1

Protein structure

Stephen J. Smerdon
National Institute for Medical Research, London, UK

CHAPTER OUTLINE

1.1 Introduction

Key concepts

- Proteins are large, complex polymers. Their three-dimensional structures dictate their biological function.
- The three-dimensional structure of proteins and their complexes provides a framework that is essential for a full comprehension of their myriad biochemical activities.
- The size and spatial separation of atoms that make up molecular structures is too small to be directly observable by, for example, light microscopy.
- At present, three methods are available for protein structure determination: X-ray crystallography, nuclear magnetic resonance spectroscopy, and electron microscopy.

The word 'protein' derives from the Greek *proteos*, meaning 'first' or 'of first importance.' In the early part of the last century, proteins rather than nucleic acids were widely regarded as the repository of hereditary information—the 'genetic material.' Classic biochemical experiments that disproved this are now the stuff of biological folklore and are described on this site and elsewhere. Over the past 75 years or so these early misconceptions have been replaced by an appreciation of the importance of proteins as the 'molecular workhorses' of the cell.

In 1926, James Sumner showed that urease from jack beans could be highly purified by crystallization, enabling him to demonstrate that enzymes were proteins. Nevertheless, at the time proteins were still thought of as heterogeneous substances with random structure. This dogma was challenged in 1934 when John Desmond Bernal and a graduate student, Dorothy Crowfoot, demonstrated that a crystal of the proteolytic enzyme pepsin produced a pattern of discrete diffraction spots on a film when exposed to a beam of X-rays. This experiment showed unequivocally that proteins possess an ordered, well-defined arrangement of atoms, and the field of structural biology was born.

Proteins are a diverse class of biological polymers that play an extraordinary variety of functional roles. In the form of enzymes, proteins catalyze most of the chemical reactions that take place in the cell. Protein function is not, however, limited to chemical catalysis. For example, interactions between protein hormones and receptors are responsible for the transmission of many developmental and physiological signals and represent just one of many activities that are mediated through highly specific protein binding events.

A chemist can utilize an almost limitless variety of conditions to increase the efficiency of chemical reactions in the laboratory. In contrast, synthetic (anabolic) and degradative (catabolic) processes, along with the host of specific, high-affinity interactions necessary for life, must occur in a largely aqueous environment and within a rather narrow range of temperatures. To a great extent, these constraints have driven the evolution of the large and complex protein molecules that we observe in living organisms. Clearly, biological activity derives directly from the relative spatial arrangement of the atoms and chemical groups from which proteins are constructed. For this reason, biological mechanisms can only be truly understood in the light of the three-dimensional atomic structure of the macromolecules involved. This chapter focuses on how our understanding of these fundamental molecular processes has evolved through the elucidation and analysis of the three-dimensional structure of proteins.

The primary goal of all structural techniques is the determination of the precise spatial relationship between each and every atom in the molecule of interest. In this respect it is important to recall that the chemical bonds between atoms within a protein are of the order of 10^{-10} m. Optical theory shows that in order to 'resolve' two objects, we must illuminate them with radiation of a wavelength that is no longer than about twice the distance between them. Given that the wavelengths of the visible electromagnetic spectrum are between ~400 and 800 nm, it is clear that light microscopy is not useful when investigating objects as small as proteins, and thus other methods must be employed.

At present, X-ray crystallography and nuclear magnetic resonance (NMR) are the only available techniques for the determination of macromolecular structures at high resolution. Significant advances in other areas, however—particularly electron microscopy—are providing important structural information in ever-increasing detail. A thorough treatment of the theoretical background to these methods is beyond the scope of this chapter and the interested reader is provided with references to a number of excellent textbooks, review articles, and online information. In the following three sections, the aim is instead to provide a brief historical background along with sufficient technical information to guide the reader through the various examples provided in the following sections, and to describe the technical advances in currently popular structural techniques that have resulted in the recent explosion of structural information.

The first protein structure, that of myoglobin, was reported by John Kendrew and coworkers in the late 1950s. Since then, the number of structures determined each year has increased exponentially. This expansion of structural information has occurred in parallel with, and as a result of, advances in the fields of molecular biology and physics (*Section 1.2, X-ray crystallography and structural biology*). In 1971, the Protein Data Bank (PDB) was established as an international repository for structural data. At present, a total of ~40,000 structures have been deposited and structures are currently being determined at the rate of ~5000 to 6000 per year (**FIGURE 1.1**).

Structural methods are increasingly being incorporated into the pantheon of routine but powerful methodologies that can be brought to bear on an experimental system. Furthermore, the scope of biological questions that can be asked has been fundamentally changed. The new field of *structural genomics* (*Section 1.16, What's next? Structural biology in the postgenomic era*) has emerged with the goal of determining the structures of all proteins from a number of target organisms ranging from simple prokaryotes to humans. Given that the human genome encodes upward of 30,000 proteins, this undertaking is ambitious. If successful, though, the benefits to basic biological science and to medicine (*Section 1.15, Structure and medicine*) could be considerable.

1.2 X-ray crystallography and structural biology

Key concepts

- At present, X-ray crystallography is the primary method for investigating macromolecular structure at atomic resolution.
- Diffraction from a crystal produces a diffraction pattern that can be related to the electron densities of each atom in the molecule by a Fourier transform.
- Phase information crucial for reconstructing an image of the molecule within the crystal is lost in the diffraction experiment but can be recovered by techniques of isomorphous replacement, molecular replacement, and anomalous scattering.

X-ray crystallography is, by far, the most effective and widely employed method for high-resolution structure determination. In light of the size, flexibility, and complexity of proteins, which will become apparent in later sections, it is perhaps amazing that these molecules can be enticed to form highly ordered three-dimensional crystalline arrays that are the first basic requirement of the method. In fact, this phenomenon was first documented in 1847 by the embryologist Karl Reichart, who observed crys-

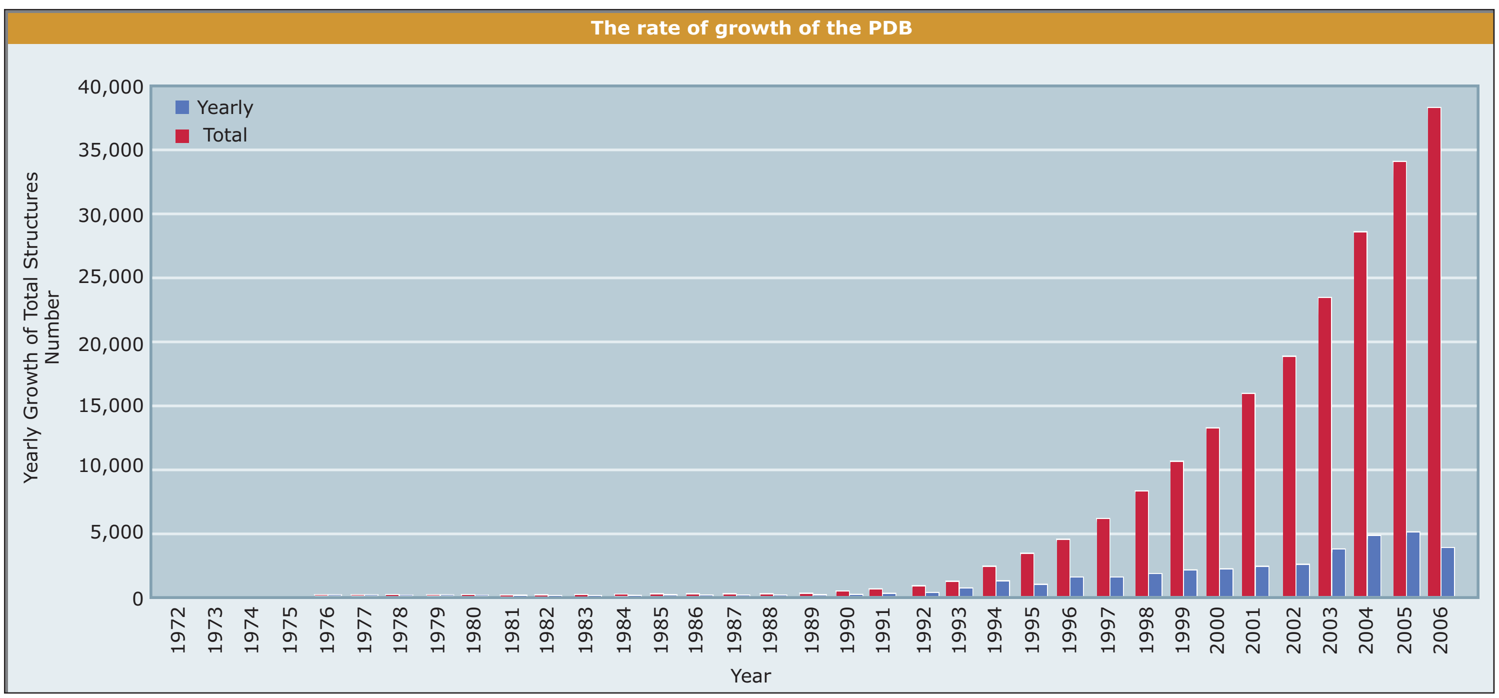

FIGURE 1.1 The rate of growth of the Protein Data Bank.

tallization of the oxygen transporter hemoglobin, a molecule that was to play a central role in the development of modern X-ray crystallographic methods. On December 28, 1895, Wilhelm Conrad Roentgen gave his preliminary report, entitled *Über eine neue Art von Strahlen (On a New Kind of Rays)*, to the president of the Wurzburg Physical-Medical Society, accompanied by experimental radiographs of his wife's hand. Not everyone shared Roentgen's enthusiasm for his discovery, particularly the mathematician and engineer Lord Kelvin who, in 1897, infamously pronounced X-rays to be an 'elaborate hoax'—certainly not the highlight of an illustrious career!

