allows the structure to be assigned. Interpretation of electron density maps requires knowledge of the primary sequence since the arrangement of heavy atoms for the side chains dictates the fitting process. Resolution is not a constant even between similarly sized proteins because crystal lattices have motion from thermal fluctuations and mobility contributes to the final overall resolution. One estimate of mobility within a crystal is the B factor (Debye–Waller factor) that reflects spreading or blurring of electron density. The B factor represents the mean square displacement of atoms and has units of Å^2.

Structures before refinement are often at a resolution >4.5 Å where only α helices are observed and the identification of side chains is unlikely. At higher resolutions of 2.5–3.5 Å the polypeptide backbone is

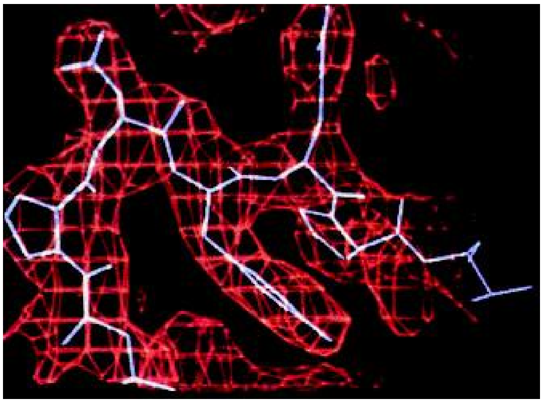


Figure 10.16 An electron density map

traced via electron density located primarily on carbonyl groups and helices, strands and aromatic side chains such as tryptophan are defined. Around 2.0 Å almost all of the structure will be identified including the conformations associated with side chains. Specialized computer programs fit electron density maps and the process is assisted by assuming standard bond lengths and angles. Refining models in an iterative fashion progressively improves the agreement with experimental data. A structure is judged by the crystallographic R-factor, defined as the average fractional error in the sum of the differences between calculated structural factors (F_{cal}) and observed structural factors (F_{obs}) divided by the sum of the observed structural factors.

$$\mathrm{R} = \sum |F_{\mathrm{obs}} - F_{\mathrm{cal}}| / \sum F_{\mathrm{obs}} \qquad (10.17)$$

A value of 0.20 is often represented as an R factor of 20 percent and 'good' structures have R-factors ranging from 15 to 25 percent or approximately 1/10th the resolution of the data. A structure of resolution 1.9 Å is expected to yield an R factor of <0.19. One result of protein structure determination is the generation of a file that lists x, y, and z coordinates for all heavy atoms whose locations are known. These files are deposited in protein databanks and the PDB files listed throughout this book represent the culmination of this analysis.

Protein crystallization

One of the slowest steps in protein crystallography is the production of protein crystals. The methods employed in crystal production rely on the ordered precipitation of proteins. The first protein (urease) was crystallized by James Sumner as long ago as 1926, and was followed by the crystallization of pepsin in 1930 by John Northrop. In the next 20 years over 40 additional proteins were crystallized, including lyzosyme, trypsin, chymotrypsin, catalase, papain, ficin, enolase, carbonic anhydrase, carboxypeptidase, hexokinase and ribonuclease, although structure determination had to wait until the advances of Perutz and Kendrew (Figure 10.17).

Crystallization requires the ordered formation of large (dimensions greater than 0.1 mm along each axis), stable crystals with sufficient long-range order to diffract X-rays. Structures produced by X-ray diffraction are only as good as the crystals from which they are derived.

To form a crystal protein molecules assemble into a periodic lattice from super-saturated solutions. This involves starting with solution of pure protein at a concentration between 0.5 and 200 mg ml^{-1} and adding reagents that reduce protein solubility close to the point of precipitation. These reagents perturb protein–solvent interactions so that the equilibrium shifts in favour of protein–protein association. Further concentration of the solution results in the formation of

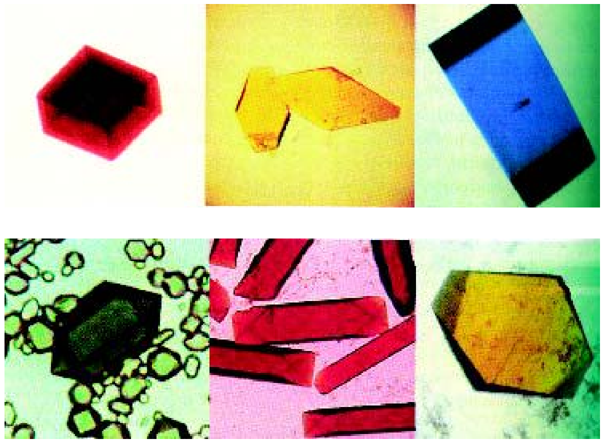


Figure 10.17 Examples of protein crystals. Top row: azurin from *Pseudomonas aeruginosa*, flavodoxin from *Desulfovibrio vulgaris*, rubredoxin from *Clostridium pasteurianum*. Bottom row: azidomethemerythrin from the marine worm *Siphonosoma funafatti*, lamprey haemoglobin and bacteriochlorophyll a protein from *Prosthecochloris aesturii*. The beauty of the above crystals is their colour arising from the presence of light absorbing co-factors such as metal, heme, flavin or chlorophyll (reproduced with permission from Voet, D., Voet, J.G. & Pratt, C.W. *Fundamentals of Biochemistry* John Wiley & Sons Ltd. Chichester, 1999)

nucleation sites, a process critical to crystal formation and is the first of three basic stages of crystallization common to all systems. It is followed by expansion and then cessation as the crystal reaches a limiting size.

Two experimental methods used to form crystals from protein solutions are vapour diffusion and equilibrium dialysis. Vapour diffusion is the standard method used for protein crystallization. It is suitable for use with small volumes, easy to set-up and to monitor reaction progress. A common format involves 'hanging drops' containing protein solution plus 'precipitant' at a concentration insufficient to precipitate the protein (Figure 10.18). The drop is equilibrated against a larger reservoir of solution containing precipitant and after sealing the chamber equilibration leads to supersaturating concentrations that induce protein crystallization in the drop.

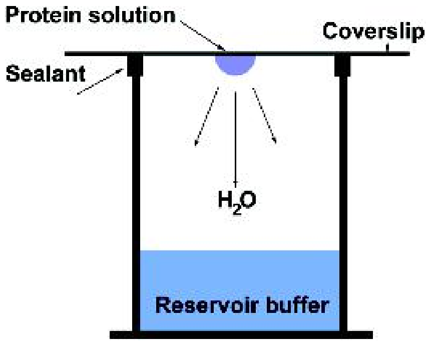


Figure 10.18 The 'hanging drop' or vapour diffusion method of protein crystallization. As little as 5 μl of concentrated solution (protein + solvent) may be suspended on the coverslip

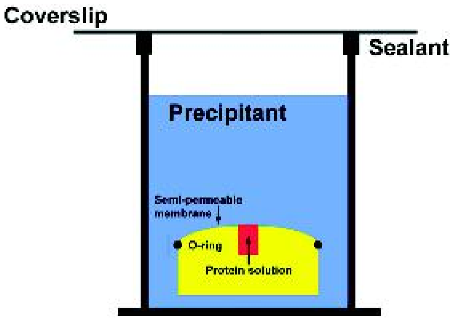


Figure 10.19 Equilibrium dialysis can be achieved with many different 'designs' although the basic principle involves the separation of protein solution from the precipitant by a semipermeable membrane. Diffusion across the membrane promotes ordered crystallization

A second method is equilibrium dialysis and is used for crystallization of proteins at low and high ionic strengths (Figure 10.19). Small volumes of protein solution are placed in a container separated from precipitant by a semi-permeable membrane. Slowly the precipitant causes crystal formation within the well containing the protein solution.

To perform large numbers of crystallization trials it is common to use robotic systems to automate the process and to eliminate labour-intensive manipulations whilst crystallization is performed in temperature-controlled rooms, free of vibration, leading to crystals appearing over a period of 4–10 days. When the above techniques work successfully X-ray diffraction yields unparalleled resolution for protein structures. There is little doubt that the technique will continue to represent the main method of protein structure determination in the exploration of new proteomes.

Nuclear magnetic resonance spectroscopy

NMR spectroscopy as a tool for determining protein structure arose more recently than X-ray crystallography. With the origins of X-ray diffraction lying in the discoveries of Röentgen, Laue and the Braggs at the beginning of the 20th century it was not until 1945 that a description of the NMR phenomenon was given by Felix Bloch and Edward Purcell. NMR spectra are observed upon absorption of a photon of energy and the transition of nuclear spins from ground to excited states. The observation of signal associated with these transitions and the use of radio frequency irradiation to elicit the response marked the start of NMR spectroscopy.

This discovery would have remained insignificant except for subsequent observations showing that nuclear transitions differed in frequency from one nucleus to another but also showed subtle differences according to the nature of the chemical group. So for the proton this property meant that proteins exhibited many signals with, for example, the methyl protons resonating at different frequencies to amide protons which in turn are different to the protons attached to the α or β carbons. In 1957 the first NMR spectrum of a protein (ribonuclease) was recorded but progress as a structural technique remained slow until Richard Ernst described the use of transient techniques. Transient signals produced after a pulse of radio frequency radiation are converted into a normal spectrum by the mathematical process of Fourier transformation – this technique would pave the way towards important advances, particularly multi-dimensional NMR spectroscopy, that are now the basis of all biomolecular structure determination.

Table 10.3 Spin properties and abundance of important nuclei in protein NMR studies

Nuclei	Spin	Abundance (% total)	Magnetogyric ratio (γ) ($\times 10^7$ $T^{-1}s^{-1}$)	Ratio (γ_n/γ_H)
^{1}H	1/2	99.985	26.7520	1.00
^{2}H	1	0.015	4.1066	0.15
^{13}C	1/2	1.108	6.7283	0.25
^{14}N	1	99.630	1.9338	0.07
^{15}N	1/2	0.370	2.712	0.10
^{18}O	5/2	0.037	3.6279	0.14
^{31}P	1/2	100.000	10.841	0.41

NMR phenomena

Underlying the NMR phenomenon is a property of all atomic nuclei called 'spin'. Spin describes the nature of a magnetic field surrounding a nucleus and is characterized by a spin number, I, which is either zero or a multiple of 1/2 (e.g. 1/2, 2/2, 3/2, etc.). Nuclei whose spin number equals zero have no magnetic field and from a NMR standpoint are uninteresting. This occurs when the number of neutrons and the number of protons are even. Nuclei with non-zero spin numbers have magnetic fields that vary considerably in complexity (Table 10.3). Spin 1/2 nuclei represent the simplest situation and arise when the number of neutrons plus the number of protons is an odd number.[2] The most important spin 1/2 nucleus is the proton with a high natural abundance (~100 %) and its occurrence in all biomolecules.

For nuclei such as ^{12}C the most common isotope is NMR 'silent' and the 'active' spin 1/2 nucleus (^{13}C) has a low natural abundance of ~1.1 percent. Similarly ^{15}N has a natural abundance of ~0.37 percent. However, advances in molecular biological techniques enable proteins to be expressed in host cells grown on media containing labelled substrates. Growing bacteria, yeast or other cell cultures on substrates containing ^{15}N or ^{13}C (ammonium sulfate and glucose are common sources) allows uniform enrichment of proteins in

[2] A third possibility exists when the number of neutrons and the number of protons are both odd and this leads to the nucleus having an integer spin (i.e. I = 1, 2, 3, etc.).

these nuclei. Biomolecular NMR spectroscopy requires proteins enriched with ^{13}C or ^{15}N or ideally both nuclei.

Although the shape of the magnetic field of a nucleus is described by the parameter I (spin number) the magnitude is determined by the magnetogyric ratio (γ). A nucleus with a large γ has a stronger magnetic field than a nucleus with a small γ. The proton (^{1}H) has the largest magnetogyric ratio and coupled with its high natural abundance this contributes to its popularity as a common form of NMR spectroscopy.

When samples are placed in magnetic fields nuclear spins become polarized in the direction of the field resulting in a net longitudinal magnetization. In the normal or macroscopic world two bar magnets can be aligned in an infinite number of arrangements but at an atomic level these alignments are governed by quantum mechanics and the number of possible orientations is $2I + 1$. For spin 1/2 nuclei this gives two orientations that in the absence of an external magnetic field are of equal energy. For spin 1/2 nuclei such as ^{1}H, ^{13}C, or ^{15}N application of a magnetic field removes degeneracy and the energy levels split into parallel and antiparallel orientations (Figure 10.20).

Spins aligned parallel with external magnetic fields are of slightly lower energy than those aligned in an antiparallel orientation. The concept of two energy levels allows one to envisage transitions between lower and higher energy levels analogous to that seen in other forms of spectroscopy. The difference in population ($n_{\text{upper}}/n_{\text{lower}}$) between each level is governed by Boltzmann's distribution

$$n_{\text{upper}}/n_{\text{lower}} = e^{-(\Delta E/k_B T)} \qquad (10.18)$$

When $\Delta E \sim k_B T$, as would occur with two closely spaced energy levels, the ratio $n_{\text{upper}}/n_{\text{lower}}$ approaches 1. At thermal equilibrium the number of nuclei in the lower energy level slightly exceeds those in the higher energy level. As a result of this small inequality it is possible to elicit transitions between states by the application of short, intense, radio frequency pulses.

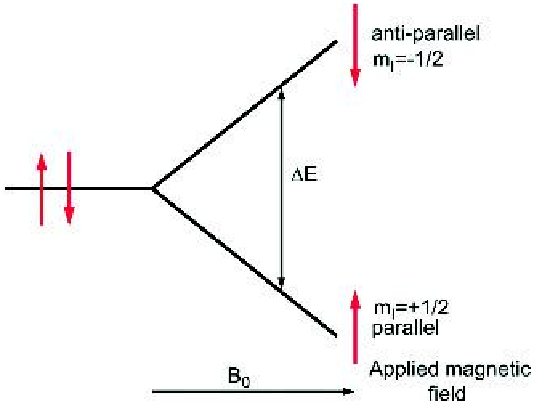


Figure 10.20 An energy level diagram reflecting the alignment of spin 1/2 nuclei in applied magnetic fields

Instead of considering a single spin it is more useful to consider the magnetic ensemble of spins. In the presence of the applied magnetic field (B_0) spin polarization occurs and a vector model of NMR views the net magnetization lying in the direction of the z-axis. Irradiation of the sample by a radiofrequency (rf) field denoted as B_1 along the x-axis rotates the macroscopic magnetization into the xy plane (Figure 10.21).

Most frequently the pulse length is calculated to tip 'magnetization' through 90° ($\pi/2$ radians). The xy plane lies perpendicular to the magnetic field and causes transverse magnetization to precess under the influence of the applied magnetic field. Precession in the xy plane at the Larmor frequency of the nuclei under investigation induces a current in the detector coil that is the *observable* signal in all NMR experiments.

Transverse magnetization decays exponentially with time in the form of a signal called a free induction decay (FID) eventually reaching zero. All NMR spectra (such as the one shown in Figure 10.22) are derived by converting the FID signal of intensity versus time into a profile of intensity versus frequency via Fourier transformation (FT).

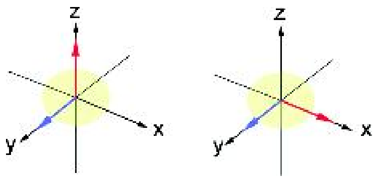


Figure 10.21 Application of a radiofrequency pulse (blue arrow) along the y-axis rotates macroscopic magnetization (red arrow) into the xy plane

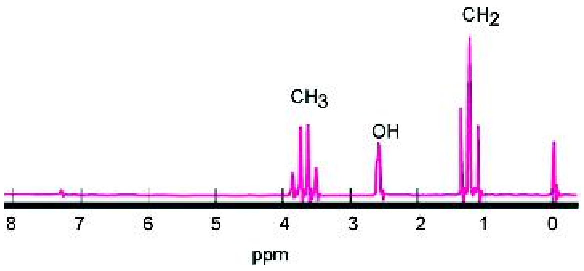


Figure 10.22 A simple NMR spectrum for the molecule ethanol

Simple NMR experiments involve repetitive application of rf pulses with a suitable recovery period between pulses allowing the return of magnetization to equilibrium. Each FID is acquired, stored on computer and added to the previous FID permitting signal averaging. In this manner high signal to noise ratio spectra are acquired on protein samples in a few minutes. Unlike most forms of spectroscopy where the incident radiation is slowly scanned through the spectral range FT-NMR involves the application of rf pulses that excite all nuclear spins simultaneously.

In the construction of magnets for use in NMR spectroscopy the magnetic field is achieved through the use of superconducting materials operating at liquid helium temperatures. Magnets are described in terms of their field strength with designations such as 14.1 T (where 1 T or Tesla is equivalent to 10^4 Gauss and 1 G is equivalent to the earth's magnetic field). The Larmor frequency is obtained from the relationship $\omega_o = -\gamma B_0$ where $\omega_o = 2\pi\nu_0$ and ν_0 is the Larmor frequency. A quick calculation of the Larmor frequency of the proton (where B_0, the field strength = 14.1 T; γ, the magnetogyric ratio of $^1H = 26.7520 \times 10^7\ T\,s^{-1}$) shows that ν_0 (^1H) occurs at a frequency of ~600 MHz. As a result spectrometers are often referred to as 400, 600, 750, 800 or 900s reflecting the proton Larmor frequency in MHz at a given magnetic field strength.