X-rays have extremely short wavelengths of the order of one **Ångstrom** (Å) and in 1912, the physicist Max von Laue suggested that they might be used to investigate the atomic structure within crystals of small molecules. Lawrence Bragg was able to demonstrate X-ray diffraction by crystals of sodium chloride and solve its crystal structure. Diffraction of X-rays occurs as a result of the scattering of X-rays by the electrons that orbit each atom. Bragg formalized the diffraction of X-rays in terms of the reflection of incident radiation from imaginary planes of atoms that result from their periodic arrangement in a crystal. For this reason, diffraction 'spots' are now known as 'reflections' and the famous equation he derived ($n\lambda = 2d\sin\theta$) is known as Bragg's Law (FIGURE 1.2).

Crystals are periodically repeating arrays (Figure 1.4), and as a result the pattern and the relative intensities of diffracted spots are related to the underlying arrangement of atoms by a mathematical summation known as the **Fourier transform** (FIGURE 1.3). The intensities of the reflections are simply the Fourier transform of the electron density around each atom; the pattern that they form on the detector (photographic film in the case of Bragg) is the transform of the crystalline lattice within which the atoms are arranged. The observed diffraction pattern is a product (strictly a convolution) of the two. Importantly, this dictates that each atom in the crystal contributes, to a greater or lesser extent, to every reflection. In order to calculate the electron density within the crystal, and thus the atomic positions, the relative phase angles of each reflection resulting from the constructive interference between scattered X-rays must be known. These are, however, completely lost in the experiment. This is a fundamental difference between diffraction and microscopy, where phase information is preserved through the use of lenses that are not available for very short wavelength X-rays. This information loss is known as the 'phase problem.'

As mentioned earlier, in 1934 Bernal and Crowfoot had shown that protein crystals, like small inorganic compounds, had sufficient internal order to diffract X-rays. Nevertheless, these and subsequent experiments dramatically illustrated the technical problems of investigating molecules of the size of proteins by diffraction methods. Crystals are formed from basic repeating motifs or unit cells that are related to each other by translation only (FIGURE 1.4). Within a unit cell, the individual molecules (or

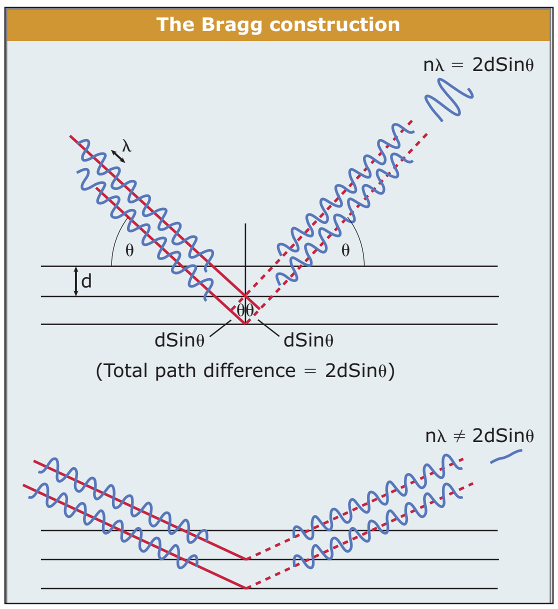

FIGURE 1.2 The Bragg construction.

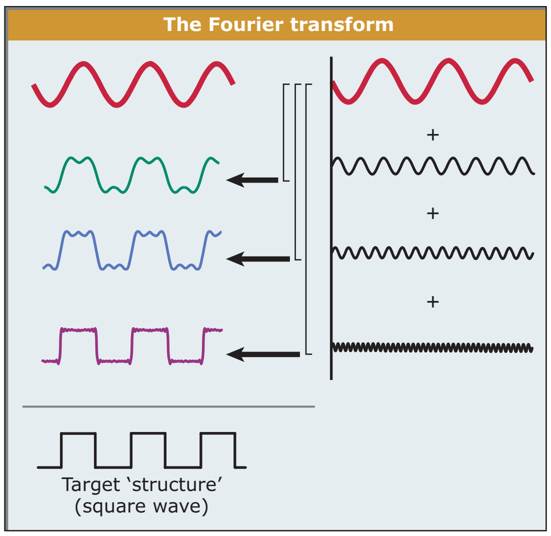

FIGURE 1.3 The Fourier transform. The addition of waves of different phase, frequency, and amplitude results in improved approximation of the square wave.

groups thereof) that constitute the crystal's asymmetric unit are arranged according to a total of 230 possible symmetries or space-groups that are described in terms of different rotational (two-, three-, four-, or six-fold rotations), translational, and mirror operations. In fact, as we will see, proteins are chiral and for this reason can only take a subset of these, representing a total of 65 available space-group symmetries.

Only the size of the unit cell decrees the number of X-ray reflections that can be observed at any given resolution. X-ray data from protein crystals may involve the measurement of tens of thousands of reflections, even for a small protein crystallized in a low symmetry space-group (**FIGURE 1.5**). The resolution of a structure is directly related to the level of accuracy at which atomic positions are known. From Bragg's Law, we know that the more finely the unit cell is sampled, the closer together the Bragg planes become. At smaller *d*-spacings, the Bragg requirement that, for a spot to be observed, the total path difference must be an integral number of wavelengths (Figure 1.2) will only be fulfilled at progressively higher values of θ. Thus the resolution is defined by the minimum value of d (in Å units) for which reflections are represented in the final set of diffraction data. In terms of the Fourier transform, the higher-resolution reflections are those that contribute the highest-frequency terms in the summation, and therefore contribute the most detailed structural information. **FIGURE 1.6** shows how the final calculated electron density varies with data resolution. Initial estimates of the phase angle for each reflection are generally poor. Remember, however, that the diffraction pattern is a Fourier transform of the contents of the asymmetric unit, and therefore we can calculate a theoretical diffraction pattern once we know the locations of the atoms in the crystal. This is the basis of crystallographic refinement, where the atomic model is adjusted using molecular graphics and computational procedures (*Section 1.5, Protein structure representations—A primer*) so as to maximize the agreement of the calculated pattern with the experimentally observed diffraction data.

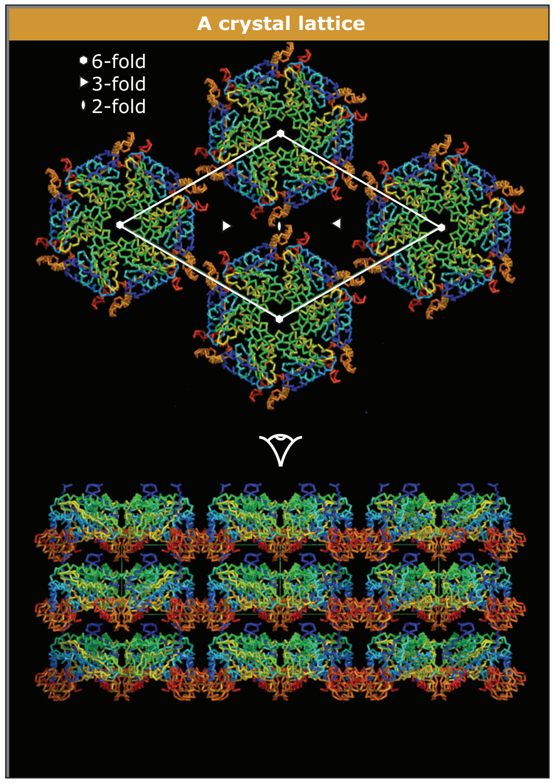

FIGURE 1.4 A crystal lattice.

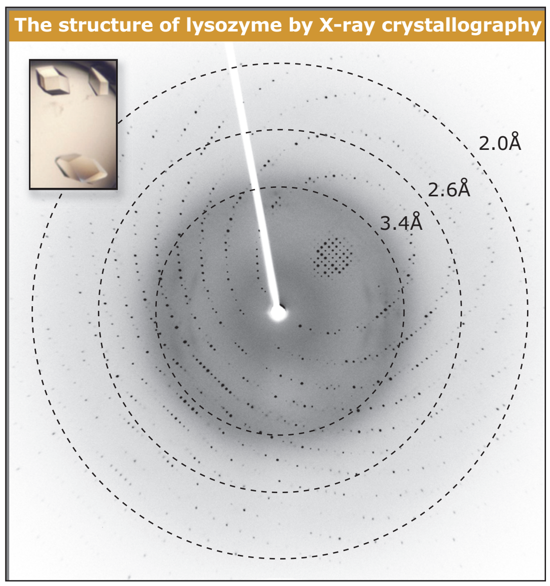

FIGURE 1.5 A part of the diffraction pattern obtained from crystals of hen egg-white lysozyme (inset), the first enzyme structure to be solved by X-ray crystallography. Circles show different limits of resolution.

Such were the technical difficulties that the first structure determination of a protein by X-ray crystallography did not happen for more than 20 years. The structure of myoglobin, a small molecule of 153 amino acids that acts as a store of molecular oxygen, was truly a revelation, showing for the first time many of the fundamental architectural principles of protein structure that we now take for granted. The crucial technical advances that enabled this extraordinary achievement, however, were developed on a related but much larger molecule, hemoglobin, which had been first crystallized in the early part of the nineteenth century. Vernon Ingram (working with Max Perutz, who had already been laboring for many years on

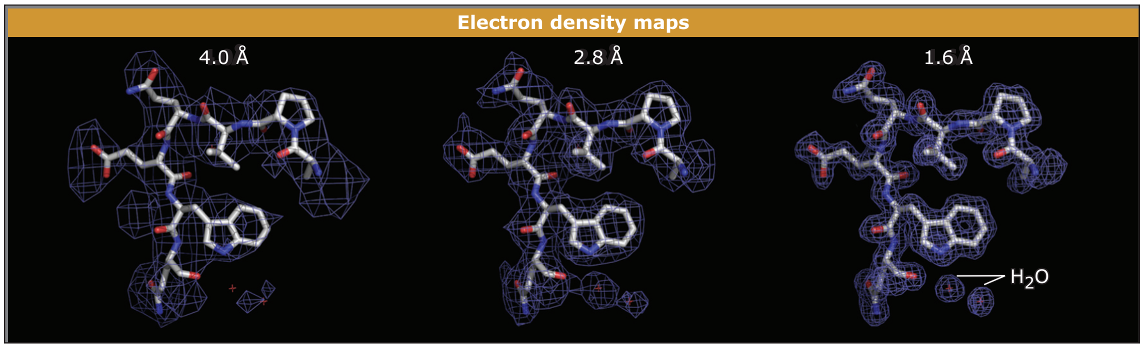

FIGURE 1.6 Electron density maps at increasingly high resolution.

the hemoglobin problem) was able to soak crystals of the protein in dilute solutions of heavy (i.e., electron dense) metal salts and collect X-ray data from them. In most cases this treatment resulted in large changes in the pattern of spots or a complete loss of diffraction. Occasionally, though, the data collected were similar (or **isomorphous**) enough with those derived from unsoaked, *native* crystals that further analysis was possible. Perutz used a mathematical procedure related to the Fourier transform called a **Patterson function** to determine the positions of bound heavy atoms in the derivatized crystal. Isomorphism is important here because the success of the procedure relies on the fact that the differences in the Patterson function of the underivatized and derivatized crystal data derive solely from the addition of the heavy atom. From this knowledge of the heavy atom positions, an approximate phase angle for each reflection could be calculated, enabling a map of the electron density of the hemoglobin molecule to be produced at 5.5 Å resolution. The smaller size of myoglobin (~17 kDa, compared with hemoglobin at~ 65 kDa) meant that John Kendrew, working in the same department as Ingram, was able to solve its structure well before the first structure of hemoglobin was published in the early 1960s.

The contributions of both Perutz and Kendrew were acknowledged in 1962 by the award of the Nobel Prize for Chemistry.

In total, the structure of hemoglobin had taken some 30 years to solve. How is it that structures are now being determined at a rate of more than 5000 per year? By and large, this phenomenal increase in speed can be attributed to four major developments that have gradually come into common use over the last 10 to 15 years: recombinant DNA technology; cryo-crystallography; multiwavelength anomalous scattering methods; and the availability of high-brightness, tuneable synchrotron X-ray sources.

The extent of the improvements in the X-ray crystallographic method have meant that the rate-limiting steps in structure determination are now the ability to grow well-diffracting crystals, and therefore, the availability of *crystallizable* samples. The considerable difficulties associated with the need to purify potentially scarce proteins from the cells/tissues within which they naturally reside have been largely removed by recombinant DNA technology. It is now possible to produce tens or hundreds of milligrams of highly pure protein by expressing its cloned gene in a variety of host cells, most commonly the bacterium *Escherichia coli*, but also in cultured yeast, insect, and mammalian cells. Crystallization can now be performed automatically with a variety of commercially available crystallization 'screens' and robotic liquid-handling devices.

Having produced well-diffracting single crystals, data can now be collected at a number of high-intensity synchrotron radiation sources around the world (**FIGURE 1.7**). These large installations produce hard (i.e., short wavelength) X-rays as a by-product of accelerating packets of electrons at velocities approaching the speed of light, in a circular orbit with a diameter measured in the hundreds of meters. As the electrons are forced to follow a circular path under the influence of a high magnetic field, energy is lost as electromagnetic radiation at wavelengths ranging from γ-rays into the ultraviolet region. Using sophisticated optics, a beam of X-rays, with an intensity that may exceed that available from a laboratory source by several orders of magnitude, can be focused onto a protein crystal with great precision. This, in combination with modern electronic CCD (charge-coupled device) detectors in place of X-ray film, results in a considerable reduction in the time required to collect diffraction data. Complete data sets that would otherwise require days to collect can now be measured in a matter of minutes.

Unfortunately, the use of radiation of this intensity (up to 10^{14} X-ray photons/mm^2 or more) would destroy many protein crystals in a few seconds due to effects of localized heating and the production of chemically reactive free radicals. Fortunately, radiation damage can be largely eliminated by preserving crystals at liquid nitrogen temperatures (100 K or −173°C) during exposure to the X-ray beam.

In cases where the structure of a homologous protein is available, a structure solution can be achieved using the technique of molecular replacement. This method attempts to place a known structure (the search model) into the crystal of the unknown protein by comparison of the Patterson functions calculated from the search structure and the target diffraction data. Here, the rotational orientation and the translation of the search model that best fit the observed diffraction data are determined and applied, providing an approximate starting structure for model building and crystallographic refinement. Obviously, in many cases, homologous structures may not be available. However, a combination of the use of recombinant DNA technology, cryo-crystallography, and synchrotron radiation sources has enabled the phase problem to be directly solved rather trivially using a technique known as multiwavelength anomalous diffraction (MAD). This method derives from the fact that, at characteristic wavelengths, chemical elements interact with X-rays in such a way that the resultant scattered wave gains a shift in its phase.

Although laboratory X-ray sources are limited to X-rays of a single, fixed wavelength, synchrotron radiation can be 'tuned' to supply X-rays at different, but well-defined, wavelengths over a useful range of 0.5 to 2.5 Å. In 1990, Wayne Hendrickson and colleagues showed that phases could be determined directly from a single crystal by exploiting the anomalous scattering of selenium atoms introduced by expressing a recombinant protein in bacteria grown on broth containing selenomethionine as the sole source of methionine. Anomalous scattering from sulfur atoms found in the amino acid cysteine had been used previously to determine the structure of a small protein crambin. Until recently, this was not considered to be a generally applicable approach because the anomalous scattering effect is rather small for sulfur. Nevertheless, the experiment was successful largely due to the high degree of order for the crambin crystals that enabled extremely accurate data to be collected at very high resolution. Selenium is a much more effective anomalous scatterer, and the ability to produce derivatized protein straightforwardly considerably facilitated the determination of phases and thus structure determination. In practice, the experiment involves the collection of data sets at and around the wavelength that corresponds to the peak of the anomalous differences. Although the theoretical details of the method are beyond the scope of this chapter, it can be conceptualized as a kind of isomorphous replacement experiment in which the necessary intensity differences are produced by physics (i.e., variable wavelength X-rays) rather than chemistry (i.e., the addition of heavy atoms). All data are collected from one crystal, and as a result the problems of lack of isomorphism that frustrated crystallographers for so long are effectively removed to the point that in a favorable case, protein structures can now be determined in a matter of hours.

A modern synchrotron source

FIGURE 1.7 A modern synchrotron source. Courtesy of the Diamond Light Source, UK.

1.3 Nuclear magnetic resonance

Key concepts

- NMR is a powerful method for investigating structures of proteins and their complexes in solution.
- NMR methodologies can provide detailed information about macromolecular dynamics that are difficult or impossible to extract from X-ray crystallographic data.

As mentioned earlier in this chapter, X-ray crystallography is not the only means of determining the structures of proteins, and over the last 20 years or so enormous progress has been made in the use of NMR to examine biomolecular structures *in solution*, and at resolutions comparable to those derived for the more 'traditional' X-ray diffraction techniques.

The phenomenon of nuclear magnetic resonance was predicted by quantum mechanics before its first experimental observation by Isidor Isaac Rabir in the 1930s, and subsequently in solution by Felix Bloch and Edward Mill Purcell in the 1940s. The physics underlying the technique involves the realization that many atomic nuclei (those with spin quantum number > 0) possess a magnetic moment as a consequence of possessing both spin and charge. When placed in an external magnetic field, the magnetic moment adopts one of a fixed number of orientations, as its behavior is quantized. In biomolecular studies, we are concerned almost exclusively with nuclei (^{1}H, ^{13}C, ^{15}N) that are spin-$^1/_2$, and adopt two possible orientations that correspond to low and high energy states. Irradiation with electromagnetic radiation of appropriate wavelength leads to transitions between these states, giving an absorption spectrum. NMR transitions lie in the radiofrequency region of the spectrum, with wavelengths in the MHz range. The experimental realization of nuclear magnetic resonance was recognized by the award of Nobel Prizes for Physics to Rabir (1944) and to Bloch and Purcell (1952).