Parameters governing NMR signals

The use of NMR spectroscopy as a tool to determine protein structure is based around several related parameters that influence the observation of signals. These parameters include the chemical shift (δ), spin-spin coupling constants (J), the spin lattice or T_1 relaxation time (sometimes denoted as R_1 the spin lattice relaxation rate = $1/T_1$), the spin–spin or T_2 relaxation time ($R_2 = 1/T_2$), the peak intensity and the nuclear Overhauser effect (NOE). Since these effects are vital to all aspects of NMR spectroscopy a brief description is given to highlight their respective importance.

The peak intensity represented formally by the integrated area reflects the number of nuclei involved in the signal. The ^1H NMR spectrum of ethanol illustrates this by containing three resonances due to the hydroxyl group, the methylene group and the methyl group. The integrated areas under each line are in the ratio of 1:2:3. Within proteins methyl resonances exhibit intensities three times greater than the Hα proton. Although the peak height is frequently used as an indicator of intensity many parameters can modify peak height so the integrated area is always the best indicator of the number of protons forming each peak.

The chemical shift denotes the position of a resonance along a frequency axis and is uniquely sensitive to the environment in which a nucleus is located. The chemical shift is defined relative to a standard such as 2,2-dimethyl-2-silapentane-5-sulfonate (DSS) and is quoted in p.p.m. units. Since resonant frequencies are directly proportional to the static field minor variations between instruments make it very difficult to compare spectra obtained on different spectrometers. To avoid this problem all resonances are measured relative to a standard defined as having a chemical shift of 0 p.p.m. For a resonance the chemical shift (δ) is measured as

$$\delta = (\Omega - \Omega_{ref})/\omega_0 \times 10^6 \qquad (10.19)$$

where Ω and Ω_{ref} are the offset frequencies of the signals of interest and reference respectively. The resonance attributable to the protons of water occurs at ~4.7 p.p.m. at ~20 °C.

The fundamental equation of NMR, $\omega = -\gamma B_o$, suggests that all nuclei subjected to the same magnetic field will resonate at the same frequency. This does not occur – nuclei experience different fields due to their local magnetic environment. This allows the equation to be recast as

$$\omega = \gamma(B_o - \sigma B_o) \qquad (10.20)$$

where σ represents a screening or shielding constant that reflects the different magnetic environments found in molecules. Another way of envisaging this effect is to use an effective field at the nucleus B_{eff} that is reduced by an amount proportional to the degree of shielding.

$$B_{\text{eff}} = B_0(1 - \sigma) \qquad (10.21)$$

Spin–spin coupling constants or J values are defined by interactions occurring through bonds. These scalar interactions are field independent and occur as a result of covalent bonds between nuclei linked by three or less bonds. The scalar coupling constants contribute to the fine structure observed for resonances in ^{1}H NMR spectra of small molecules although they are rarely resolved in studies of proteins. Scalar coupling leads to the splitting of the methylene signal of ethanol into a quartet as a result of interactions with each of the three protons of the methyl group. Similarly the methyl proton resonance is seen as a triplet due to its interactions with each methylene proton. As the molecular weight increases this fine structure is frequently obscured by line broadening effects.

A correlation exists between the magnitude of the spin–spin coupling constant (^3J) and the torsion angles found for vicinal protons in groups of the polypeptide chain. The Karplus equation derived from theoretical studies of ^{1}H–C–C–^{1}H couplings suggests a relationship of the form

$$^3J = A\cos^2\theta + B\cos\theta + C \qquad (10.22)$$

and in proteins the $^3J_{\text{NH}\alpha}$ coupling constant and torsion angle φ are related by

$$^3J_{\text{NH}\alpha} = 6.51\cos^2\theta - 1.76\cos\theta + 1.60$$
$$(\text{where } \theta = \phi - 60) \qquad (10.23)$$

If the coupling constant $^3\text{J}_{\text{NH}\alpha}$ is measured with sufficient accuracy the torsion angle can be estimated and used as a structural restraint. Many coupling constants between nuclei are measured using heteronuclear and homonuclear methods (Table 10.4).

The longitudinal relaxation time (T_1) reflects the rate at which magnetization returns to the longitudinal axis after a pulse and has the units of seconds. If M_{o} is the magnetization at thermal equilibrium then at time t after a pulse the recovery of M_{o} is expressed as

$$M_z = M_{\text{o}}(1 - e^{-t/T_1}) \qquad (10.24)$$

where T_1 is the longitudinal relaxation time. In solution T_1 is correlated with the overall rate of tumbling of a macromolecule but is also modulated by internal molecular motion arising from conformational flexibility. Both T_1 and T_2 depend on the correlation time τ_c a factor closely linked with molecular mass, and estimated assuming a spherical protein of hydrodynamic radius r in a solution of viscosity, η, via the Stokes equation as

$$\tau_c = 4\pi\eta r^3/3k_BT \qquad (10.25)$$

Macromolecules such as proteins have longer correlation times than small peptides. A typical value for a 100 residue spherical protein is 3–5 ns.

The transverse or spin–spin relaxation time (T_2) describes the decay rate of transverse magnetization in the xy plane. T_2 is always shorter than T_1 and is correlated with dynamic processes occurring within a protein. T_2 governs resonance linewidth and decreases with increasing molecular mass. For

Table 10.4 One, two and three bond coupling constants

Coupling	Magnitude (Hz)	Coupling	Magnitude (Hz)
^{1}H–^{13}C (^1J)	110–130	^{13}Cα –^{13}CO (^1J)	55
^{1}H–^{15}N (^1J)	89–95	^{13}Cα –Hα and ^{13}Cβ –Hβ (^1J)	130–140
H–C–C–H vicinal (^3J)	2–14	HN–HA (^3J)	2–12
H–C–H geminal (^3J)	−12–−15	^{13}Cα –^{15}N (^1J)	45
^{15}N–^{13}CO	15	^{13}Cα –^{15}N (^2J)	70

Lorentzian lineshapes the linewidth at half maximum amplitude is

$$\Delta\nu_{1/2} = 1/\pi T_2 \qquad (10.26)$$

with decreases in T_2 leading to broader lines. NMR spectroscopy of large proteins is often described as 'limited by the T_2 problem'. Many factors influence T_2 and the most important are molecular mass, temperature, solvent viscosity, and exchange processes.

Probably the most important measurable parameter in NMR experiments for the determination of protein structure is the nuclear Overhauser effect (NOE). This is the fractional change in intensity of one resonance as a result of irradiation of another resonance. As a result of dipolar or 'through space' interactions the irradiation of one resonance perturbs intensities of neighbouring resonances. The NOE is expressed as

$$\eta = (I - I_o)/I_o \qquad (10.27)$$

where I_o is the intensity without irradiation and I is the intensity with irradiation. The NOE effect is rapidly attenuated by distance and declines as the inverse sixth power of the distance between two nuclei.

$$\eta \propto r^{-6} \qquad (10.28)$$

The NOE phenomenon, like the relaxation times T_1 and T_2, varies as a function of the product of the Larmor frequency and the rotational correlation time. Armed with an appreciation of these parameters it is possible to extract much information on the structure and dynamics of regions of the polypeptide chain in proteins.

Practical biomolecular NMR spectroscopy

NMR signals are obtained by placing samples into strong, yet highly homogeneous, magnetic fields. These magnetic fields arise from the use of superconducting materials (niobium–tin and niobium–titanium alloy wires) wound around a 'drum' maintained at temperatures of ~4 K. High field strengths (11.7–21.5 T) arise from the influence of current flowing through the wires and the constant temperature ensures the field strength does not fluctuate. Samples contained within a quartz tube are lowered via a cushion of compressed air into the centre (bore) of the magnet where a 'probe' is maintained at a consistent temperature (±0.1 °C), usually between 5 and 40 °C. Electronics located in the probe allow excitation of nuclei and detection of signals and are linked to computer-controlled devices that permit the generation of rf pulses of defined timing, phase, amplitude and duration whilst allowing the detected signal (FID) to be amplified, filtered and subjected to further processing (Figure 10.23). This processing includes storage of the raw data as well as Fourier transformation.

Experiments may continue for 3–4 days and require protein stability for this period. To observe magnetization involving exchangeable protons such as amides (NH) in proteins it is necessary to perform experiments in water. This brings additional problems of intense signals due to the high concentration of water protons (110 M) that far exceeds the concentration of 'signals' from the protein (~10^{-3} M). Sophisticated methods of solvent (water) suppression allied to post-acquisition processing eliminate these signals effectively from protein spectra.

Chemical shifts reflect the magnetic microenvironments of groups. The amide (NH) proton of a polypeptide backbone has a chemical shift between 8.0 and 9.0 p.p.m., methyl groups have chemical shifts between 0 and 2 p.p.m. whilst the Hα proton has a value between 4.0 and 4.6 p.p.m. ^{1}H chemical shifts were derived from an analysis of short unstructured model peptides of the form Gly-Gly-X-Ala, where X was each of the 20 residues (Table 10.5).

Within *folded* proteins some chemical shifts deviate from their expected values reflecting magnetic environments that depend on conformation. This is seen in the ^{1}H NMR spectrum of ubiquitin where peaks above 9.0 and below 0 p.p.m. reflect tertiary structure although many resonances are found close to their 'random coil' conformations. The dispersion over a much wider frequency range offers the possibility of identifying resonances in a protein – a process called assignment.

The assignment problem in NMR spectroscopy

The assignment problem for a protein of ~100 residues and perhaps 700 protons requires identifying which resonance belongs to a particular proton. The assignment problem remains the crux of determining protein

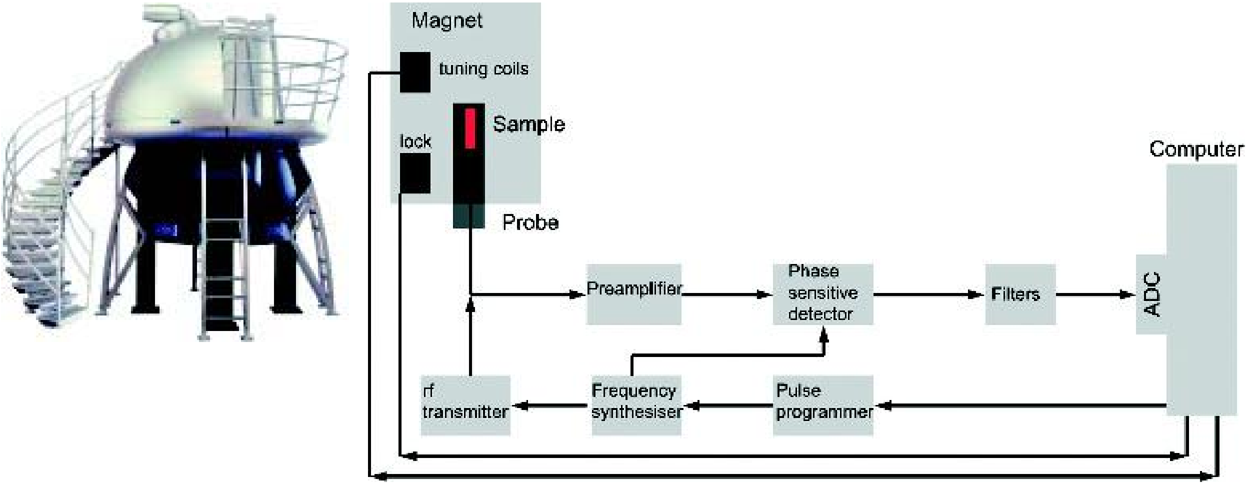


Figure 10.23 A state-of-the-art NMR spectrometer operating at 900 MHz and a block diagram representing a NMR spectrometer. The major components are the magnet containing the probe together with 'shim' or tuning coils to maintain homogeneity and electronics to 'lock' the field at a given frequency. Outside of the magnet are radio frequency transmitters with pulses designed via a pulse programmer and a frequency synthesizer together with a detection system of numerous amplifiers, filters and analogue–digital converters. All systems are controlled via a computer (Reproduced courtesy Varian Inc)

structure by NMR spectroscopy and is perhaps analogous to the phase problem of X-ray crystallography. Without first 'assigning' the protein we cannot move on to derive restraints important in the derivation of molecular structure. The starting point involves assigning a signal to a specific atom or group of atoms (e.g. HA proton, CH_3 protons, etc.). Spectral overlap in 1D spectra usually prevents complete assignment even in small proteins and as shown in the ^1H NMR spectra of ubiquitin (Figure 10.24) – a protein with 76 residues – many overlapping peaks preclude absolute identification.

Historically the major advance in this area was to spread conventional 1D spectra into a second dimension that reduces spectral overlap. In the last two decades these methods have expanded dramatically and almost all structural investigations of proteins start with the acquisition of 2D spectra. A normal 1D NMR spectrum involves a pulse followed by measurement of the resulting FID. 2D NMR spectroscopy involves the application of successive pulses that lead to magnetization transfer between nuclei. The mechanism of transfer proceeds either by a through bond (scalar mechanism) or by a through space (dipolar) interaction. A 2D experiment consists of four time periods; the preparation, evolution, mixing and acquisition periods often preceeded by a *relaxation* delay. The preparation period consists of a single pulse or a series of pulses and delays and its purpose is to create the magnetization terms that will make up the indirect dimension. These terms evolve during the *evolution* or t_1 period. The evolution of magnetization during the t_1 period is followed by a second pulse that initiates the mixing period and results in magnetization transfer to other spins, normally via through bond or through space interactions; transfer can also occur via chemical exchange. The decay of the FID is detected during the *acquisition* or t_2 period. The whole pulse scheme is repeated to allow for signal averaging and other instrumental factors including relaxation. The experiment is then repeated by incrementing the t_1 period to build up a series of FIDs recorded at different t_1 intervals. A 2D experiment contains two time variables, t_1 and t_2, and Fourier transformation of this dataset, $S(t_1, t_2)$ yields a 2D contour plot, $S(\omega_1, \omega_2)$, where the precession

Table 10.5 The ^1H chemical shifts of the amino acids residues in a random coil conformation

	Chemical shift (p.p.m.)			
Residue	HN	HA	HB	others
Ala	8.25	4.35	1.39	–
Asp	8.41	4.76	2.84, 2.75	–
Asn	8.75	4.75	2.83, 2.75	7.59, 6.91 (sc amide)
Arg	8.27	4.38	1.89, 1.79	1.70 (HG), 3.32 (HD), 7.17, 6.62 (sc NH)
Cys	8.31	4.69	3.28, 2.96	–
Gln	8.41	4.37	2.13, 2.01	2.38 (HG), 6.87,7.59 (sc NH_2)
Glu	8.37	4.29	2.09, 1.97	2.31,2.28 (HG)
Gly	8.39	3.97	–	–
His	8.41	4.63	3.26, 3.20	8.12 (2H), 7.14 (4H)
Ile	8.19	4.23	1.90	1.48, 1.19 (HG=CH_2), 0.95 (HG=CH_3), 0.89 (HD)
Leu	8.42	4.38	1.65	1.64 (HG), 0.94,0.90 (HD)
Lys	8.41	4.36	1.85, 1.76	1.45 (HG), 1.70 (HD), 3.02 (HE), 7.52 (sc NH_3)
Met	8.42	4.52	2.15, 2.01	2.64 (HG), 2.13(HE)
Phe	8.23	4.66	3.22, 2.99	7.30 (2,6H), 7.39 (3,5H), 7.34 (4H)
Pro	–	4.44	2.28, 2.02	2.03 (HG), 3.68,3.65 (HD)
Ser	8.38	4.50	3.88, 3.88	–
Thr	8.24	4.35	4.22	1.23
Trp	8.09	4.70	3.32, 3.19	7.24(2H), 7.65(4H), 7.17(5H), 7.24(6H), 7.50(7H), 10.22 (indole NH)
Tyr	8.18	4.60	3.13, 2.92	7.15 (2,6H), 6.86 (3,5H)
Val	8.44	4.18	2.13	0.97,0.94 (HG)

Adapted from *NMR of Proteins and Nucleic Acids*. Wuthrich, K. (ed.) Wiley Interscience, 1986.

frequencies occurring during the evolution and detection periods determine peak positions (ω_1 and ω_2) in the 2D plot.

In general, three homonuclear ^1H-NMR experiments are used for assignment of proteins. The COSY (correlated spectroscopy) and TOCSY[3] (total correlated spectroscopy) describe magnetic interactions between scalar coupled nuclei – resonances linked via 'through bond' interactions (Figure 10.25). When the COSY experiments are performed on protein dissolved in H_2O, cross peaks are observed in 2D spectra reflecting, for example, connectivity between NH and HA protons.

In ^1H NMR through bond interactions (Figure 10.26) as a result of coherence transfer are limited to two or three bonds and are restricted to *intra*-residue correlations. The four-bond interaction between the $HN_{(i+1)}$,$HA_{(i)}$ is too weak to be observed. Cross peaks can also occur between HA and HB protons via ^3J couplings and it is soon apparent that characteristic patterns of connectivity are observed for different residues (Figure 10.27).

Performing experiments in D_2O (^2H_2O) removes most HN signals since protons exchange for deuterons leading to the loss of the left-hand side of the 2D spectra. These protons are described as labile.