First viewed as a physicist's tool for extracting the magnitudes of the magnetic moments for atomic nuclei, NMR soon became an indispensable technique for chemists, following the observation that the exact resonance frequency of a nucleus was exquisitely sensitive to its local chemical environment. For example, when the NMR spectrum of ethanol (CH_3CH_2OH) is recorded at sufficiently high resolution (**FIGURE 1.8**), separate signals can be seen for each of the three chemically distinct types of proton (hydrogen nucleus) present. Each compound therefore gives a characteristic NMR "fingerprint," making NMR an invaluable analytical tool in chemical investigations. In addition to the different resonance positions ("**chemical shifts**"), NMR signals contain fine structure, arising from through-bond communication between the magnetic moments ("J-coupling"). Analysis of these two effects allows the different signals to be attributed, or *assigned*, to the hydrogen type (i.e., the C**H**$_3$, C**H**$_2$, or O**H** moiety).

The utility of NMR spectroscopy in protein structure investigations was not immediately obvious. The low energies of the NMR transitions render it an insensitive technique requiring large quantities of sample, and the complexity of the NMR spectrum of even a small protein, with say 500 hydrogen types, was considered intractable. The 800 MHz NMR spectrum of the 14 kDa protein lysozyme (Figure 1.8) contains several hundred peaks.

Several technical and methodological advances were central to the development of NMR as a tool to investigate the structures of biomolecules. Richard Ernst introduced Fourier transform NMR, which increased the sensitivity of the technique by orders of magnitude. Ernst also addressed the issue of complexity in his introduction of multidimensional NMR methods, which spread the signals out into a second (or higher) frequency dimension. An example of such a two-dimensional spectrum can be seen in **FIGURE 1.9**. The conventional one-dimensional spectrum lies along the diagonal of the two-dimensional spectrum. The off-diagonal peaks represent correlations between different hydrogen types. In this instance the correlations are a result of the **nuclear Overhauser effect** (NOE), and they signify that the protons sharing the correlation lie within ~5 Å of each other in the tertiary structure. In addition, the sensitivity and resolution problems have been lessened somewhat by the availability of ever-increasing external magnetic field strengths, from a few tenths of a Tesla produced by a permanent magnet in the early days of NMR up to

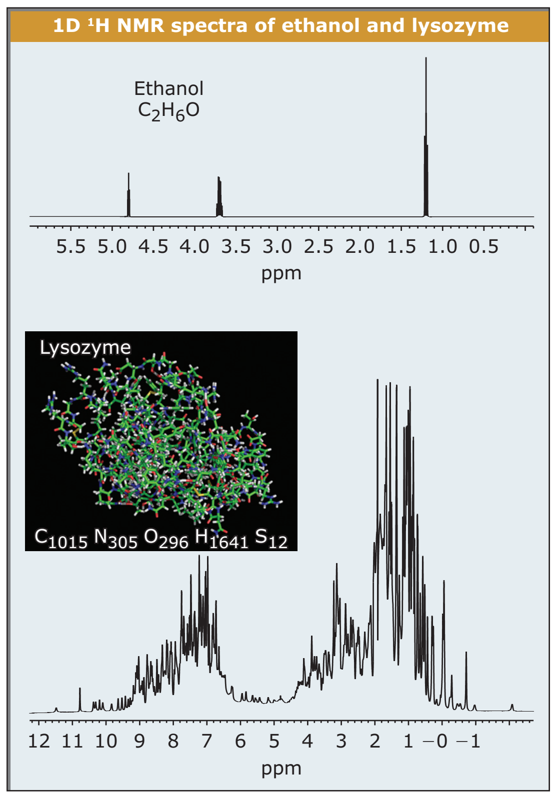

FIGURE 1.8 One-dimensional ^{1}H NMR spectra of ethanol (top) and a small protein lysozyme (bottom).

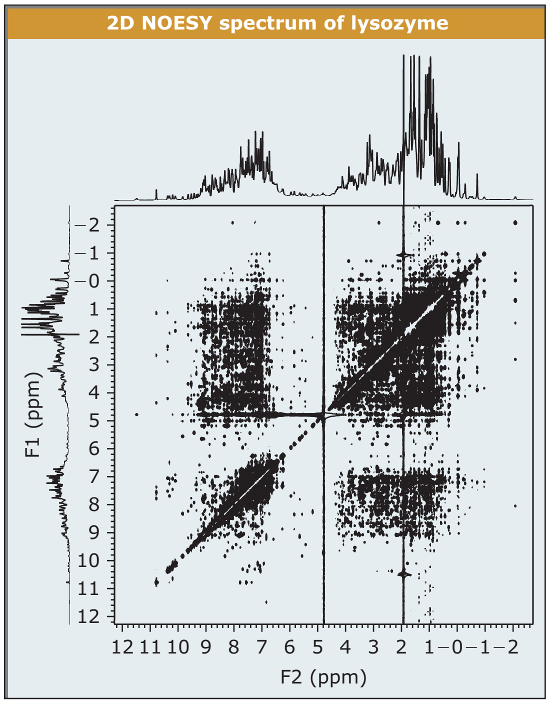

FIGURE 1.9 Two-dimensional NOESY spectrum of lysozyme collected at highfield strength (800 MHz), with the one-dimensional spectrum for each separate dimension shown alongside and above the x and y axes.

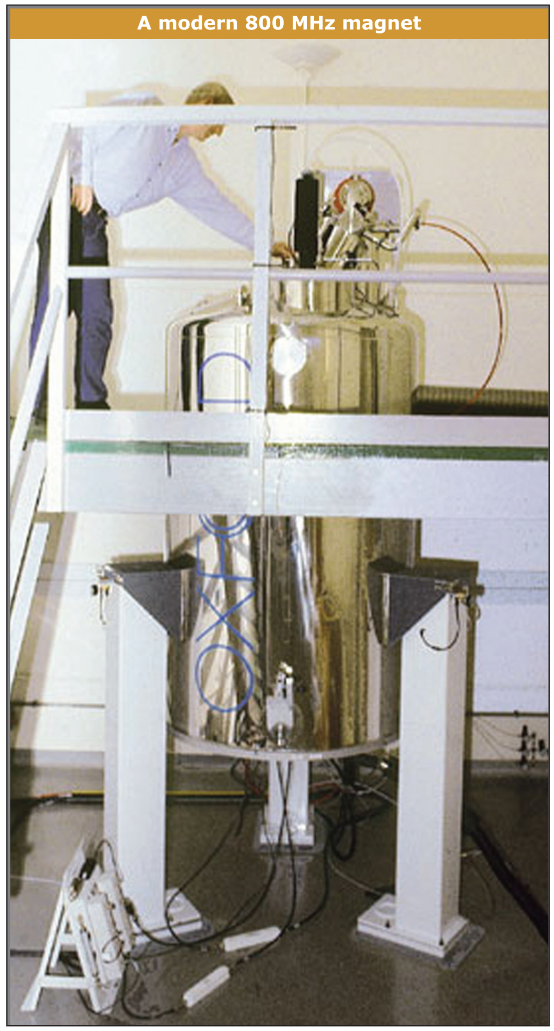

FIGURE 1.10 A modern 800 MHz magnet.

the 20T and larger fields accessible using superconducting magnets today (**FIGURE 1.10**). Ernst's contributions to the development of NMR spectroscopy were recognized with the award of the Nobel Prize for Chemistry in 1991.

The use of these methods in protein structure determination was pioneered by Kurt Wuthrich and coworkers, who used a combination of the **J-coupling** (through-bond) information with 'through-space' information from the NOE to assign the NMR spectrum of proteinase inhibitor IIa. They then went on to use the NOE information to calculate its three-dimensional structure. Initial skepticism was allayed by a blind trial in which the structure of the α-amylase inhibitor tendamistat was solved independently using X-ray crystallography and NMR spectroscopy. Wuthrich's realization of the potential of NMR to solve the three-dimensional structures of proteins, together with his development of methodologies toward this goal, was also acknowledged with the award of the Nobel Prize for Chemistry in 2002.

The use of ^{1}H NMR spectroscopy for the assignment of a protein's spectrum, and the elucidation of its three-dimensional structure, remained a daunting undertaking. This situation was transformed in the early 1990s by employing molecular biology techniques for heterologous protein expression in bacterial hosts. This enabled the production of isotopically labelled protein samples using expression in *Escherichia coli* cultured on a minimal growth medium supplemented with ^{13}C-labeled glucose and ^{15}N ammonium chloride as the sole carbon and nitrogen sources. (The common isotopes of these nuclei, ^{12}C and ^{14}N, are not amenable to study by high-resolution NMR techniques.) This facilitated the development of a huge arsenal of "triple resonance" (^{1}H/^{13}C/^{15}N) NMR methods, notably by Ad Bax and coworkers. These NMR methods permitted much more efficient through-bond communication between nuclei, as one-bond ^{1}H-^{15}N, ^{15}N-^{13}C, and ^{13}C-^{13}C J-couplings could now be used instead of three-bond ^{1}H-^{1}H J-coupling. In addition, they allowed the extension of multidimensional NMR to include ^{13}C and/or ^{15}N frequencies. Hitherto the assignment procedure was predicated on the observation of NOEs between sequential residues—a painstaking process fraught with ambiguity. The use of triple-resonance techniques allows an unambiguous step-by-step journey along the polypeptide backbone.

Armed with a complete or near-complete assignment of the ^{1}H, ^{13}C, and ^{15}N nuclei in the protein, it is possible to extract a huge amount of structural information from various NMR parameters. The two most fruitful sources traditionally have been NOEs (the observation of an NOE between two ^{1}H nuclei, and its magnitude, are constraints on the maximum distance between the nuclei) and coupling constants—the values of three-bond coupling constants, e.g., $^{3}J(H_{N}\text{-}H_{\alpha})$—which are functions of the intervening dihedral angle. These structural constraints, if sufficient in quantity, can be included as additional energy terms, along with known covalent bond lengths and angles, in restrained molecular dynamics protocols with simulated annealing schedules to calculate the structure. The progress of such a calculation is depicted in **FIGURE 1.11**. Owing to the nonexact and possibly incomplete nature of the experimental constraints, the calculation is performed many times. The results are superimposed to give a family of structures (**FIGURE 1.12**), all of which are compatible with the experimental data; this gives some impression of the precision of the structure determination.

Although the many examples of structures determined by both NMR and crystallography show that proteins in solution and in the hydrated crystalline state are, by and large, very similar, NMR has a real and important advantage over X-ray methods in its ability to access the dynamics of biomolecules in the solution state. It is possible to infer dynamics from crystal structures, but the return to equilibrium of NMR signals contains direct information about atomic motions. The relaxation of ^{15}N nuclei has been most widely exploited in this regard. The relaxation of the ^{15}N nucleus is dominated by its attached amide hydrogen nucleus, and the efficiency of this interaction is governed by the rate of reorientation of the $^{1}H\text{-}^{15}N$ bond vector with respect to the external magnetic field. The resulting analysis gives exquisite, residue-specific information on protein dynamics over a wide range of time scales, from picoseconds (fast internal motion of a residue with respect to the protein overall) to nanoseconds (overall tumbling) to micro/milliseconds (conformation exchange processes).

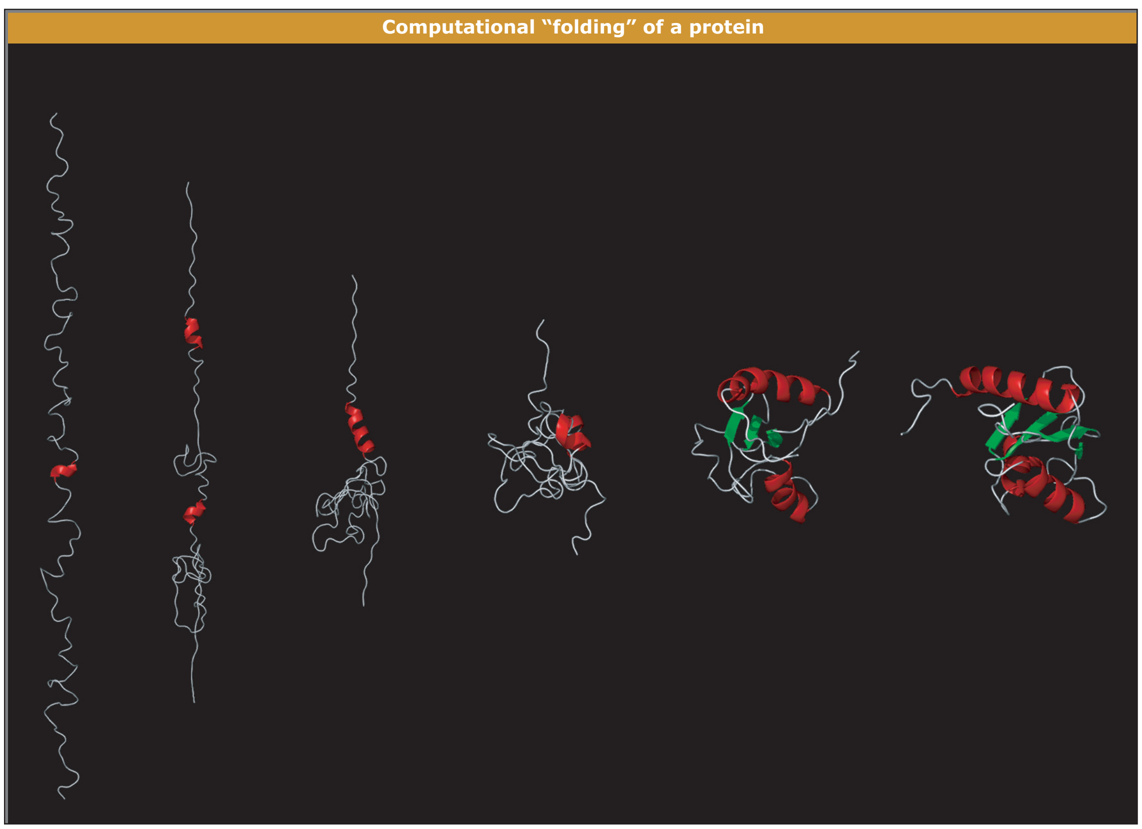

FIGURE 1.11 Computational 'folding' of a protein into a structure consistent with restraints derived from a variety of NMR experiments.

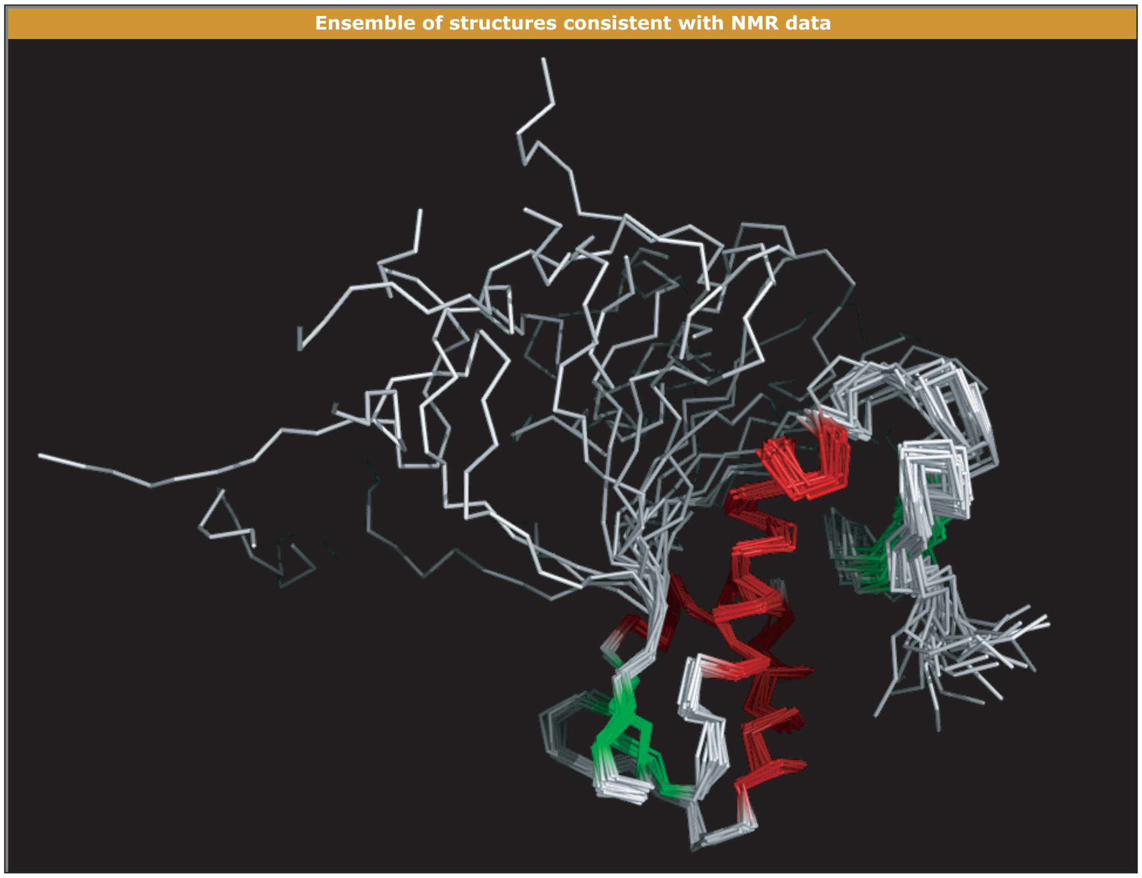

FIGURE 1.12 An ensemble of structures, all of which are consistent with the experimental NMR data. Image generated from Protein Data Bank file IG2K.

A powerful use of NMR spectroscopy is the rapid analysis of protein–protein or protein–ligand interactions. These experiments are most commonly performed using HSQC (heteronuclear single quantum coherence) spectra acquired from ^{15}N-labeled protein. Peaks in the HSQC spectrum are derived from protons attached to nitrogen atoms. Thus there is at least one peak for almost every residue in a protein (from the amide hydrogen of the peptide bond, except proline) and additional peaks for any side-chain NHs. The ^{15}N-HSQC spectrum of a 10 kDa protein is shown in FIGURE 1.13. Addition of an unlabeled binding partner contributes no new peaks, but will affect the peaks corresponding to protein residues at the binding site. These peaks will either shift gradually as the binding partner is added, or gradually disappear and reappear elsewhere in the spectrum, according to the kinetics of the complex in question. Mapping of the perturbed residue positions on the overall structure can then provide an excellent picture of the interaction surface, without the need to laboriously determine the structure of the complex. Quantitative analysis of the pattern of shifts may allow determination of dissociation constants. This approach is also used extensively to screen for binding in a drug discovery context. Analogously, conducting a pH titration of the protein sample allows residue-specific pKa values to be determined—information difficult to obtain by other methods.