[3] The TOCSY experiment is sometimes called HOHAHA (Homonuclear Hartmann-Hahn) after the discoverers of an original solid state experiment demonstrating cross-polarization between nuclei.

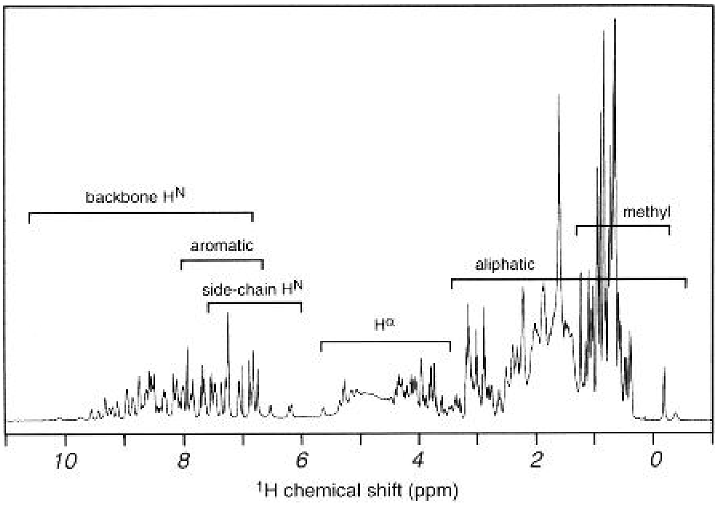


Figure 10.24 The ^1H NMR spectra of ubiquitin, a protein containing 76 residues. (Reproduced with permission from Cavanagh, *J. et al. Protein NMR Spectroscopy: principles and practice*. Academic Press, 1996)

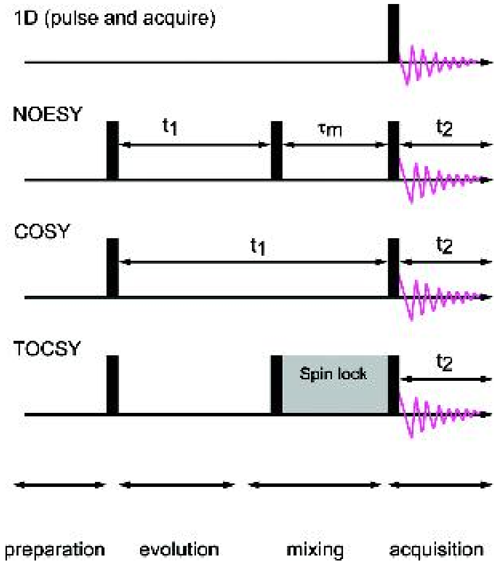


Figure 10.25 Schematic representing 1 and 2D homonuclear pulse sequences used in protein NMR spectroscopy. The pulse schemes have been given names such as COSY (correlation spectroscopy), TOCSY (total correlation spectroscopy) and NOESY (nuclear Overhauser effect spectroscopy) that define the basic mode of magnetization transfer

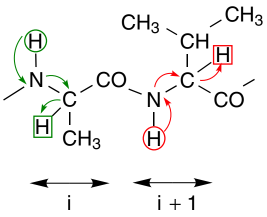


Figure 10.26 Through bond interactions in residue i and i + 1 of a polypeptide chain (... Ala-Val...). The amide proton and HA proton of the same residue are linked via a three-bond coupling

The TOCSY experiment relies on cross-polarization, and during a complicated pulse sequence the application of a 'spin locking' field leads to all spins becoming temporarily equivalent. In a condition of "equivalence" magnetization transfer occurs when the mixing time is equivalent to 1/2 *J*. For proteins containing many residues with different side chains there is no single value for 1/2 *J* that allows all connectivities to be observed. Instead TOCSY experiments are repeated with different mixing times and the ideal situation involves observing complete spin systems from the HN to the HA and HB and on to the remaining resonances.

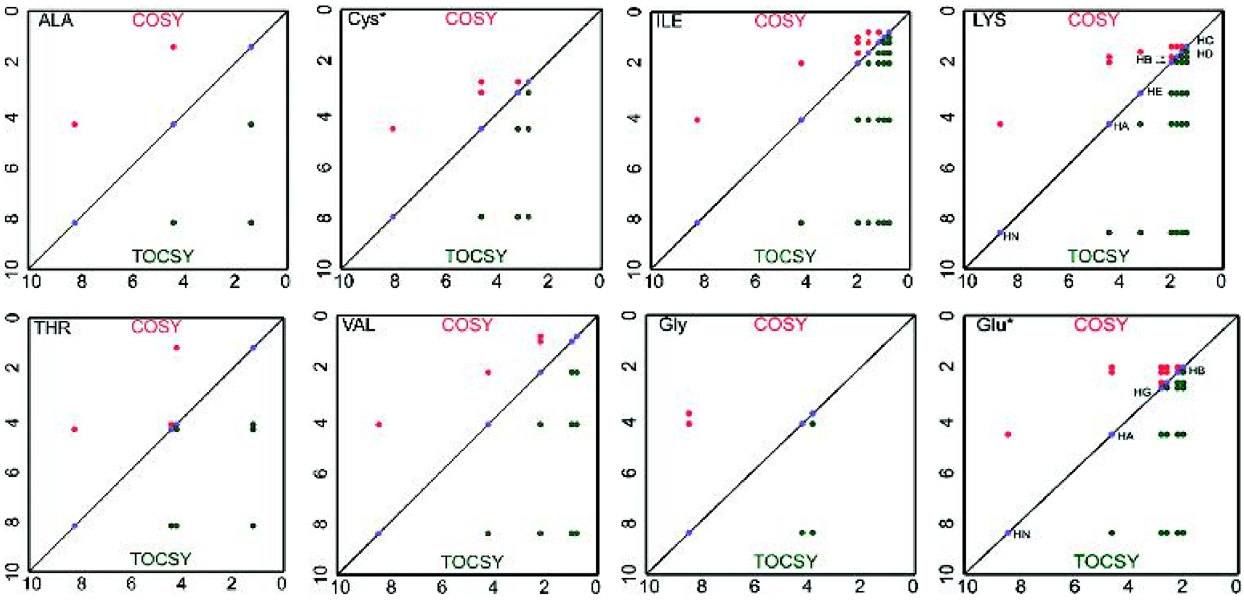


Figure 10.27 Patterns of connectivity seen in 2D COSY and TOCSY spectra of eight of the 20 amino acid residues. *The pattern shown by Cys is similar for all other residues in which the spin system involves a HN–HA and two non-degenerate HB protons. This is likely to include the aromatic residues Phe, Tyr and Trp as well as Asp, Asn and Ser. Similarly, the profile shown by Glu is shared by Gln and Met. The pattern does not take any account of peak fine structure that occurs in COSY type experiments

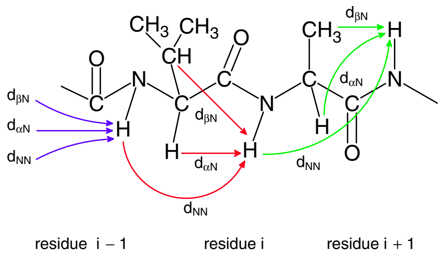


Figure 10.28 The pattern of sequential NOEs expected within a polypeptide chain for three successive residues

Connectivity patterns are sometimes diagnostic for individual residues and having identified spin system type it is necessary to establish its location within the primary sequence (Figure 10.28). This is achieved through NOESY experiments. The 2D homonuclear NOESY experiment relies on *through space* interactions between nuclei (^{1}H) separated by less than 6 Å. By establishing sequential connectivity between resonances of a known spin system type it is possible to identify clusters of residues that are unique within the primary sequence. In Figure 10.28 valine and alanine exhibit sequential connectivity and if these residues occur sequentially only once in the primary sequence then these residues are identified unambiguously. If the dipeptide Val-Ala occurs in more than one region of the polypeptide chain then the identity of residues i – 1 and i + 2 becomes important in establishing conclusive sequence specific assignments.

Some sequential NOEs are indicative of particular units of secondary structure (Table 10.6). For example, the regular periodicity of helices leads to the close approach of HA and HN between residues (i,i + 2), (i,i + 3) and (i,i + 4) that do not occur in antiparallel or parallel β strands. More significantly the observation of $d_{\alpha N}$ (i,i + 4) is strongly indicative of a regular α

Table 10.6 Regular secondary structure gives characteristic NOEs

Interaction	α helix	3_{10} helix	Antiparallel β strand	Parallel β strand	Type I turn	Type II turn
$d_{\alpha N}$	2.7	2.7	2.8	2.8	2.8	2.7
$d_{\alpha\beta}$	2.2–2.9	2.2–2.9	2.2–2.9	2.2–2.9	2.2–2.9	2.2–2.9
$d_{\beta N}$	2.2–3.4	2.2–3.4	2.4–3.7	2.6–3.8	2.2–3.5	2.2–3.4
$d_{\alpha N}(i,i+1)$	3.5	3.4	2.2	2.2	3.4	2.2
$d_{NN}(i,i+1)$	2.8	2.6	4.3	4.2	2.6	4.5
$d_{\beta N}(i,i+1)$	2.5–2.8	2.9–3.0	3.2–4.2	3.7–4.4	2.9–4.1	3.6–4.4
$d_{\alpha N}(i,i+2)$	4.4	3.8	–	–	3.6	3.3
$d_{NN}(i,i+2)$	4.2	4.1	–	–	3.8	4.2
$d_{\alpha N}(i,i+3)$	3.4	3.3	–	–	3.1–4.2	3.8–4.7
$d_{\alpha\beta}$ (i,i+3)	2.5–4.4	3.1–5.1	–	–	–	–
$d_{\alpha N}(i,i+4)$	4.2	–	–	–	–	–

Distances involving β protons vary due to the range of distances possible. The first three sets of distances refer to intra-residue connectivity. All distances are in Å

helix whilst the presence of only $d_{\alpha N}$ (i,i + 2) and $d_{\alpha N}$ (i,i + 3) may indicate more tightly packed 3_{10} helices. Similarly, intense NOEs between $d_{\beta N}$ (i,i + 1) indicates a β strand. Since both strands and helices tend to persist for several residues the observation of blocks of sequential NOEs with these connectivities allows secondary structure to be defined.

The NOESY spectrum not only resolves sequentially connected residues but also represents the basis of protein structure determination using NMR data. Integrating the volumes associated with the cross peaks in NOESY spectra quantifies this interaction and from intra-residue and sequential NOEs allows the volumes to be converted into distances, since regular secondary structure is associated with relatively fixed separation distances (see Table 10.6). Some NOE cross peaks arise as a result of long-range interactions (i.e between residues widely separated in the primary sequence) and these cross peaks play a major role in determining the overall fold of a protein since by definition they represent separation distances of less than 6 Å. Using this approach Kurt Wuthrich pioneered protein structure determination for BUSI IIA – a proteinase inhibitor from bull seminal plasma in 1984 – and it marked a landmark in the progression of NMR spectroscopy from analytical tool to structural technique (Figure 10.29).

Table 10.7 Some of the first proteins whose structures were determined by homonuclear NMR spectroscopy

Protein	Date	Mass
BUSI IIA	1985	6050
Lac repressor headpiece	1985	5500
BPTI	1987	6500
Tendamistat	1986	8000
BDS-I	1989	5000
Human complement protein C3a	1988	8900
Plastocyanin	1991	10 000
Thioredoxin	1990	11 700
Epidermal growth factor	1987	5800

Between 1984 and 1990 structures for many small (<10 kDa) proteins were determined via 2D NMR techniques (Table 10.7). To compare the power of crystallography and NMR spectroscopy Wuthrich and

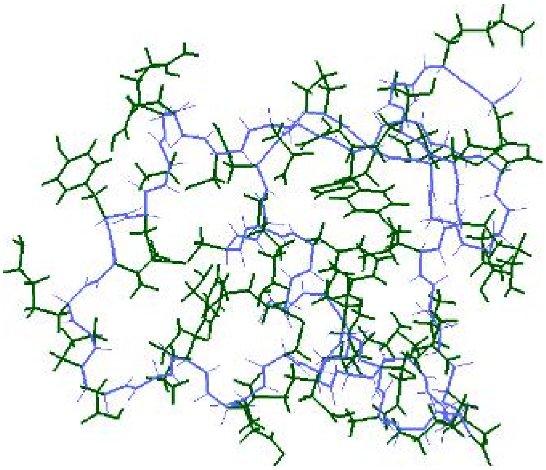


Figure 10.29 The structure of BUSI IIA – the first protein structure determined using NMR spectroscopy. Backbone atoms are shown in blue, side chains in green

Robert Huber independently determined the structure of tendamistat, a protein inhibitor of α-amylase with 74 residues (Figure 10.30). The results showed emphatically that crystal and solution state structures were comparable in almost every respect and testified to the ability of NMR spectroscopy in deriving 3D structures.

However, larger proteins containing more than 100 residues presented new and increasingly difficult problems that arose from the greater number of resonances. This resulted in spectral overlap with increased correlation times producing broader linewidths. The answer was heteronuclear NMR spectroscopy.

Heteronuclear NMR spectroscopy

Resonance assignment is vital to protein structure determination. Additional spin 1/2 nuclei allow alternative assignment strategies and new pulse techniques exploiting heteronuclear scalar couplings present in $^{13}C/^{15}N$ enriched proteins (Figure 10.31). Experiments involved the transfer of magnetization from ^{1}H to ^{13}C and/or ^{15}N through large one-bond scalar couplings. The magnitude of these couplings (30–140 Hz) is much greater than the ^{1}H–^{1}H ^{3}J couplings (2–10 Hz).

The simplest heteronuclear 2D experiments correlate the chemical shift of the ^{15}N nucleus with its attached proton. In 2D ^{15}N–^{1}H heteronuclear spectra cross peaks are spread out according to the nitrogen chemical shift

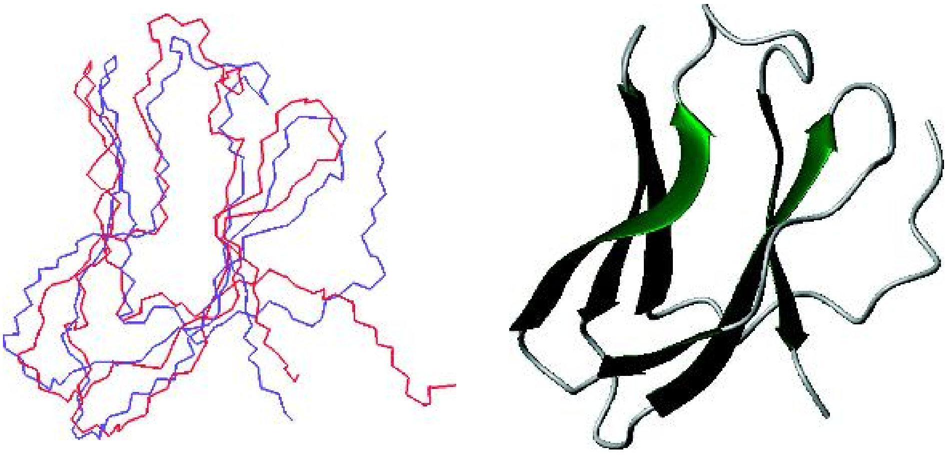


Figure 10.30 The structures of tendamistat derived by NMR and crystallography are superimposed for the polypeptide backbone. The NMR structure is shown in red and the crystallographic structure is shown in blue. (PDB files: 2AIT and 1HOE). With the minor exception of the N terminal region the two structures agree very closely. Tendamistat is a protein based largely on β sheet (right)

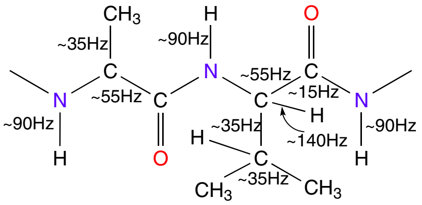


Figure 10.31 Heteronuclear coupling between ^{13}C, ^{1}H and ^{15}N within proteins together with the magnitude of these one bond coupling constants (^{1}J)

Table 10.8 Nitrogen chemical shifts for the backbone and side chain amides of residues found in proteins

Residue	^{15}N	Residue	^{15}N
Ala	125.04	Leu	122.37
Arg	121.22	Lys	121.56
Asn sc	119.02	Met	120.29
Asp	119.07	Phe	120.69
Cys	118.84	Pro	–
Glu	120.23	Ser	115.54
Gly	107.47	Thr	111.9
Gln sc	120.46	Trp indole	122.08
His	118.09	Tyr	120.87
Ile	120.35	Val	119.31

Determined from position in a sequence AcGGXGG-NH$_2$ in 8 M urea. Adapted from *J. Biomol. NMR* 2000, **18**, 43–48.

as well as the proton chemical shift (Figure 10.32). Many residues have characteristic ^{15}N and ^{13}C chemical shifts (Tables 10.8 and 10.9). For example, Gly, Ser and Thr residues show ^{15}N chemical shifts below 115 p.p.m. along with the side chain amides of Gln and Asn. Observation of a cross peak in heteronuclear spectra in these regions is most probably attributed to one of these residues.