When compared with X-ray crystallography, NMR spectroscopy of proteins remains an immature discipline, in which significant methodological and technical advances are still common. Recent years have seen the introduction of new types of structural constraints (residual dipolar couplings) to augment those conventionally employed. Transverse relaxation optimized spectroscopy (TROSY) has permitted the extension of conventional methods to the study of proteins and their complexes up to molecular weights of ~100 kDa, and the introduction of cryogenic detector electronics has yielded a gain in sensitivity of approximately threefold. Structural genomics initiatives have prompted new, faster data acquisition schemes and ever-increasing automation of the data analysis and structure calculation procedures. Most recently, advances in NMR methodology (CRINEPT, CRIPT) have provided some insight into structural aspects of systems as large as the GroEL/GroES system with a molecular weight of ~1 MDa.

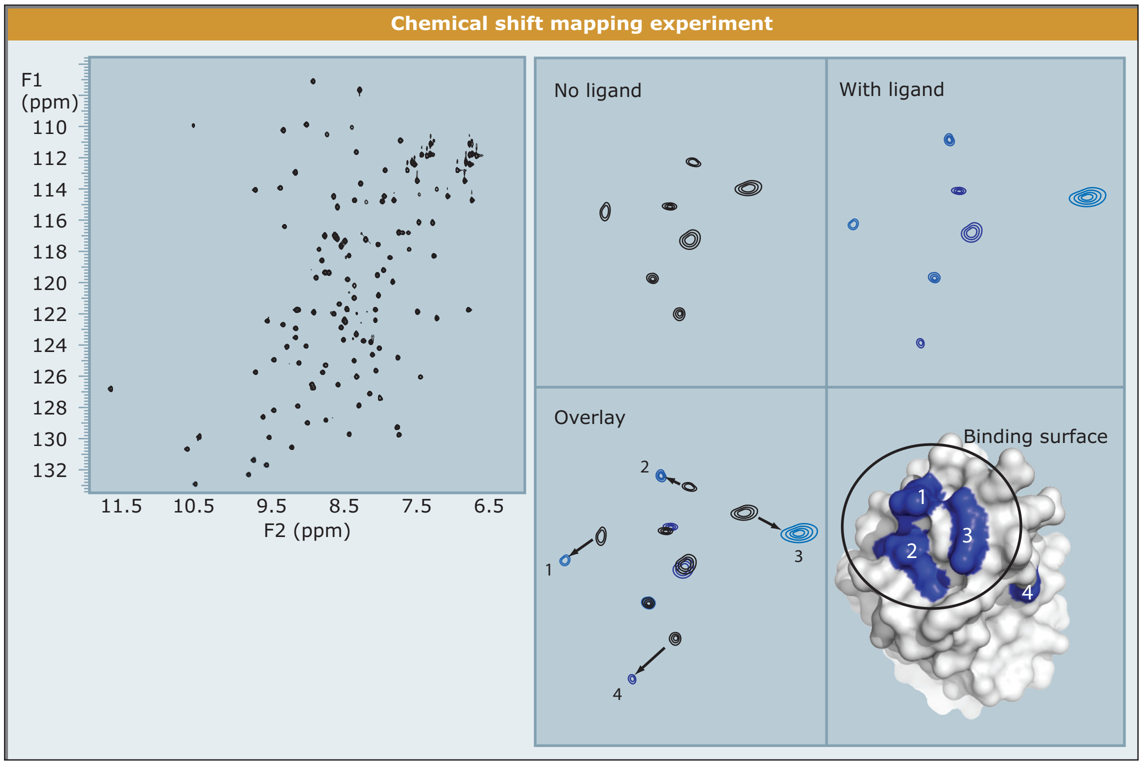

FIGURE 1.13 Chemical shift mapping experiment. The HSQC spectrum of the unliganded protein is shown on the right. The changes in position and intensity of assigned chemical shifts that occur as a ligand is added can be 'mapped' onto the molecular surface to reveal the potential binding site.

1.4 Electron microscopy of biomolecules and their complexes

Key concepts

- Cryo-electron microscopy is capable of imaging macromolecular complexes that may be too large or flexible for X-ray diffraction or NMR approaches.
- In favorable cases, EM methods can produce structural information at or near atomic resolution.

Electron micrographs are a familiar sight in most biology textbooks, and electron microscopy has been used for over 50 years to image biological samples at a resolution that far exceeds anything possible with light microscopy. Beyond its obvious use in imaging cell ultrastructure, EM has come to the fore as an increasingly powerful method for investigating macromolecular structure. For a long time, EM studies of proteins and complexes were limited to resolutions of the order of 20 to 30 Å, but more recent developments in hardware and in experimental approach are providing information at much higher resolution, in some special circumstances approaching that of X-ray crystal structures.

Remarkably, modern **transmission EMs** are still built following the original concepts laid out in the first electron microscope designed in the 1930s by Ernst Ruska. The electron source is generally a tungsten or lanthanum hexaboride filament from which electrons 'boil' off at very high temperatures in a process called thermionic emission. Higher-resolution studies employ a field-emission gun (FEG) that provides a bright beam of approximately coherent (i.e., parallel and in-phase) electrons that have a very narrow distribution of energies. The high energies, and therefore short wavelengths, of these electron beams explain why EM is capable of much higher resolution than optical microscopy.

For imaging, EMs employ a system of high-field magnets that act as lenses to focus the electrons, in the same way that glass lenses are used in light microscopy. The obvious advantage here is that unlike in the X-ray diffraction experiment, the phases of the electron waves are maintained throughout the process, allowing direct imaging of the sample on a suitable detector, most commonly photographic film or charge-coupled devices (CCDs). Lastly, beam characteristics such as size and coherence are controlled by a series of apertures situated above and below the sample stage. Electrons are scattered by air, and therefore the entire electron path, including the sample stage, is maintained at a high vacuum. This necessitates fixation of biological samples in order to maintain, as closely as possible, their native structure.

In general, protein structure analysis by EM is carried out with samples prepared in one of two ways. Negative staining involves the deposition of heavy metal salts, usually uranyl acetate or phosphotungstic acid, on and around the molecules of interest spread onto a carbon support. These heavy elements interact with the electron beam very strongly, and as a result this method provides extremely high-contrast images. In addition, the presence of the metal largely protects the protein complex from the damaging effects of electron bombardment and allows high electron doses to be employed. The sample is essentially coated in heavy salts, though, and because of this details of internal structure are lost (FIGURE 1.14). This leaves an image of the outline of the biomolecule and limits the attainable resolution to around 10 to 30 Å. In addition, the effects of heavy metal binding and dehydration may result in some distortion of the native structure.

Clearly, it is desirable to maintain the sample in as close to physiological conditions as possible, and to be able to investigate its overall architecture, both external and internal. This has been made possible by the development of cryo-EM methods. Here, the sample molecules are spread, under largely native solution conditions, onto a carbon grid and flash frozen by being plunged into a reservoir of liquid ethane. The rapidity of the freezing process prevents formation of ice and instead results in vitrification, where the water molecules adopt an amorphous or 'glasslike' noncrystalline state (Figure 1.14). In order to maintain this frozen state, the sample stage in a cryo-EM is maintained at low temperatures with liquid nitrogen (~80 K) or helium (~5 K). This limits the damaging effects of the electron beam, but the effects are not eliminated and cryo-EM studies must be carried out using electron doses that are much lower than are possible in negative stain experiments. This limitation is exacerbated by the fact that images of unstained samples have a much lower contrast due to the poor interaction of electrons in the carbon, nitrogen, and oxygen from which organic molecules are made. The resulting low signal–noise ratio complicates interpretation and makes assignment of orientation difficult, placing a lower limit on the molecular weight of around 200 to 300 kDa in the single-particle approaches described below.

The highest-resolution structures determined thus far have been derived from two-dimensional arrays (or crystals) of identically oriented protein complexes. This, then, allows both electron diffraction and imaging experiments to be carried out. Here, the advantage over X-ray diffraction is that the phases of the measured amplitudes can still be directly determined and the image reconstructed by Fourier transform. **Electron crystallography** is not, however, without disadvantages and problems of crystal quality, sample preparation, and other technical issues make electron crystallography a rather challenging endeavor. Most notably, a single diffraction image collected at a single orientation of the two-dimensional crystal will only allow a two-dimensional image, or projection, of the sample to be generated. In order to extend the information to three dimensions, a number of images must be collected where the sample is tilted with respect to the electron source. In practice, physical limitations prevent tilt angles of more than about 60°, which results in a cone of missing data. Inevitably, the resulting three-dimensional structure will be less well defined (i.e., at a lower resolution)

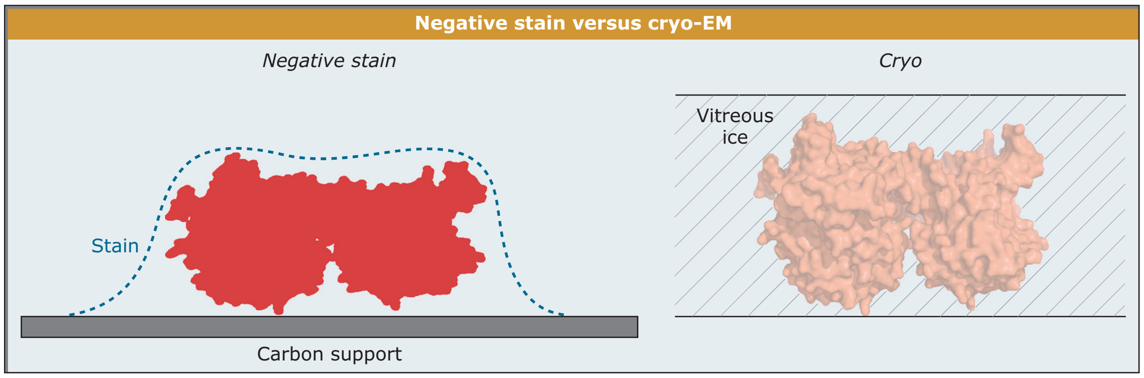

FIGURE 1.14 Negative stain versus cryo-EM.

along the direction of the electron beam than in the plane of the sample stage. In spite of the experimental difficulties, this method has been used with considerable success to obtain the structures of several membrane proteins, most notably bacteriorhodopsin and other helical assemblies at effective resolutions of 3 to 4 Å (FIGURE 1.15).