Heteronuclear 2D versions of the NOESY, COSY and TOCSY are often recorded but spectral overlap can remain a problem especially in NOE spectra where both intra- and inter-residue cross peaks are observed. This problem was resolved by further pulse sequences that involved three time variables (t_3, t_2 and t_1) and where Fourier transformation leads to a cube or 3D plot. The major advantage of these techniques is that 2D spectra are extended into a third dimension usually the ^{15}N or ^{13}C chemical shift. A panoply of new techniques based on heteronuclear J couplings massively enhanced NMR spectroscopy as a tool for ‘large’ protein structure determination. Ambiguities seen previously were eliminated by new pulse sequences that identified spin systems with high sensitivity. These

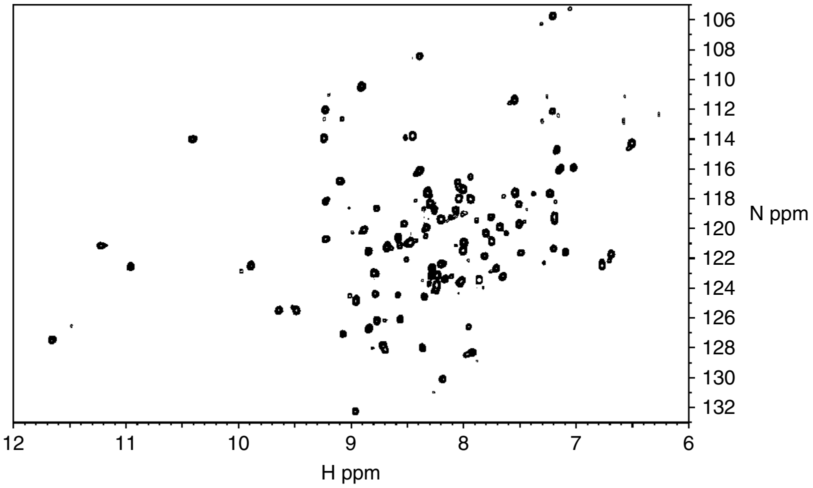


Figure 10.32 A 2D heteronuclear ^{15}N–^{1}H correlation spectra for a protein

Table 10.9 ^{13}C chemical shifts for the backbone and side chain amides of residues found in proteins

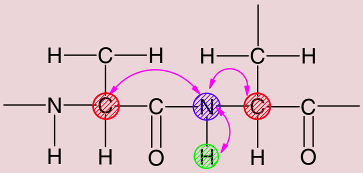
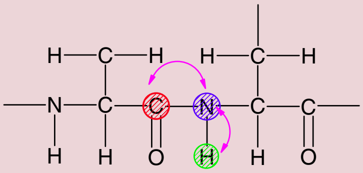
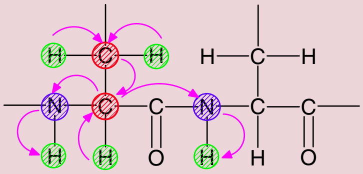
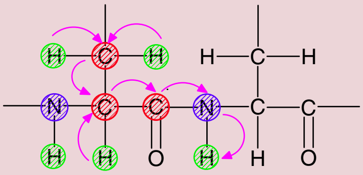
Residue	CO or C′	C_α	C_β	C_γ	C_δ	Other
Ala	175.8	50.8	17.7			
Arg	175.0	54.6	28.8	25.7	41.7	
Asn	173.1	51.5	37.7			175.6 (amide)
Asp	174.2	52.7	39.8			178.4 (carboxylate)
Cys	175.7	57.9	26.0			
Gln	174.0	54.1	28.1	32.2		179.0 (amide)
Glu	174.8	54.9	28.9	34.6		182.8 (carboxylate)
Gly	172.7	43.5				
His	172.6	53.7	28.0			135.2 (C2), 118.7 (C4), 130.3 (C5)
Ile	174.8	59.6	36.9	25.4, 15.7	11.3	
Leu	175.9	53.6	40.5	25.2	23.1, 21.6	
Lys	174.7	54.4	27.5	23.1	31.8	
Met	175.0	53.9	31.0	30.7		15.0 (C_ε)
Phe	176.0	57.4	37.0			136.2 (C1), 130.3 (C2/C6), 130.3 (C3/C5), 128.6 (C4)
Pro	175.2	61.6	30.6	25.5	48.2	*trans* configuration,
		61.3	33.1	23.2	48.8	*cis* shows small differences with exception of C_β
Ser	172.6	56.6	62.3			
Thr	172.7	60.2	68.3	20.0		
Trp	176.7	56.7	27.4			C3 (108.4), 112.8 (C7), 137.3 (C8), 127.5 (C9)
Tyr	176.0	57.4	37.0			128.0 (C1), 130.0 (C2/C6), 117.0 (C3/C5), 156.0 (C4)
Val	174.9	60.7	30.8	19.3, 18.5		

The chemical shifts were determined in linear pentapeptides of the form GGXGG. Data adapted from Wuthrich, K. *NMR of Proteins and Nucleic Acids*. John Wiley & Sons, 1986.

techniques carry names that highlight the different correlation made during experiments. For example a 3D HNCO correlates the NH of residue i with the CO of the proceeding residue (i − 1). Triple resonance experiments (Table 10.10) greatly increase the number of assignments, and armed with a larger number of assignments comes the possibility of deriving greater numbers of structural restraints from NOEs and coupling constant data.

The use of multi-dimensional NMR methods has seen larger proteins assigned and it is now feasible to attempt to completely assign proteins of molecular mass in excess of 30 kDa. As of 2003 the NMR derived structures (coordinates) of over 2500 proteins have been deposited in the Protein Databank. Of these proteins the vast majority have masses below 15 000 and a search of databases suggests that structures for proteins with more than 150 residues are increasingly steadily. In 2002 a backbone assignment for a 723-residue protein (malate synthase) was reported and bears testament to the effectiveness of these new multi-dimensional NMR methods. Assignment is, however, only the first step in structure determination and for any protein (large or small) it is necessary to obtain

Table 10.10 A small selection of common triple resonance experiments used for sequential assignment in proteins

Experiment	Observed correlation	Magnetization transfer	J couplings
HNCA	$^{1}H^{N}_{i}$–$^{15}N_{i}$–$^{13}C^{\alpha}_{i}$ $^{1}H^{N}_{i}$–$^{15}N_{i}$–$^{13}C^{\alpha}_{i-1}$		$^{1}J_{NH}$, $^{1}J_{NC}$, $^{2}J_{NC\alpha}$
HNCO	$^{1}H^{N}_{i}$–$^{15}N_{i}$–$^{13}CO_{i-1}$		$^{1}J_{NH}$, $^{1}J_{NCO}$
CBCANH	$^{13}C^{\beta}_{i}/^{13}C^{\alpha}_{i}$–$^{15}N$–$^{1}H^{N}_{I}$ $^{13}C^{\beta}_{i}/^{13}C^{\alpha}_{i}$–$^{15}N_{i+1}$–$^{1}H^{N}_{I+1}$		$^{1}J_{CH}$, $^{1}J_{C\alpha C\beta}$, $^{1}J_{NC\alpha}$, $^{2}J_{NC\alpha}$, $^{1}J_{NH}$
CBCA(CO)NH	$^{13}C^{\beta}_{i}/^{13}C^{\alpha}_{i}$–$^{15}N_{i+1}$–$^{1}H^{N}_{I+1}$		$^{1}J_{CH}$, $^{1}J_{C\alpha C\beta}$, $^{1}J_{C\alpha CO}$, $^{1}J_{NC}$, $^{1}J_{NH}$

a sufficient number of structurally significant NOEs or other conformational restraints to define structure with reasonable precision.

Generating protein structures from NMR-derived constraints

The information derived from NMR spectroscopy about protein conformation can be divided into two major classes. One class of conformational restraints reflect angular information whilst a second class are distance dependent. It is self evident that if sufficient angles and distances between atoms are defined then, in theory at least, it is possible to define overall conformation. The majority of conformational restraints are obtained via homo- and heteronuclear NOESY experiments and convert the NOE cross peak volumes derived from the relationship

$$\mathrm{NOE}_{ij} \propto 1/r_{ij}^{6} \qquad (10.29)$$

into distances that represent the separation between two nuclei i and j. Most schemes categorize NOE cross peak volumes into three classes of restraints (strong, medium and weak with sometimes a fourth group termed very weak). Strong NOEs represent distances between nuclei that are close together ranging from 1.9 to 2.9 Å; medium NOEs represent distances from 2.9–3.5 Å and weak NOEs represent distances from 3.5–5.0 Å. Very weak NOEs are generally assumed to include separation distances in a range 5.0–6.5 Å. In effect these distances represent upper and lower limits for the separation distances. Until recently NOEs were the only experimental restraints used to determine protein structure since the number of torsion angles measurable with accuracy was often limited. However, heteronuclear NMR methods have permitted increased numbers of restraints based on torsion angles derived from the measurement of coupling constants between backbone atoms as well as some side chains.

Most approaches (see Figure 10.33) derive protein structures from NMR data using distance geometry and simulated annealing. In all cases the methods satisfy the conformational data yielding a structure consistent with experimental data and parameters governing bond lengths and angles. Unfortunately almost all NMR restraints include a range of possible values. So for example, an NOE cross peak volume may be consistent with distances ranging between 3.5 and 5.0 Å whilst a $^3J_{\mathrm{HNHA}}$ coupling constant of <5 Hz is consistent with a torsion angle (ϕ) between 0 and −120°. When this is repeated for every NOE and torsion angle many protein conformations are consistent with the data and there is no uniquely defined structure.

The approach used to calculate structures from NMR data involves the generation of an ensemble of ‘low energy’ conformations all consistent with the data that are progressively refined during calculation. Structure calculations based on experimental data derived from NMR spectroscopy yield a family of closely related structures all of which are consistent with the input data. The three-dimensional structures of interleukin-1β and thioredoxin derived using NMR spectroscopy are shown in Figure 10.34, and like their crystallographic counterparts these structures are models of clarity providing a detailed insight into the mechanism of protein function.

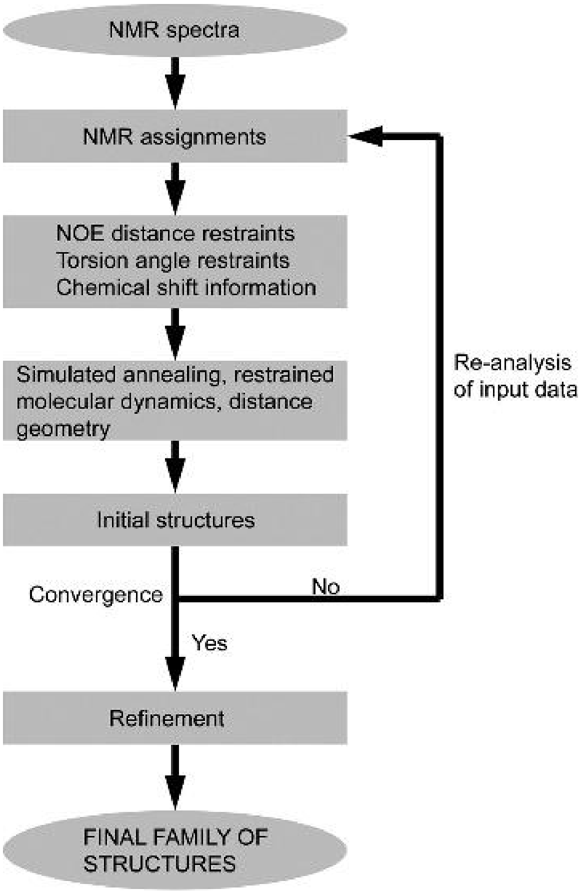

NMR spectra
NMR assignments
NOE distance restraints
Torsion angle restraints
Chemical shift information
Simulated annealing, restrained molecular dynamics, distance geometry
Initial structures
Convergence
No
Yes
Re-analysis of input data
Refinement
FINAL FAMILY OF STRUCTURES

Figure 10.33 The approach used to derive protein structures from torsion angle and distance restraints obtained via NMR spectroscopy

Structure calculations yield families of related conformations all of low overall energy. Although a mean structure calculated from a family of 20–50 closely related structures is often shown this structure is of no more relevance than any of the other structures of low energy generated via structure calculation. The generation of a mean structure allows a root mean square deviation to be calculated from differences in the centre of mass of atoms in a standard (mean) structure with that in family of molecules. A small difference in position for each atom between the mean and test structures leads to a low rmsd value and

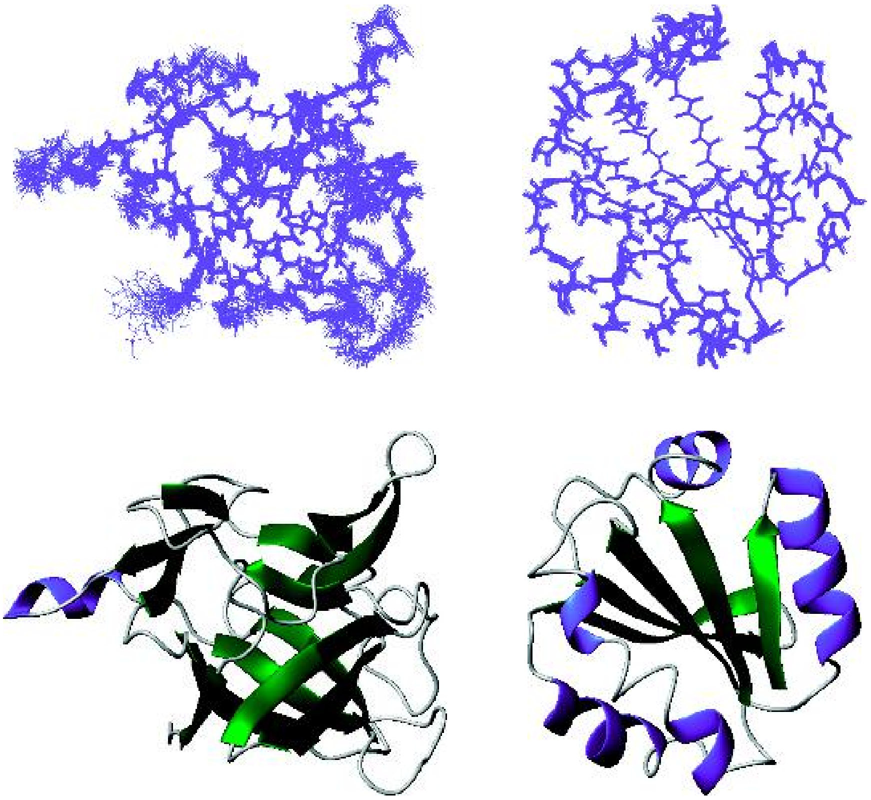


Figure 10.34 The solution structures of interleukin-1β and thioredoxin derived using NMR spectroscopy. The structures show the 50 lowest energy structures for the polypeptide backbone together with the secondary structure distribution of the lowest energy structure (bottom row) (PDB: 7I1B and 1TRU). Interleukin-1β contains ~150 residues and has a mass of ~17 500, whilst thioredoxin contains 105 residues and has a mass of ~11 500. Reproduced courtesy of FEI Company www.feicompany.com

implies structures are similar.

$$\mathrm{rmsd} = \sqrt{\frac{1}{N}\sum_{\mathrm{i}=1}^{N}(r_{\mathrm{mean}} - r_{\mathrm{test}})^2} \qquad (10.31)$$

The precision of NMR structures is ultimately related to the number of experimental restraints defining conformation. If there are no restraints between side chains then structure is poorly defined. Most frequently backbones are defined with highest precision followed by the side chains. In high quality NMR derived structures as many as 50 restraints per residue may be obtained leading to a backbone rmsd of 0.3–0.5 Å. Advances in molecular biology, instrumentation, heteronuclear pulse sequences and computational methods are now permitting structures to be determined using NMR spectroscopy of equal precision to those obtained by X-ray diffraction. However, one group of proteins for which NMR and to a lesser extent X-ray diffraction struggle to provide detailed structures are large asymmetric complexes or macromolecular assemblies.

Cryoelectron microscopy

Visualizing structure has provided enormous impetus to understanding biological processes, and electron microscopy has provided many views of cells and subcellular organelles through the use of a transmission electron microscope (Figure 10.35; TEM).

Pioneering work using electron microscopy to study viral organization in the tail assembly of bacteriophage T4 by Aaron Klug laid the foundation for determining 3D structure from series of 2D images or projections of an object's electron density formed at an image plane (Figure 10.35). Three-dimensional information could be recovered if a number of views of the object were recorded at different angles of observation. This is usually done with a tilting stage that elevates specimens to angles of ~70°. For structures of high rotational symmetry such as helices it is possible to use just a single or small number of orientations to reconstruct a 3D model. The discovery had wider implications since it was exploited in the field of medicine as a medical imaging technique called computerized axial tomography (CAT). This work also allowed Henderson and Unwin to determine the orientation of helices in bacteriorhodopsin and pointed the way forward towards the use of electron microscopy to study proteins at higher levels of atomic resolution. From continuous development of instrumentation, improved specimen preparation, massive advances in data processing and increased computational power electron microscopy has become a technique capable of providing structural information. The technique is useful for proteins not amenable to study by X-ray crystallography or NMR.