It has been suggested that the preparation of highly ordered two-dimensional protein crystals of high-molecular-weight protein complexes presents as many technical challenges as growing three-dimensional crystals for X-ray analysis. The requirement for ordered arrays, however, can be circumvented by using single-particle methods that were originally developed for negative stain imaging, but which are now being applied to great effect in cryo-EM. In this approach, individual proteins are scattered onto a carbon surface, where they sit in a range of orientations. Thus different 'faces' of the molecule complex are imaged as a series of two-dimensional projections, and from knowledge of the relative orientations of each of these a three-dimensional view of the molecule can be built (FIGURE 1.16). Although conceptually simple, in practice this method is technically challenging and somewhat laborious. In addition to and, in part as a result of, the low contrast, deriving the relative orientation of each particle is demanding. The simplest assumption is that all molecules have identical structure and are related by a set of simple rotations. In practice, this may not be the case, and conformational heterogeneity arising from intrinsic disorder/flexibility or from the existence of different liganded states of molecules within the sample may be

Bacteriorhodopsin structure

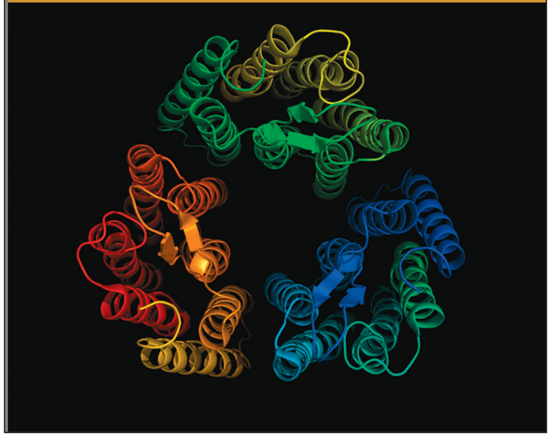

FIGURE 1.15 Bacteriorhodopsin structure determined by electron crystallography. Image generated from Protein Data Bank file IFBB.

difficult to account for, and the problem becomes increasingly acute as higher resolutions are sought. Also, a complete high-resolution image requires a complete sampling of possible orientations, a situation that is rarely achieved because any asymmetric objects randomly scattered onto a surface will inevitably settle most often in the most stable orientations. This problem can be partially circumvented by use of 'holey' carbon grids, which have pores that contain particles in suspension such that they can adopt essentially random orientations (Figure 1.16). Nonetheless, the technical difficulties mean that the most complete and highest-resolution cryo-EM reconstructions may require analysis of tens or hundreds of thousands of individual particles. An example of a field of single particles in a cryo-EM image is shown in FIGURE 1.17.

Probably the most common use of single-particle cryo-EM is in the investigation of the structures of large, multiprotein complexes that may be difficult or impossible to crystallize for X-ray crystallographic analysis, or that are too large for NMR studies. In many cases, individual proteins within such complexes may be amenable to X-ray or NMR methods, and high-resolution structures may be available. In such cases, it may be possible to orient or 'dock' the structure of the isolated subunit into the lower-resolution EM envelope, potentially providing valuable information about the structural and functional roles of individual components in the context of the biologically relevant complex. Although these docking procedures may be carried out manually, a variety of computational fitting procedures have now been developed to guide and accelerate the process. This combined approach has had a number of notable successes, none more impressive than the reconstruction of an atomic model for the colivirus T4 by Rossmann and coworkers.

Finally, a recently developed and exciting use of cryo-EM called cryo-electron tomography is beginning to reveal the structures and distribution of large protein complexes in cells. In this method, individual cells are frozen on a support, as for single-particle studies, and a set of EM snapshots is then collected as the frozen specimen is rotated by small angular increments. Each image therefore represents a two-dimensional projection of the sample 'viewed' from different directions, and from this series the original image can be reconstructed in three dimensions by back-projection. The basic principle is shown schematically in FIGURE 1.18 for an isolated pro-

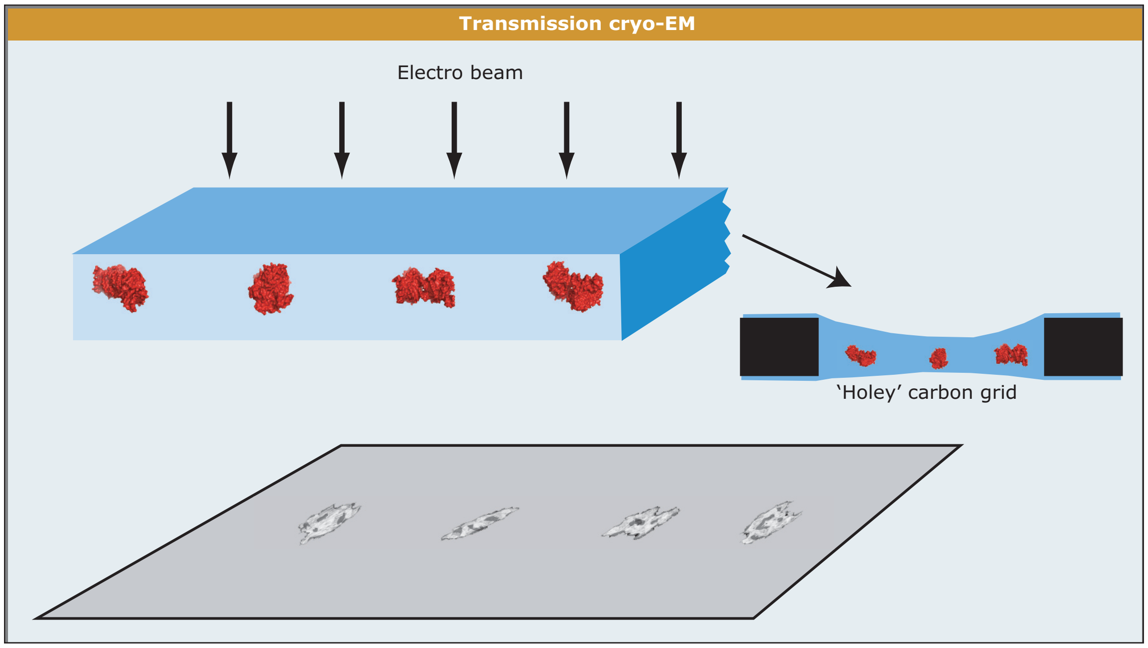

FIGURE 1.16 Transmission cryo-EM.

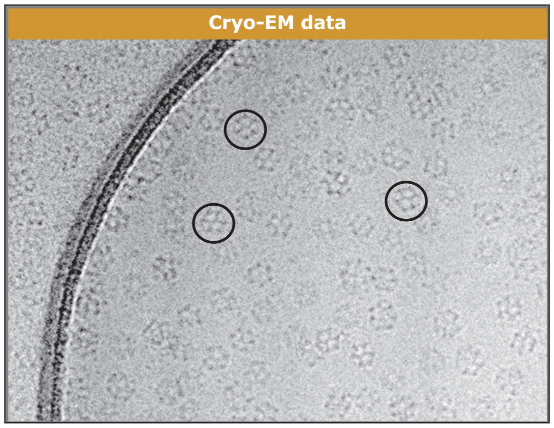

FIGURE 1.17 Cryo-EM data. Three similarly oriented particles are highlighted. Courtesy of Dr. Peter Rosenthal.

tein complex as a simple example. Clearly, the amount of information potentially contained in the reconstructed image is truly immense, representing the three-dimensional arrangement of every molecule in the cell! Extracting information about specific complexes, however, represents a formidable challenge. In theory, specific labeling of particular complexes would be helpful, but is extremely difficult to achieve for complexes within a cell. An alternative approach is to use structural 'templates' derived from complexes of known structure to search the cytoplasmic milieu. This method has already been used to locate large symmetric complexes such as the 26S proteasome in cryo-preserved cells.

1.5 Protein structure representations—A primer

Key concepts

- Proteins are three-dimensional objects, and even a relatively small example of ~10 kDa molecular weight will contain upward of 1000 atoms. This causes considerable difficulties in presenting structural data in a clear and understandable way.
- Displaying each atom and chemical bond certainly conveys the degree of complexity in proteins, but little other useful information is discernable. For this reason, a number of different schematic representations have been devised in order to illustrate different features of protein structure.