Early electron microscopy studies involved 'fixing' biological preparations using cross-linking agents followed by staining with heavy metal (electron dense) compounds; treatments that could distort structure by introducing artifacts. This problem persisted until Kenneth Taylor and Robert Glaeser reported electron diffraction data recorded at a temperature of ~100 K from frozen thin films containing hydrated catalase crystals. In one swoop the technique advanced allowing biological 'units' to be imaged under representative conditions. The technique became known as cryo-electron microscopy (cryo-EM). Sample preparation

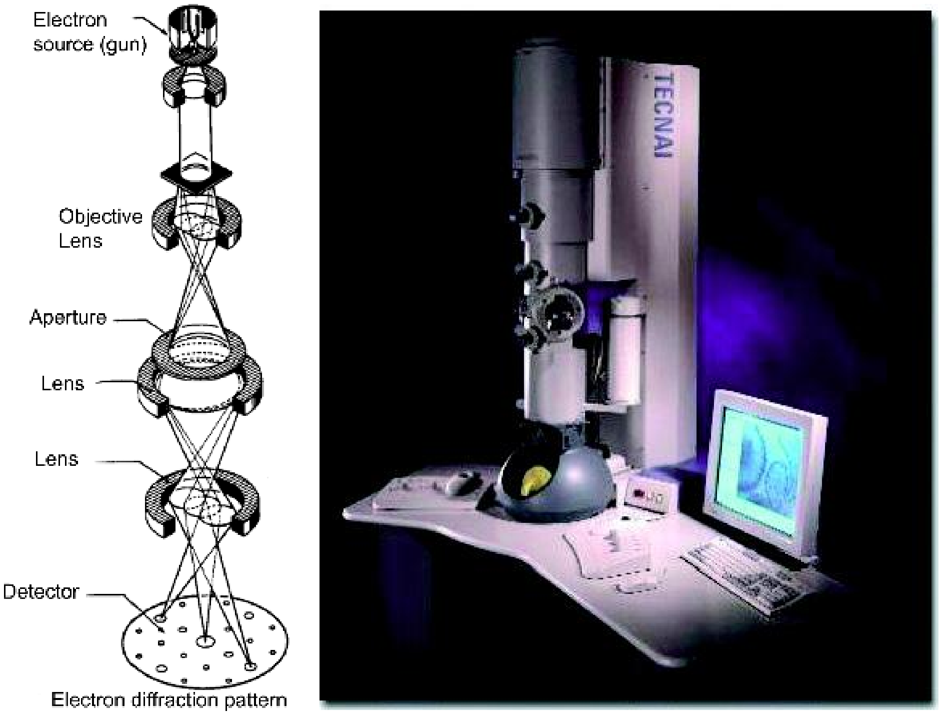


Figure 10.35 An electron diffraction pattern together with a picture of a modern TEM used for structural analysis. (Reproduced by permission of FEI Company.)

is important to cryo-EM and plunging a thin layer supported on a carbon grid into a bath of ethane cooled by liquid nitrogen leads to rapid and efficient cooling. Liquid ethane, a very efficient cryogen, produces a metastable form of water called the vitrified state. The vitrified state is amorphous and lacks the crystalline ice-like lattice. Although metastable the vitrified form is maintained indefinitely at low temperatures (~77 K) and all cryo-EM methods use this approach. Images are recorded at ~100 K using 'low dose' techniques (0.1 electrons $Å^{-2}$) to identify areas of interest and to limit radiation damage. Areas of interest are then captured at greater magnification with a higher dose of electrons (5–10 $Å^{-2}$). The data has a low signal/noise ratio and requires extensive averaging and processing in deriving structural information.

The early successes of cryo-EM centred mainly on viral structure determination and many viral capsids have been defined. One example involved determination of the hepatitis B capsid protein structure by the groups of R.A. Crowther and A. Steven (Figure 10.36). At a resolution of 7.6 Å this study identified elements of secondary structure within subunits.

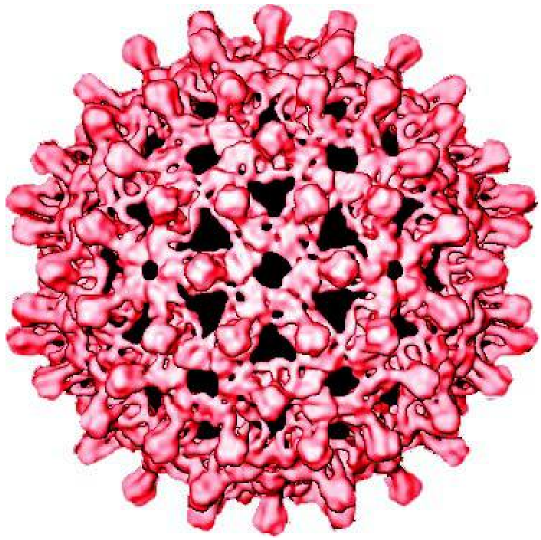


Figure 10.36 The reconstructed 3D structure of the hepatitis B capsid (adapted with permission from Baumeister, W. & Steven, A.C. *Trends Biochem. Sci.* 2000, **25**, 625–631. Elsevier)

Single-particle analysis deals with large proteins or protein complexes and is growing in its application to biological problems largely through the efforts of Joachim Frank in the study of ribosome structure. In this method macromolecular complexes are trapped in random orientations within vitreous ice and images from all possible orientations are combined and averaged to provide enhanced views.

From a sample preparation point of view single particle analysis is straightforward – it requires a well-dispersed, homogeneous complex frozen in amorphous ice and aligned over the holes of a carbon grid. Structure determination from differently oriented particles requires the combination of data from many images using methods that distinguish 'good' from 'bad' particles, determine the orientation relative to a reference point and evaluate the quality of the resulting data. These methods are intensely computational and beyond the scope of this text. Although the method has not produced data comparable in resolution to X-ray diffraction or NMR spectroscopy methods the technique has allowed ribosomal subunits to be determined at resolutions approaching 7.0 Å. Significantly, the results of crystallography and/or NMR can often be assimilated into the cryo-EM data to build models based on more than one structural technique.

The power of electron crystallography has been demonstrated for bacteriorhodopsin but success has also been achieved with a variety of proteins including tubulin, light harvesting complexes associated with the reaction centres of photosynthetic organisms, aquaporin a transport protein of the red blood cell membrane and the acetylcholine receptor of neurones.

Tubulin is a key protein of the microtubule assembly regulating intracellular activity particularly cytoskeletal function and eukaryotic cell division where it facilitates the separation of chromatids as part of mitosis. In this role the ability of microtubules to polymerize and depolymerize is important. The microtubule contains 13 protofilaments arranged in parallel to form a hollow structure approximately 24 nm in diameter. The protofilaments are composed of alternating α- and β-tubulin monomers each composed of a single polypeptide chain (the α and β subunits) of ~55 kDa. Within the microfibril each protofilament is staggered

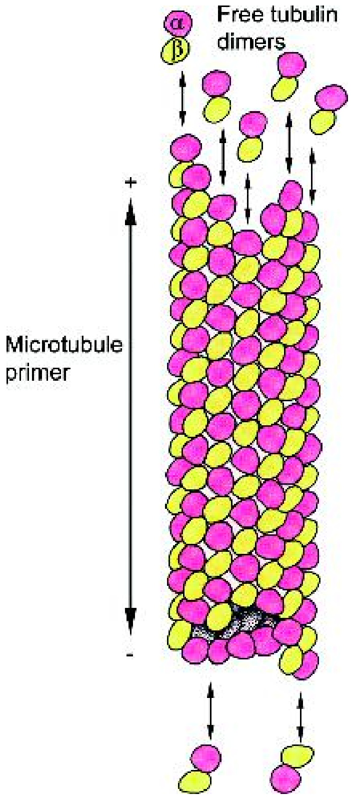


Figure 10.37 The arrangement, assembly and disassembly of tubulin dimers within a microtubule. Assembly is primarily from a preformed microtubule primer. The growing (+) end represents rapid addition of dimers when compared with rates of removal whilst at the (−) end the converse is true. Assembly is enhanced at 37 °C and by the presence of GTP whilst the absence of nucleotide and lower temperatures drives disassembly (adapted from Darnell *et al. Molecular Cell Biology*, 2nd edn. Scientific American Books, 1990.)

with respect to the next to give a characteristic helical pattern (Figure 10.37).

Assembly involves GTP binding to α and β subunits. Subunit addition brings β -tubulin into contact with the α subunit promoting GTP hydrolysis on the interior β-tubulin with GTP bound to α-tubulin unhydrolyzed during polymerization. In the presence of Zn ions purified tubulin assembles into a two-dimensional ordered sheet ideal for electron crystallography (Figure 10.38). In these sheets the protofilaments are similar in appearance to those in microtubules with the exception that they are associated in an antiparallel array. The polymerization reactions make studies using NMR spectroscopy unfeasible whilst the inherent mobility of the filaments has prevented crystallization suitable for X-ray diffraction.

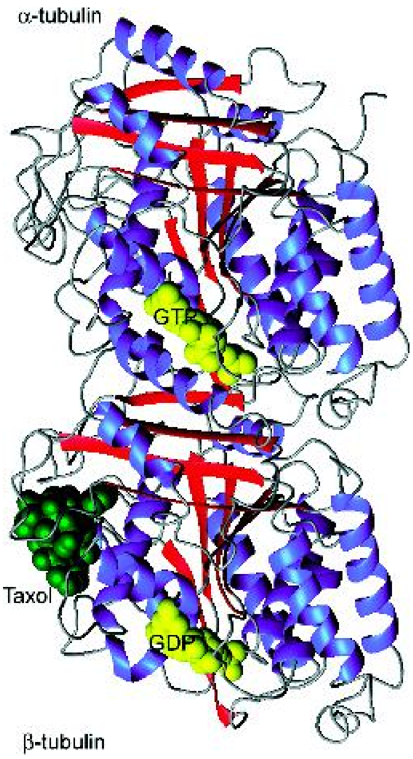


Figure 10.38 The structure deduced for the αβ tubulin dimer determined at a resolution of 0.37 nm by electron crystallography (PDB: 1TUB reproduced with permission from Nogales, E. *et al. Nature* 1998, **391**, 199–203. Macmillan Publishers Ltd)

The electron density profile determined directly from cryo-EM studies for the αβ dimer allowed the sequence of each subunit to be fitted directly with little ambiguity. Each monomer was a compact structure containing a core of β strands assembled into a sheet surrounded by α helices. A conventional Rossmann fold was found at the N-terminal region and contained a nucleotide-binding region. An intermediate domain contained a four-stranded β sheet and three helices along with a Taxol-binding site. A third domain of two antiparallel helices crossing the N- and intermediate regions represented a binding surface for additional motor proteins. A Mg ion was bound near the GTP binding site whilst a binding site for Zn ions was identified at the interface between protofilaments.

Cryo-EM methods are now firmly established as a third method for macromolecular structure determination. The list of successes is continuing to expand with increasingly detailed pictures of the ribosome, the thermosome, viral capsids, transport proteins, amyloid fibres and receptors obtained using cryo-EM methods. There is little doubt that cryo-EM will increase in importance as a method of three-dimensional structure determination for macromolecular complexes.

Neutron diffraction

Neutron diffraction requires the crystallization of proteins followed by measurement of the diffraction of neutrons. Neutron beams were traditionally generated *via* atomic reactors but increasingly sources are derived *via* synchrotron beam lines. Until the introduction of synchrotron radiation a major problem was achieving sufficiently high neutron fluxes to enable adequate data collection times. Although less widely used the technique has proved very useful for locating hydrogen atoms within protein crystals something that is difficult to achieve with X-rays.

The usefulness of neutron diffraction to biological structure determination in the post-genomic era involves its ability to probe dynamic properties involving protons. Hydrogen exchange underscores many biological reactions and neutron diffraction is adept at locating protons that exchange for deuterium within enzyme active sites. Such observations help to establish catalytic mechanisms and the exchange of hydrogen for deuterium leads to a large change in neutron scattering factors.

Two noteworthy examples where neutron diffraction has proved of value are the catalytic mechanisms of serine proteases and lysozyme. In serine proteases such as trypsin the catalytic mechanism revolves around the catalytic triad of invariant Ser, His and Asp residues (Figure 10.39). The nucleophilic Ser residue forms two tetrahedral intermediates during bond cleavage and is assisted by increased imidazole basicity caused by hydrogen bonding to the aspartate. Hydrogen bonding could arise from protonation of the imidazole group or protonation of aspartyl side chains.

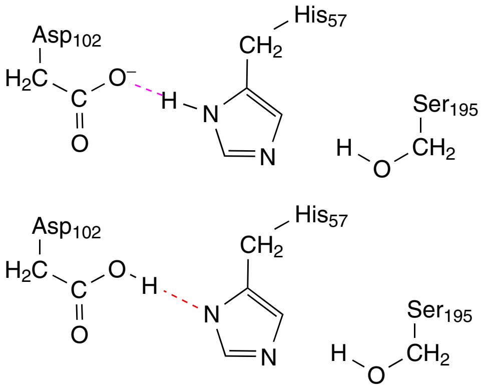


Figure 10.39 The exchangeable proton in the catalytic site of serine proteases was shown to remain attached to His57 and was not transferred to Asp102

Neutron protein crystallography showed that the proton remained attached to the imidazole ring of His 57 and not the carboxylate of Asp 102 and clarified the reaction mechanism. Similarly in lysozyme, the deuterium atom was found to reside on the carboxyl side chain of Glu 35, rather than Asp 52, again emphasizing a key aspect of the catalytic mechanism.

Optical spectroscopic techniques

Absorbance

Transitions between different electronic states occur in the ultraviolet, visible and near infrared regions of the electromagnetic spectrum and are widely used for studying protein structure. Absorbance involves transitions of outer shell electrons between various electronic states and is governed by the rules of quantum mechanics. The absorbance of light excites an electron from the ground state to a higher excited state (Figure 10.40).

Absorbance is governed by several rules; the frequency of the incident radiation must match the

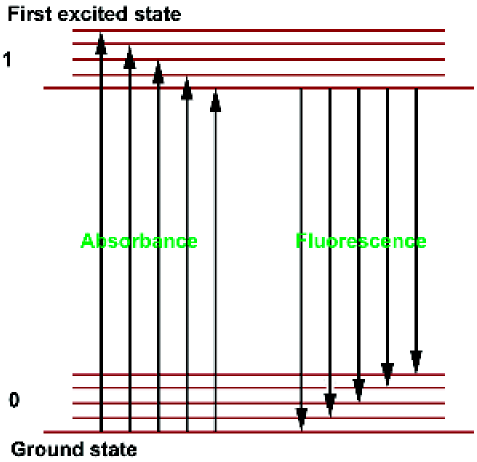


Figure 10.40 Absorbance and fluorescence resulting from excitation and emission between the ground and first excited state

quantum of energy necessary for transition from ground to excited states (i.e. $h\nu = \Delta E$). When the resonance condition is satisfied transitions can occur if further complex selection rules are obeyed. Selection rules are divided into high probability or allowed transitions and forbidden transitions of much lower probability. Within the latter set of transitions are spin-forbidden and symmetry-forbidden transitions. Spin forbidden transitions involve a change in spin multiplicity defined as $(2S + 1)$ where S is the electron spin number (analogous to I, the nuclear spin quantum number). Spin multiplicity reflects electron pairing (see Table 10.11). For a favourable transition there is no change in multiplicity ($\Delta S = 0$).

Table 10.11 Spin multiplicity for atoms and molecules

No. of unpaired electrons.	Electron spin S	$(2S + 1)$	Multiplicity
0	0	1	Singlet
1	1/2	2	Doublet
2	1	3	Triplet
3	3/2	4	Quartet

Symmetry-forbidden transitions reflect redistribution of charge during transitions in a quantity called the transition dipole moment. Differences in dipole moment arise from the different electron distributions of ground and excited states and when these differences approximate to zero the transition is forbidden. To express this slightly differently absorbance requires a change in dipole moment.

Absorbance is normally observed by peaks of finite intensity and definitive linewidth in plots against frequency. The intensity is reflected by a Boltzmann distribution that describes the population of each energy level. In molecules atomic interactions lead to the association of electrons with neighbouring nuclei whilst the nuclei move relative to each other in the form of quantized vibrational or rotational motion. As a result molecules possess ground and excited states that split into a variety of sub-states, each corresponding to different vibrational modes, and then to further rotational sub-states. Absorbance of light leads to transitions from the lowest ground state to a variety of energy levels in the first excited state. In the absence of chemical reactions or energy transfer the electrons return to the ground state. Absorbance spectra of proteins, when recorded at room temperature, consist of several closely related transitions each of similar, but non-identical, frequency leading to distinct bandwidths, reflecting the statistical sum of all individual transitions. Despite broad bands absorbance spectra are frequently characterized by maxima diagnostic of a particular transition and chromophore (e.g. A_{280} for proteins). In proteins transitions often involve aromatic side chains (Figure 10.41) but in the near UV/visible regions (350–750 nm) co-factors such as heme, flavin, or chlorophyll will also give intense absorbance bands. Using molecular orbital theory electrons are defined according to the orbitals in which they reside as either σ, π or n (non-bonding) with the corresponding anti-bonding orbitals denoted as σ^*, π^* or n*. Transitions between $\sigma \rightarrow \sigma^*$ lie in the far UV region (large energy difference) and are not normally observed by optical methods, but transitions between $\pi \rightarrow \pi^*$ and $n \rightarrow \pi^*$ are frequently observed in the UV and visible region of the electromagnetic spectrum. Physicists and chemists plot the spectra of molecules in the form of intensity versus frequency, where the frequency is expressed as

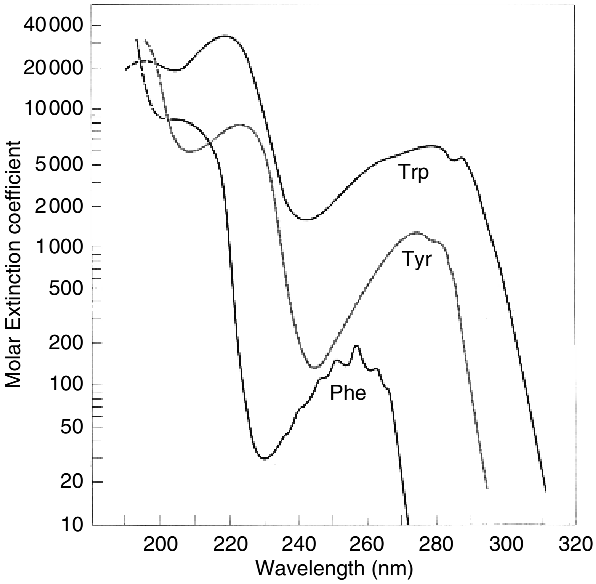


Figure 10.41 Detailed absorbance spectra in the UV visible region for the aromatic amino acids Tyr, Trp and Phe. In proteins a summation of these profiles is observed reflecting the number of residues (reproduced and adapted with permission from Wetlaufer, D.B. *Adv. Prot. Chem.* 1962, **17**, 304–390. Academic Press)

a wavenumber. The wavenumber ($\bar{\nu}$) is defined as

$$\bar{\nu} = 1/\lambda = \nu/c \qquad (10.31)$$

with molecular transition frequency (λ) expressed in cm^{-1}, and where c is the velocity of light. In contrast, biochemists usually plot intensity versus wavelength, although in view of the relationship $c = \nu\lambda$ it is straightforward to convert these equations.

Fluorescence

Fluorescence is the process by which electronically excited molecules decay to the ground state via the emission of a photon without any change in spin multiplicity. Emission is detected by spectrofluorimeters and occurs at longer wavelengths than the corresponding absorbance band (Figures 10.42 and 10.43).

The quantum yield of fluorescence emission is defined as the ratio of photons emitted through fluorescence to the number of photons absorbed. Its maximum value is 1, although other processes contribute to deactivation of excited molecules and lower values are observed.

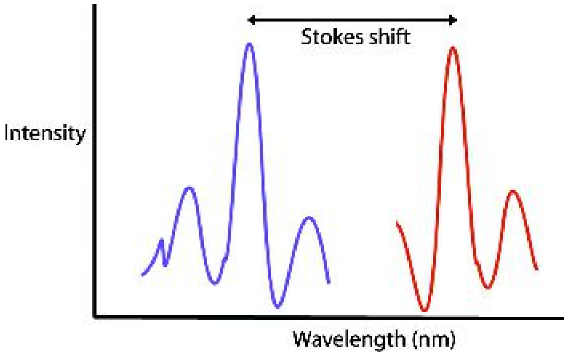


Figure 10.42 Absorbance and fluorescence spectra for a chromophore. The difference between the peaks of maximum intensity is called the Stokes shift

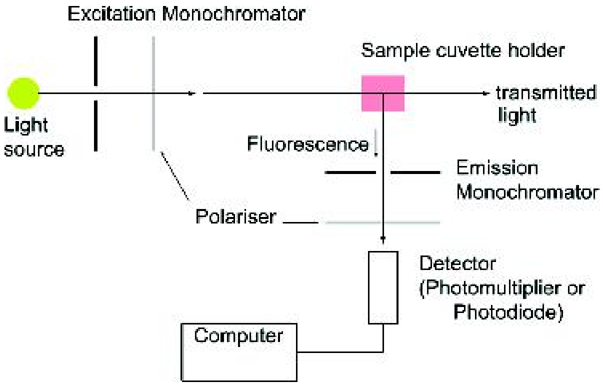


Figure 10.43 A basic instrumental arrangement for detection of fluorescence. The light source is a deuterium- or xenon-based lamp with the wavelength created by the use of diffraction gratings. Light is passed through one or more adjustable excitation monochromators and strikes a sample cuvette. Fluorescence, measured at right angles to the incident beam, is recorded by a photomultiplier sensitive between ~200–1000 nm. Polarizers can be placed into the light path of both excitation and emission beams for measurement of fluorescence polarization

In proteins the principal fluorophore is the indole side chain of tryptophan with a contribution that exceeds the other aromatic residues, and dominates the fluorescence of proteins between 300 and 400 nm. Folded states of proteins show different fluorescence spectra to unfolded states due to the influence of local environment and solvent. In the folded state tryptophan emission shows a blue shift towards 330 nm from a 'free' or unfolded value of ~350 nm. Changes in intensity and wavelength reflect the influence of local environment on emission. Other aromatics side chains or molecules such as heme may quench fluorescence whilst collisional quenching by small molecules such as oxygen, iodide and acrylamide is an important effect.

Collisional quenching leads to excited state deactivation and is a nanosecond process occurring via diffusion-controlled encounters. Aromatic molecules on protein surfaces are quenched more effectively than buried groups with the fluorescence intensity decaying exponentially with time. The intensity at time t is given by

$$I = I_o e^{-t/\tau} \tag{10.33}$$

where I_o is the intensity at time $t = 0$ and τ is the mean lifetime of the excited state. The mean lifetime is therefore the time decay for the fluorescence intensity to fall to 1/e. Measurement of fluorescent lifetimes is possible with sensitive photon-counting fluorimeters and is reflected by time constants ranging from a few picoseconds to hundreds of nanoseconds. These measurements shed light on molecular motion and fluorophore environments.

Collisional quenching can determine the accessibility of fluorophores in proteins via the Stern–Volmer analysis. It is modelled as an additional process deactivating the excited state. Collisional quenching depopulates the excited state leading to decreased fluorescence lifetimes. The dependence of emission intensity, F, on quencher concentration [Q] is given by the Stern–Volmer equation

$$F_o/F = \tau_o/\tau = 1 + k_q\tau_o[\mathrm{Q}] \tag{10.34}$$

$$= 1 + K_{SV}[\mathrm{Q}] \tag{10.35}$$

where τ and τ_o are the lifetime in the presence and absence of quencher, k_q is the bimolecular rate constant for the dynamic reaction of the quencher with the fluorophore and the product $k_q\tau_o$ is called the Stern–Volmer constant or K_{SV}. A low value for K_{SV} indicates that residues (fluorophores) have low solvent exposure and are buried within a protein.

In many proteins tyrosine fluorescence is effectively quenched by tryptophan as a result of energy transfer occurring primarily from dipole–dipole (acceptor–donor) interactions where the magnitude is governed by the distance, intervening media and orientation. Semiquantitative models of fluorescence resonance energy transfer allow estimation of distances between donor and acceptor. Energy transfer involving a dipolar (through space) coupling mechanism of the donor excited state and the acceptor is known as Forster energy transfer and varies as the inverse sixth power of the intervening distance.

Fluorescence polarization measurement associated with emission intensity and fluorescence anisotropy provide insight into the mobility of fluorophores in proteins. Fluorescence polarization occurs when linearly polarized light is used to excite molecules. Molecules oriented in parallel to the direction of propagation are preferentially excited; for fixed or rigid molecules arranged in parallel this will yield maximum polarization whilst motion associated with the fluorophore leads to randomized orientations and depolarization. Depolarization is also sensitive to the rotational motion or 'tumbling' of molecules. This Brownian motion is described by a rotational correlation time (τ_c) that to a first approximation is estimated assuming spherical proteins of known radius. The correlation time reflects a relaxation process and if it is shorter than the fluorescence decay time leads to molecular reorientation between the events of absorbance and emission.

The degree of fluorescence polarization, P, is defined as

$$P = (I_{\parallel} - I_{\_})/(I_{\parallel} + I_{\_}) \tag{10.37}$$

where $I_{\parallel}$ is the fluorescence intensity measured with polarization parallel to the absorbed plane-polarized radiation, and $I_{\_}$ is the fluorescence intensity measured perpendicular to the absorbed radiation. A rigid system leads to a maximum value of 1.0 whilst a system exhibiting complete depolarization due to molecular tumbling has a value of −1. These values are theoretical and are rarely, if ever, observed in solution. Typical rotational correlation times for proteins in solution are of the order of 1–100 ns.

Frequently, proteins show partial polarization or are completely unpolarized as a result of tumbling. In addition, internal motion, protein concentration, solvent viscosity, temperature, as well as resonance energy transfer can lead to substantial alterations in polarization. The most common method of analysing fluorescence polarization involves the Perrin equation

$$1/P - 1/3 = (1/P_o - 1/3)(1 + 3\tau/\tau_c) \qquad (10.38)$$

where τ is the fluorescence lifetime, τ_c is a correlation time reflecting the rotational relaxation rate due to molecular tumbling and is equivalent to

$$\tau_c = 4\pi\eta r^3/k_B T \qquad (10.39)$$

where η is the solvent viscosity and r is the hydrodynamic radius. P_0 represents the intrinsic polarization observed under conditions promoting least molecular motion such as high solvent viscosity and low protein concentration. For non-spherical proteins the equation should be re-written as $\tau_c = 3\eta V/RT$ where V is the volume.

Fluorescence polarization depends on the rotational correlation time *and* the fluorescence lifetime and is difficult to use if there is no knowledge of the lifetime (τ) or the limiting value of P_o. Fluorescence anisotropy decays as the sum of exponentials and is used in time resolved measurements. Anisotropy is defined as the ratio of the difference between the emission intensity parallel to the polarization of the electric vector of the exciting light ($I_\parallel$) and that perpendicular ($I_\perp$) divided by the total intensity (I_T), where $I_T = I_\parallel + 2I_\perp$

$$A = (I_\parallel - I_\perp)/(I_\parallel + 2I_\perp) \qquad (10.40)$$

The emission anisotropy (A) is related to the correlation time of the fluorophore (τ_c) through the Perrin equation

$$A_o/A = 1 + \tau/\tau_c \qquad (10.41)$$

where A_o, is the limiting anisotropy of the fluorophore and depends on the angle between the absorption and emission transition dipoles, and τ is the fluorescence lifetime. Fluorescence anisotropy is used to probe tryptophan residues in proteins as well as motion of covalently attached fluorophores. Measurement of tryptophan anisotropy reveals that proteins often exhibit additional motion associated with Trp residues and has been attributed to rotation of the indole side chain about the CA–CB bond with models involving precession of the side chain within a cone.

Green fluorescent protein

The green fluorescent protein (GFP) is responsible for the natural green 'colour' exhibited by the jellyfish, *Aequorea victoria* and other coelenterates, where it gives cells a characteristic glow visible on the surface of sea at sunset. Details of the structure of GFP (Figure 10.44) have revealed an 11-stranded antiparallel β sheet forming a barrel-like structure of diameter 3.0 nm and length 4.0 nm and that has been likened to a 'lantern'. GFP is a remarkably stable protein, resistant to proteases and denatured only under extreme conditions such as 6 M guanidine hydrochloride at 90 °C. The lantern functions to protect the 'flame' which is positioned in a single α helix located towards the centre of the β barrel. The 'flame' responsible for the intrinsic fluorescence of GFP is *p*-hydroxy-benzylideneimidazolinone and results from cyclization of the tripeptide Ser65-Tyr66-Gly67 and 1,2-dehydrogenation of the Tyr.

GFP absorbs blue light around 395 nm preferentially but also exhibits a smaller absorbance peak at 475 nm (ε of ∼30 000 and 7000 $\text{M}^{-1}\text{cm}^{-1}$ respectively). In the coelenterate *A. victoria* GFP absorbs blue light by fluorescence energy transfer from a second protein aequorin converting the incident 'blue' light via energy transfer into green light. *Aequorea* are luminescent jellyfish observed to glow around the margins of their umbrellas as a result of light produced from photogenic cells in this region. Exposure to physical stress causes the intracellular calcium concentrations to increase leading to an excited state of aequorin generated through Ca^{2+} binding to the pigment coelenterazine. The light produced by aequorin generates GFP fluorescence by energy transfer, although the physiological basis of these reactions is unclear (Figure 10.45).

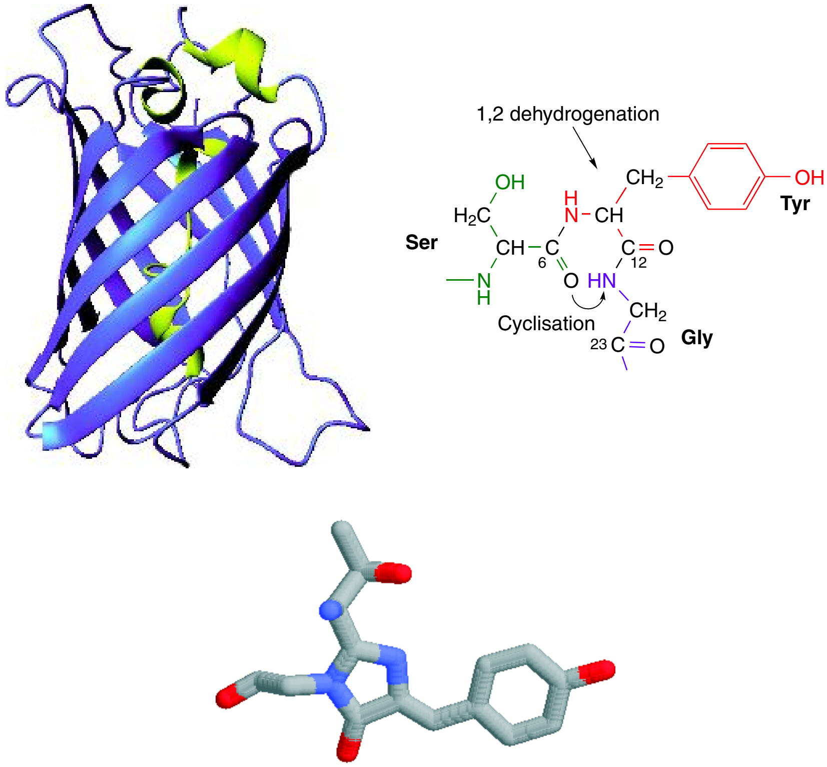


Figure 10.44 The structure of GFP showing 11-stranded β barrel (PDB:1EMA). The cyclic *p*-hydroxybenzylideneimidazolinone fluorophore

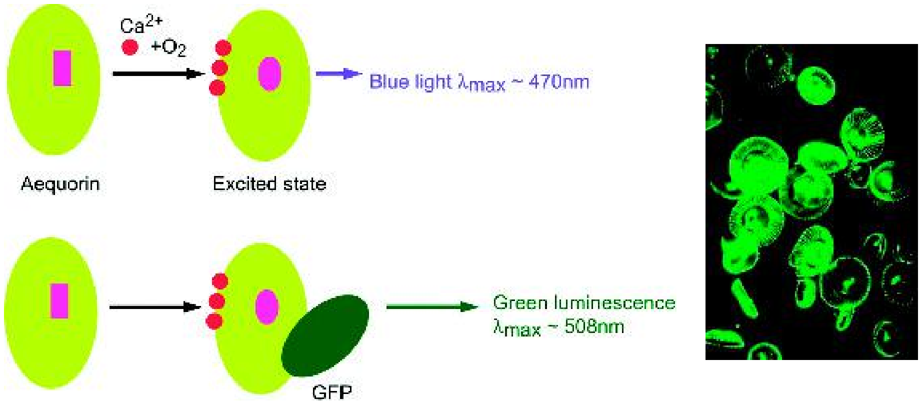


Figure 10.45 GFP fluorescence. *In vitro* mixing of aequorin and Ca^{2+} produces blue light ($\lambda_{max} \sim 470$ nm). The blue fluorescent form of aequorin represents an excited state that decays to the ground state. In the presence of GFP energy transfer occurs with a green luminescence observed comparable to that seen in the living animal (right) (reproduced with permission from Yang F., Moss L.G. & Phillips G.N. *Nat. Biotechnol.* 1996, **14**, 1246–1251)

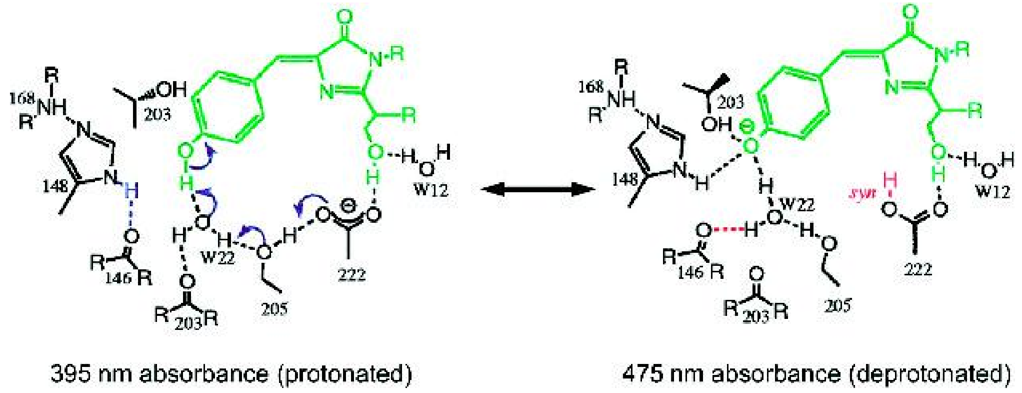

168
NH
146
148
203
W22
205
222
W12
syn
395 nm absorbance (protonated)
475 nm absorbance (deprotonated)

Figure 10.46 The structure at the active site of GFP associated with the 398 and 475 nm absorbance bands (reproduced with permission from Brejc, K. *et al. Proc. Natl Acad. Sci.* 1997, **94**, 2306–2311)

Absorbance of GFP around 398 nm leads to a fluorescence emission peak at 509 nm with a quantum yield between 0.72–0.85, and hence the green colour observed *in vivo* and *in vitro*. The intensity ratio between the absorbance peaks is sensitive to factors such as pH, temperature, and ionic strength, suggesting the presence of two different forms of the chromophore. It has been shown that the two forms differ in their protonation states with the 398 nm band reflecting a protonated fluorophore whilst the 475 nm band reflects deprotonation of this group (Figure 10.46).

The unique fluorescence of GFP allows the expressed protein to be fused to other domains creating a chimeric protein with a fluorescent tag that can be measured and has been used in N- and C-terminal fusions to follow gene expression, protein–protein interactions, cell sorting pathways, and intracellular signalling.

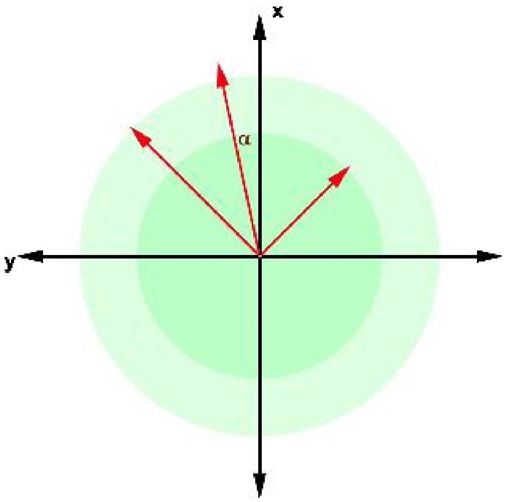


Figure 10.47 The resultant rotation of ε when ε_R and ε_L make unequal angles with the x-axis. A levorotatory chiral centre exists causing a rotation of angle α

Circular dichroism

Circularly polarized light travels through optically active media with different velocities due to different indices of refraction for right (dextro) and left (laevo) components (Figure 10.47). This is called optical rotation and the variation of optical rotation with wavelength is known as optical rotary dispersion (ORD). ORD is normally measured at a specific wavelength and temperature and is usually denoted by a parameter α, the angle of rotation of polarized light.

The right and left circularly polarized components are also absorbed differentially at some wavelengths due to differences in extinction coefficients for the two polarized components (Figure 10.48). The addition of the left and right components yields ε at every point from the relationship $\varepsilon = \varepsilon_L + \varepsilon_R$. When this light is

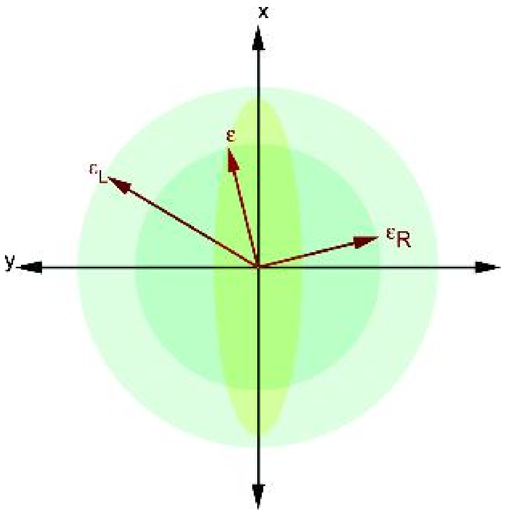


Figure 10.48 The variation of ε for a circular dichroic sample; the vector traces out an ellipse

passed through an optically active medium the left and right circularly polarized waves are rotated at different speeds due to small differences in molar absorptivity for the left and right components. The differential absorption of the left and right circularly polarized light gives rise to the technique of circular dichroism (CD).

CD spectroscopy measures differences in absorbance of right and left circularly polarized light. In proteins the optically active absorption bands tend to be located in the far-UV or near-UV regions of the spectrum and involve $\pi \rightarrow \pi^*$, $n \rightarrow \pi^*$, and $\sigma \rightarrow \sigma^*$ transitions. The chromophores responsible for these transitions are either intrinsically asymmetric or more frequently asymmetric as a result of their interactions with the protein. CD spectroscopy is of considerable importance to the study of proteins as a quantitative tool in the estimation of secondary structure. This arises because regularly repeating helix or sheet produce different CD signals. In a protein CD signals reflect the sum of the secondary structure components and it has proved possible to estimate the proportions of helix, strand or turn in a protein of unknown structure with remarkable precision.

Historically, ellipticity is the unit of CD and is defined as the tangent of the ratio of the minor to major elliptical axis. An axial ratio of 1:100 will therefore result in an ellipticity of 0.57°. In CD spectroscopy the measured parameter is the rotation of polarized light expressed with the units of millidegrees. Modern CD instrumentation is capable of precision down to thousandths of a degree (1 millidegree). There are a number of ways of expressing the CD signal of a sample although frequently many incorrect forms are used in the literature. The observed CD signal (S) is expressed in millidegrees and is normally converted to $\Delta\varepsilon_m$ or $\Delta\varepsilon_{mrw}$, where $\Delta\varepsilon_m$ is the molar CD extinction coefficient and $\Delta\varepsilon_{mrw}$ is the mean residue CD extinction coefficient, respectively. In any CD experiment it is vital to know the protein concentration extremely accurately to avoid significant error. It can be shown that

$$\Delta\varepsilon_m = S/(32980Cl) \qquad (10.42)$$

where C is the concentration in mol dm^{-3} (or M) and l is the path length in cm. The light path is usually a value between 0.1 and 1.0 cm. The units of $\Delta\varepsilon_m$ are therefore M^{-1}cm^{-1} and the analogy with the molar extinction coefficient determined via absorbance measurement is clear. Alternatively, expressing the ellipticity as a mean residue CD extinction coefficient leads to

$$\Delta\varepsilon_{mrw} = Smrw/(32980Cl) \qquad (10.43)$$

where the mean residue weight (mrw) is the molecular weight divided by the number of residues. In this instance C is expressed in mg ml^{-1}. CD intensities are sometimes reported as molar ellipticity ($[\theta]_M$) or mean residue ellipticity ($[\theta]_{mrw}$). These terms are calculated from the following equalities

$$[\theta]_M = S/(10Cl) \qquad (10.44)$$

where C is again the concentration in mol dm^{-3} or

$$[\theta]_{mrw} = Smrw/(10Cl) \qquad (10.45)$$

Both $[\theta]_M$ and $[\theta]_{mrw}$ have the units degrees cm^2 dmol^{-1}. $[\theta]$ and $\Delta\varepsilon$ may be inter-converted using the relationship

$$[\theta] = 3298\Delta\varepsilon \qquad (10.46)$$

CD extinction coefficients are the most logical unit since they are direct analogs of extinction coefficient in absorbance measurements and lead to values of $\Delta\varepsilon_{\mathrm{mrw}}$ in an approximate range ±20 whilst the corresponding values for $\Delta\varepsilon$ range from ±3000. Using the above relationships it will be apparent that values of $[\theta]_{\mathrm{mrw}}$ values are in the range ±70 000.

In the determination of secondary structure content one approach is to assume that spectra are linear combinations of each contributing secondary structure type ('pure' α helix, 'pure' β strand, Figure 10.49) weighted by its relative abundance in the polypeptide conformation. Unfortunately the problem with this approach is that there are no standard reference CD spectra for 'pure' secondary structure. More significantly synthetic homopolypeptides are poor models of secondary structure and most homopolymers do not form helices nor is there a good example of a 'model' β strand. An empirical approach involved determining the experimental CD spectra of proteins for which the structures are already known. Using knowledge of the structure the content of helix, turn and strand is defined accurately and by using a database of reference proteins these methods prove accurate and reliable when applied to unknown samples. A number of different mathematical procedures have been adopted using reference proteins of varying size and secondary structure content and all of these methods give similar results.

A variation of CD occurs when samples are placed in magnetic fields maintained at low temperatures (<77 K). Under these conditions all molecules exhibit CD spectra and the technique is called magnetic circular dichroism (MCD). In most cases the spectra of proteins are too complex to interpret fully but for metalloproteins the MCD technique provides a powerful tool with which identify ligands and metal ions.

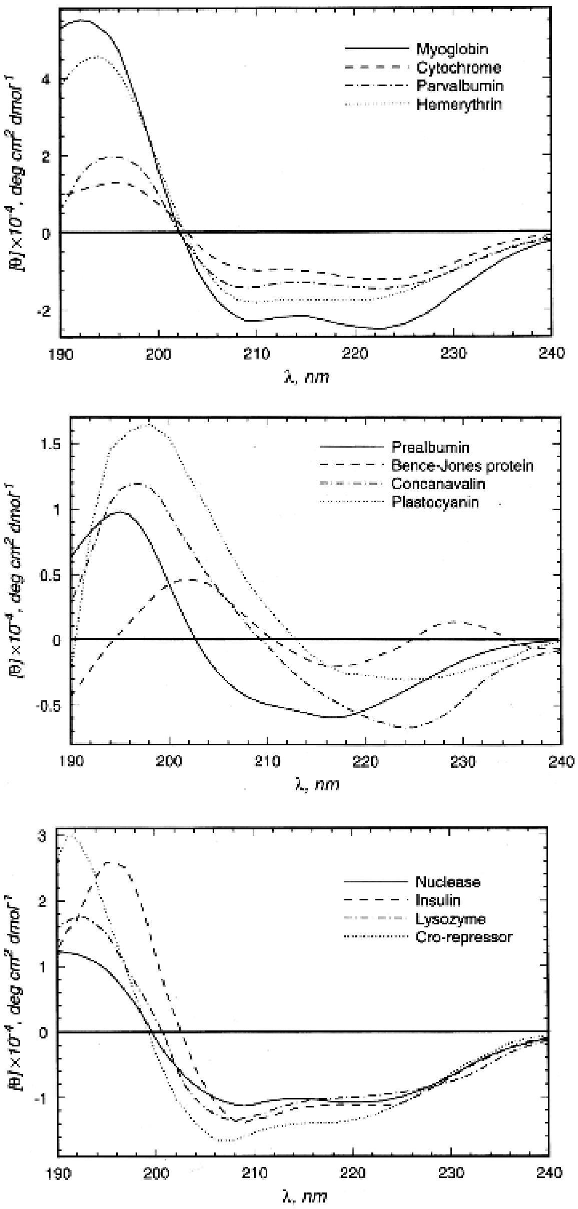


Figure 10.49 CD spectra of proteins with diverse secondary structure content. Top: extensively α helical proteins; centre: proteins rich in β strands; bottom: protein with (α + β) structure (reproduced with permission from *Circular Dichroism* – the conformational analysis of biomolecules, Fasman, G.D (ed) Kluwer Academic 1996.)

Vibrational spectroscopy

The vibrational spectra of proteins are extremely complex and lie at lower frequencies than electronic spectra within the infrared (IR) region. It is useful to divide the IR region into three sections, the near, middle and far IR regions with the most informative zone located at wavenumbers between 4000 and 600 cm^{-1} (Table 10.12). Historically, IR

Table 10.12 The relationship between wavenumber and wavelength for the near, middle and far IR regions

Region	Wavelength range (μm)	Wavenumber range (cm^{-1})
Near	0.78–2.5	12 800–4000
Middle	2.5–50	4000–200
Far	50–1000	200–1000

It is worth noting that 750 nm is located at the far red end of the visible spectrum. This is 0.75 μm or a wavenumber of $1/0.75 \times 10^{-4}$ cm^{-1} or 13 333 cm^{-1}.

spectroscopy uses wavenumbers to denote frequency. In any molecule the atoms are not held in rigid or fixed bonds but move in a manner that is reminiscent of two bodies attached by a spring. Consequently all bonds can either bend or stretch about an equilibrium position that is defined by their standard bond lengths (Figure 10.50).

Exposure to IR radiation with frequencies between 300 and 4000 cm^{-1} leads to the absorption of energy and transition from the lowest vibrational state to a higher excited state. In a simple diatomic molecule there is only one direction of stretching or vibration and this leads to a single band of IR absorption. Atoms that are held by weak bonds require less energy to reach excited vibrational modes and this is reflected in the frequencies associated with IR absorbance. As molecules increase in complexity more modes of vibrations exist with the appearance of more complex spectra. For a linear molecule with n atoms, there are $3n - 5$ vibrational modes whilst if it is non-linear it will have $3n - 6$ modes. Water is a non-linear molecule containing three atoms and has three modes of vibration in IR spectra (Figure 10.51).

There is one more condition that must be met for a vibration to be IR active and this requires changes to the electric dipole moment during vibration ($d\mu/dr \neq 0$). This arises when two oppositely charged atoms move positions. By treating a diatomic molecule as a simple harmonic oscillator Hooke's law can be used to calculate the frequency of radiation necessary to cause a transition. For a simple harmonic oscillator the

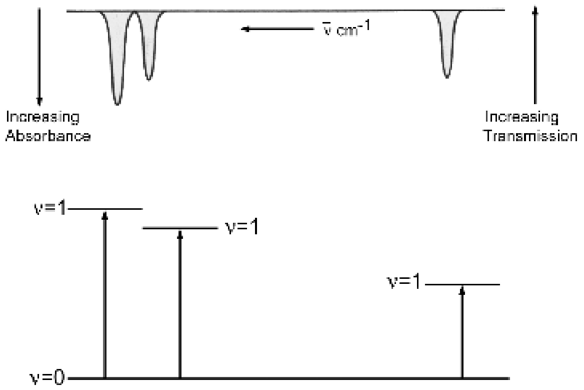


Figure 10.51 The IR spectrum of water showing three fundamental modes of vibration. At higher resolution each peak will show fine structure as a result of simultaneous transitions between different rotational levels

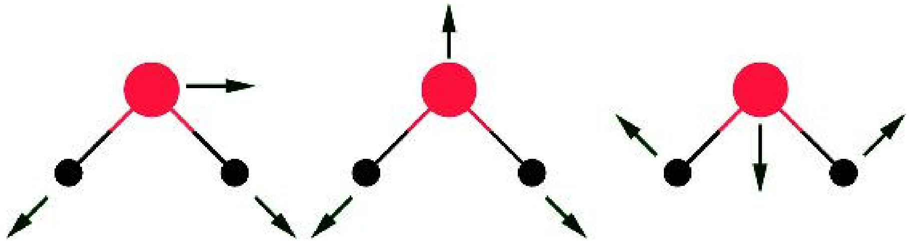


Figure 10.50 Bond stretching and bending in a simple molecule (H_2O). Ignoring fine structure arising from simultaneous transitions the IR spectrum of water has three peaks at 3943, 3832 and 1648 cm^{-1}

frequency of vibration (ν) is given by

$$\nu = \frac{1}{2\pi}\sqrt{\frac{k}{\mu}} \qquad (10.47)$$

where k is the force constant determined by bond strength, and μ is the reduced mass. The term μ, the reduced mass is equivalent to $(m_1 + m_2)/m_1 m_2$ and leads to

$$\nu = \frac{1}{2\pi}\sqrt{\frac{k\ m_1 m_2}{m_1 + m_2}} \qquad (10.48)$$

From the above equation it can be seen that a strong bond between atoms leads to bands at higher frequency; IR spectra of C≡C bonds absorb between 2300 and 2000 cm^{-1}, C=C bonds occur at 1900–1500 cm^{-1} whilst C–C occur below 1500 cm^{-1} and the spectral region becomes crowded and complex. Despite complexity it is called the fingerprint region and is diagnostic for small biomolecules.

Fourier transform IR spectroscopy measures the vibrations associated with functional groups and highly polar bonds within proteins. In proteins these groups provide a biochemical 'fingerprint' made up of the vibrational features of all contributing components. Vibrational spectra associated with proteins are complex but the common components contributing to the IR spectrum include at least seven amide bands denoted as amide I–amide VII representing different vibrational modes associated with the peptide bond. For most observation on proteins only the first three amide bands denoted as amide I, amide II and amide III are of significant interest. Of these the amide I band is most widely used for secondary structure analyses. Amide I is the result of C=O stretching of the amide group coupled to the bending of the N–H bond and the stretching of the C–N bond. These vibrational modes give IR bands between 1600–1700 cm^{-1} and are sensitive to hydrogen bonding and to the secondary structure environment. Amide I bands centred around 1650 to 1658 cm^{-1} are usually believed to be characteristic of groups located within α helices. Unfortunately, disordered regions of polypeptides as well as turns can also give amide I bands in this region and this can complicate assignment. In contrast β strands give highly diagnostic bands in the region located from 1620–1640 cm^{-1}. Parallel and antiparallel β strands are sometimes distinguishable with antiparallel regions showing a large splitting of the amide I band due to the interactions between strands.

Water (H_2O) is a very strong infrared absorber with prominent bands centred at wavenumbers of ~3800–3900 cm^{-1} (H–O stretching band), 2125 cm^{-1} (water association band) and one at 1640 cm^{-1} (the H–O–H bending vibration) lying in a major spectral region of interest (the conformationally sensitive amide I vibrations). Despite the high s/n ratios, accurate frequency determination, speed and reproducibility of modern instrumentation it is frequently necessary to perform measurements in aqueous solution around 1640 cm^{-1} in D_2O where no strong bands are close to the amide I frequency. Measurements performed on the amide I band in D_2O and hence on deuterium substituted proteins are often distinguished as amide I′. One advantage IR spectroscopy shares with CD is that the technique is readily performed on relatively small amounts of material (often ~100 μl) and represents a good starting point for new structural investigation of a protein. For both methods increasingly successful attempts to extract quantitative information on protein secondary structure have been made. In IR spectroscopy analysis of the amide I bands has proved successful and has permitted deconvolution of spectra into components attributable to helical, strand and turn elements of secondary structure. As a result IR like CD provides a means of estimating the secondary structure present in proteins.

Raman spectroscopy

Raman spectroscopy is concerned with a change of frequency of light when it is scattered by molecules. The Raman technique owes its name to Chandrasekhara Venkata Raman who discovered in 1928 that light interacts with molecules via absorbance, transmission *or* scattering. Scattering can occur at the same wavelength when it is known as Rayleigh scattering or it can occur at altered frequency when it is the Raman effect (see Figure 10.52).

Raman's discovery was that when monochromatic light is used to irradiate a sample the spectrum of scattered light shows a pattern of lines of shifted frequency known as the Raman spectrum. The shifts are independent of the excitation wavelength and are

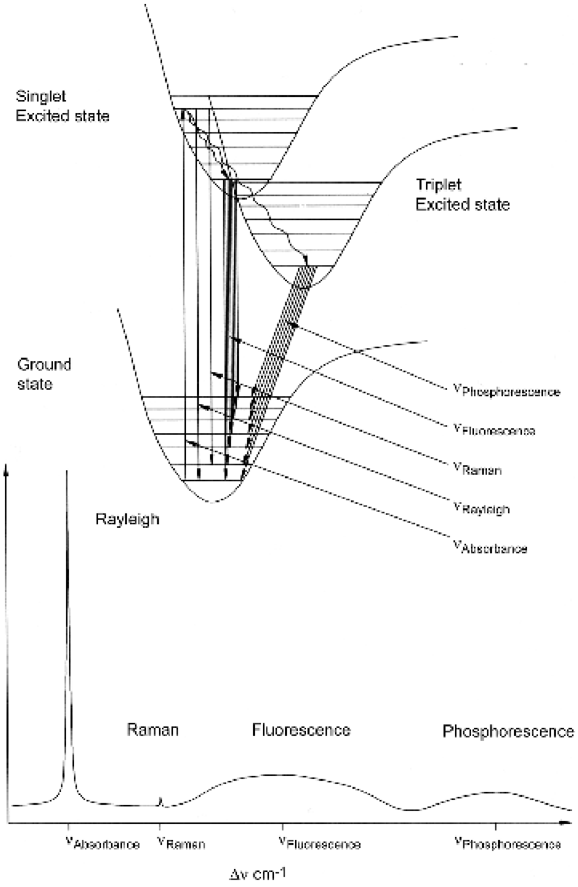


Figure 10.52 Relationship between Raman, Rayleigh, absorbance and fluorescence (adapted from Wang, Y. & van Wart H.E. *Methods Enzymol.* 1993, **226**, 319–373. Academic Press)

characteristic for the molecule under investigation. Collision of a photon with a molecule leads to elastic and inelastic scattering; the former contributes to the Rayleigh line whilst inelastic collisions cause quantum transitions of lower and higher frequency and are responsible for the negative and positive shifts observed in Raman spectroscopy corresponding to vibrational and rotational transitions within scattering molecules. The two types of Raman transitions are sometimes called Stokes lines. These changes are normally observed directly in the IR region of the spectrum but they can also be observed as frequency shifts within the more accessible UV or visible regions. Raman signals from aqueous protein samples are easily measured with signals from water being relatively weak. A much greater problem derives from fluorescence associated with proteins that swamps smaller Raman signals.

A variation on the basic technique of Raman spectroscopy is called resonance Raman spectroscopy. The technique is often applied to metalloproteins where electronic transitions associated with chromophore ligands to metal centres produce enhancements to the basic Raman effect. Vibrational modes associated with the ligand are dramatically enhanced often by factors of 10^3 or more and have proved extremely useful in examining the structural organization of metal centres within proteins. Since these centres are often found at the catalytic core of an enzyme resonance Raman has facilitated understanding the chemistry occurring within these sites. The techniques of fluorescence, Raman and absorbance are related via excitation of an electron to an excited state followed by its decay back to the ground state (Figure 10.52).

ESR and ENDOR

Electron spin resonance (ESR) spectroscopy measures a comparable process to NMR spectroscopy with the exception that changes in *electron* magnetic moment are measured as opposed to changes in nuclear spin properties. Despite this difference much of the basic theory of NMR applies to ESR spectroscopy.[4] Electrons possess a property called spin that leads to the generation of a magnetic moment when an external magnetic field is applied. Spin is quantized with an electron spin quantum number $M_s = \pm 1/2$. In the presence of an applied magnetic field degeneracy is removed and leads to transitions between the two states (Figure 10.53). The resonance condition is expressed as

$$\Delta E = h\nu = g\beta H \qquad (10.49)$$

where g is a dimensionless constant (often called the Lande g factor and is equal to 2.00023) and β is the

[4] The terms ESR and EPR are used interchangeably and both terms are acceptable and used by authors in the subject literature. EPR = electron paramagnetic resonance.

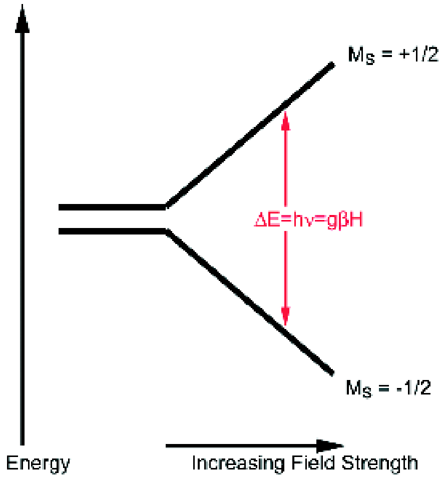


Figure 10.53 The resonance condition for ESR spectroscopy. Such a transition would give rise to a single line. For instrumental reasons ESR spectra are presented as first derivative spectra

Bohr magneton. The electron magnetic moment is 600 times greater than that of the proton. ESR experiments are normally carried out at a field strength of 3400 G (0.34 T) and a frequency of ~9 GHz (known as the X-band) that lies in the microwave region of the electromagnetic spectrum. In ESR the frequency is usually fixed and spectra are acquired by sweeping the applied magnetic field whilst measuring concomitant changes in resonance absorption.

A fundamental requirement for ESR is the presence of one or more unpaired electrons. In proteins this occurs frequently for transition metals such as Fe, Co, Mn, Ni and Cu with unpaired d electrons. Routinely ESR spectra are recorded at low temperatures, usually below 50 K, through the use of liquid helium cooled probes. Rapid spin lattice relaxation rates at elevated temperatures are slowed at lower temperatures leading to increased signal intensity and less saturation. Aqueous samples absorb microwaves strongly and freezing results in the formation of ice a less efficient microwave absorber.

Iron exists in three different oxidation states in biological systems. Usually the ferrous (d^6) and ferric (d^5) configurations are encountered but occasionally enzymes such as peroxidases or cytochromes P450 have a ferryl (d^4) state. The ferric iron form occurs as an unpaired high spin ($S = 5/2$) state or a partially paired low spin ($S = 1/2$) state that are distinguished using ESR spectrometry. In ferredoxins, metalloproteins containing iron–sulfur clusters, ESR has revealed a variety of different clusters {[Fe-Cys4], [2Fe-2S], [3Fe-3S] and [4Fe-4S] centres}, interacting redox centres, redox state transitions, as well as ligand identity, coordination and geometry (Figure 10.54).

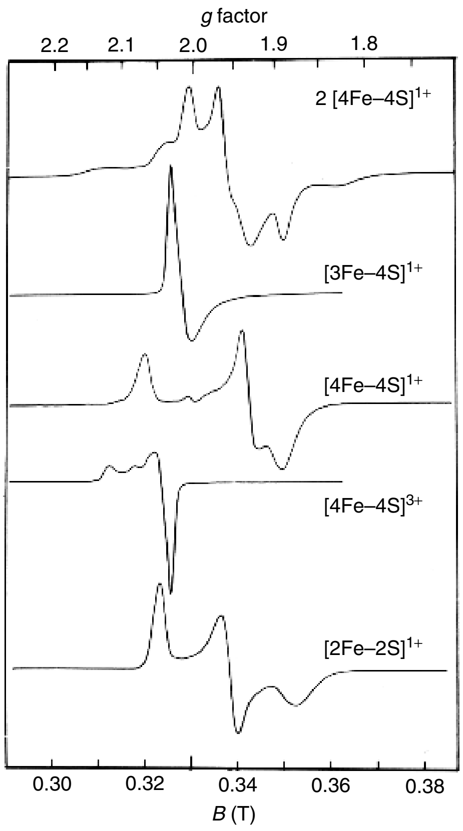


Figure 10.54 Representative ESR spectra for iron–sulfur clusters found in ferredoxins. From top to bottom the spectra reveal the properties of the iron centres in proteins from *Clostridium pasteurianum*, *Desulfovibrio gigas*, *Bacillus stearothermophilus*, *Chromatium vinosum* high potential protein and *Mastogocladus laminosus*. All spectra were recorded at X-band frequencies at a temperature between 10 and 20 K

Table 10.13 International and national facilities for X-ray crystallography and NMR spectroscopy

Facility and role	Web address
Synchrotron sources	
Daresbury UK. Crystallography and CD	http://www.srs.ac.uk/srs/
Cornell USA. Crystallography	http://www.chess.cornell.edu/
European facility Grenoble France. X-ray and Neutron diffraction	http://www.esrf.fr/
NMR facilities	
Madison USA. Ultra high field NMR spectroscopy	http://www.nmrfam.wisc.edu/
Frankfurt Germany. EU funded large scale facility for NMR and ESR	http://www.biozentrum.uni-frankfurt.de/LSF/

Although ESR spectroscopy has been most widely used to characterize iron-containing proteins it has also been used to study: (i) the metal centre of Ni^{2+} (d^8) in enzymes like urease, hydrogenases and carbon monoxide dehydrogenase; (ii) the wide variety of Cu^{2+} (d^9) centres such as those found in superoxide dismutase, tyrosinase and haemocyanin; (iii) Mo centres such as that found in enzymes like xanthine oxidase; and (iv) Mn centres found within photosynthetic systems involved in oxygen evolution.

Recent developments in ESR spectroscopy utilize advantages of pulse methods for acquiring spectra as opposed to sweeping the magnetic field. ESR spectroscopy has traditionally been a continuous wave (CW) technique but techniques such as ESEEM (electron spin echo envelope modulation) or ENDOR (electron-nuclear double resonance spectroscopy) probe electron–nuclear interactions in metalloproteins unsuitable for study using crystallography or NMR. The techniques can derive relaxation times and information on metal–ligand mobility, distance and orientation.

Armed with several of the above spectroscopic tools the structure of almost all proteins can be gradually determined. The time scale from isolation of a protein to determination of its three-dimensional structure has decreased dramatically from several years to perhaps a few months. In many cases acquisition of experimental data takes only a few days whilst analysis and interpretation of the raw data is often a longer process. Unfortunately the cost of instrumentation required to perform these studies at the highest level has become extremely prohibitive. Ultra high field NMR spectrometers, powerful electron microscopes and synchrotron beam lines are beyond the scope of most individual institutions, with a single instrument costing millions of pounds. Increasingly this instrumentation is being organized around core national, and sometimes internationally coordinated centres of excellence where these facilities serve many research groups (Table 10.13).

Summary

By exploiting different regions of the electromagnetic spectrum and the interaction of atoms with radiation considerable information can be acquired on the structure and dynamics of proteins.

The plethora of techniques available for studying proteins means that if one technique is unsuitable there is almost certainly another method that can be applied.

X-ray crystallography and multi-dimensional NMR spectroscopy yield detailed pictures of protein structure at an atomic level. X-ray crystallography relies on the diffraction of X-rays by electron dense atoms constrained within a crystal. Some proteins fail to crystallize and this represents the major limitation of the technique.

The ‘phase problem’ is overcome through the use of isomorphous replacement but a basic requirement is that addition of heavy metal atoms does not alter protein structure. The structures derived for the proteasome, the ribosome and viral capsids serve to emphasize the success of the technique.

Membrane proteins or proteins with extensive hydrophobic domains are notoriously difficult to crystallize, although several structures now exist within the Protein Databank.

NMR spectroscopy measures nuclear spin reorientation in an applied magnetic field most frequently for spin 1/2 nuclei. In proteins this centres around the ^{1}H but with isotopic labelling can also include ^{15}N and ^{13}C nuclei.

A major hurdle in NMR spectroscopy is the assignment problem – identifying which resonances belong to a given residue. Sophisticated multidimensional heteronuclear NMR methods have been developed to facilitate this process based around the interaction of nuclei ‘through bonds’ and ‘through space’.

Solution structure is defined from the use of a combination of torsion angle and distance restraints. The former are obtained from J coupling experiments whilst the r^{-6} dependence of the NOE is the basis of distance constraints.

NMR structures are usually presented as a family or ensemble of closely related protein topologies. Although most structures for proteins determined by NMR spectroscopy have masses below ~20 kDa the size limit is increasing steadily with the introduction of new methods.

The vast majority of structures deposited in the Protein Databank (>95 percent) have been determined using crystallography or NMR spectroscopy, but slowly a third technique of cryo-EM is gaining prominence.

Cryo-EM methods have the advantage of requiring little sample preparation and are particularly suitable to very large protein complexes. The technique relies on electron diffraction by particles immersed in a frozen lattice. Single particle analysis is the current ‘hot’ area and involves trapping macromolecular complexes in random orientations within vitreous ice and combining thousands of images to provide an enhanced picture of the system.

Optical techniques based around electronic transitions provide information on the absorbance and fluorescence of chromophores found in proteins usually aromatic residues. Tryptophan exhibits a significantly greater molar extinction coefficient when compared to either tyrosine or phenylalanine with the result that most of a protein’s absorbance or fluorescence reflects the relative abundance of Trp within a polypeptide chain.

Time-resolved measurements allow changes in fluorescence or absorbance to be followed on very short time scales sometimes in the sub-nanosecond range. This allows measurements of protein mobility such as the motion of aromatic side chains.

CD involves the measurement of differential absorption of right and left circularly polarized light as a function of wavelength. In proteins the far UV region from 260–180 nm is dominated by the CD signals due to elements of secondary structure such as helix, turns and strands.

Proteins composed extensively of helical structure have very different CD spectra to those containing proportionally more β strands. CD spectra allow the secondary structure content of unknown proteins to be predicted with reasonable accuracy.

IR spectroscopy of proteins has been used to measure changes in the vibrational states associated with the amide bond. Amide bonds in helices, turns and strands show characteristic transitions in a region known as the amide I band between 1600 and 1700 cm^{-1}. Experimental spectra are fitted as combinations of helices and strands and can be used to estimate secondary structure content in a comparable manner to CD spectroscopy.

Problems

1. Why are fluorescence maxima observed at longer wavelengths than absorbance maxima?
2. Why do peaks in absorbance spectra become narrower when the temperature is lowered to 77 K?

3. What is the wavelength range of IR studies conducted at 300–4000 cm^{-1}.

4. Show that an axial ratio of 1:100 leads to an observed rotation of circularly polarized light of 0.57°.

5. The absorbance of a 10 µM solution of tryptophan in buffer at 280 nm is 0.06. The buffer alone gives an absorbance of 0.004 at 280 nm. Assuming a path length of 1 cm calculate the extinction coefficient of tryptophan. What are the expected absorbance values for pathlengths of 1, 2 and 20 mm?

6. Using the data provided in Table 10.4 calculate the Larmor frequency of the following nuclei ^{1}H, ^{13}C and ^{15}N at field strengths of 20, 18.1, 16, 11.7 and 8 T.

7. Sketch the high resolution 1D ^{1}H-NMR spectra of alanine, glycine and threonine.

8. Draw the expected pattern of connectivities in 2D homonuclear TOCSY and COSY spectra for the residues not included in Figure 10.27. (ignore fine structure for cross peaks).

9. Explain how heavy metal derivatives of proteins might be detected using non-denaturing electrophoretic techniques. What might be the limitations to this technique? Using Harker diagrams show how the use of a second heavy metal derivative leads to a unique solution for the intensity and phase.

10. The structure of cytochrome c′ from *Chromatium vinosum* has been determined and the data has been deposited in protein databanks. Locate the relevant PDB file and from the primary citation identify the methods used to determine the structure together with the unit cell dimensions, the number of molecules per unit cell, the resolution achieved. Describe the structure determined and any new features discovered.