Arguably, the most obvious way in which to represent a molecular structure is to draw each atom along with the chemical bonds between atoms. This kind of representation is known as a 'ball-and-stick' type, a name that admirably describes it. Here, each atom is shown as a sphere, generally of arbitrary radius, and each bond as a cylinder (or sometimes a cone), as shown in FIGURE 1.19. These kinds of representation are most used to convey stereochemical details of, for example, active-site regions of enzymes for which details of relative atomic positions are of most interest. A related method is the 'space-filling' or CPK (for Corey, Pauling, and Koltun) representation (FIGURE 1.20). In this case, atoms are shown as larger spheres scaled according to their atomic radii. For large mol-

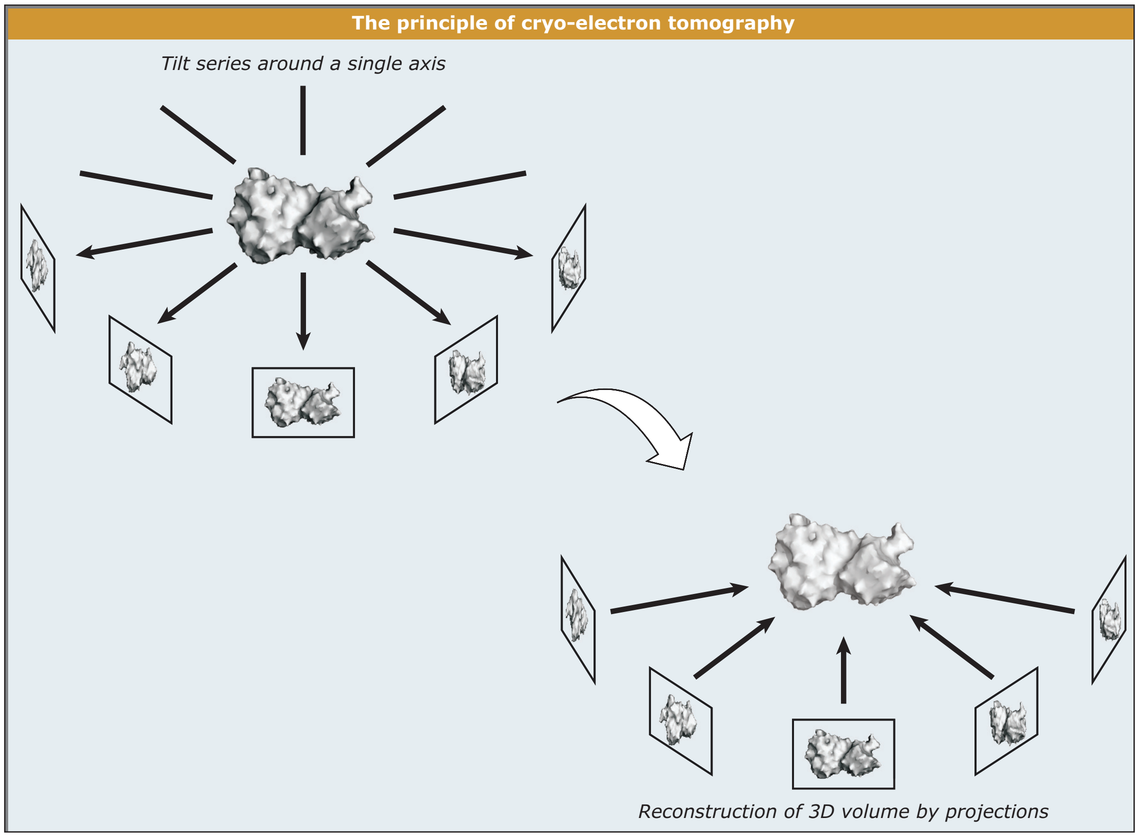

FIGURE 1.18 The principle of cryo-electron tomography.

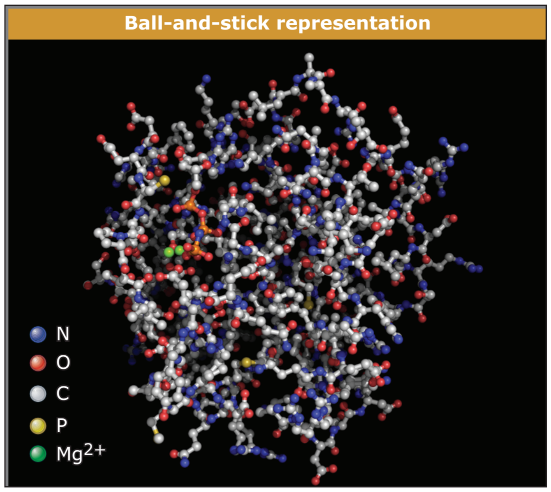

FIGURE 1.19 Ball-and-stick representation with coloring according to atom type.

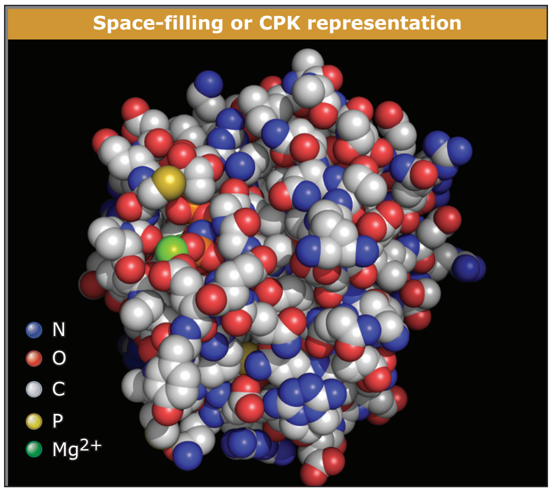

FIGURE 1.20 Space-filling or CPK representation.

ecules with thousands of atoms and bonds, the ball-and-stick representation contains far too much information and the overall architecture of a protein is largely obscured. Similarly, the space-filling representation suffers from the inevitable property that atoms 'inside' the protein are not visible at all, and only surface atoms are discernable.

By far the most popular and effective means of conveying the overall structure of a protein is the 'ribbons' representation (FIGURE 1.21), in which β-strands are shown as arrows to indicate directionality (N-terminal to C-terminal) and α-helices are represented as ribbonlike coils or as tubes. As will be seen later in this chapter, the overall path in three dimensions of a protein chain is quite accurately described by linking the Cα atoms of each consecutive amino acid—

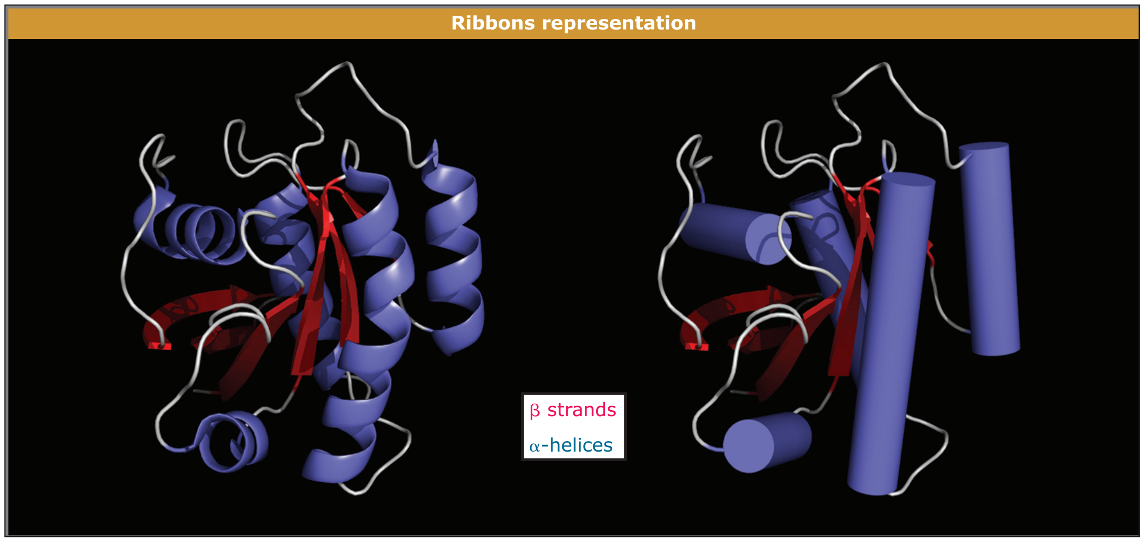

FIGURE 1.21 Ribbons representation with β strands shown as arrows (running from N-terminus to C-terminus) and α-helices shown as spirals (left) or simple tubes (right).

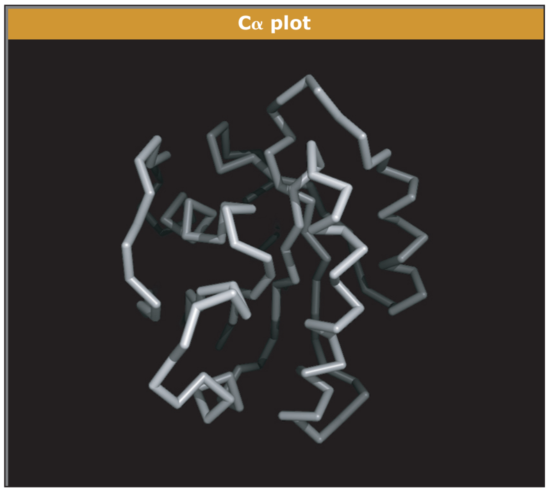

FIGURE 1.22 Cα-plot. For simplicity, it may be convenient to show a structure as a Cα plot where each α-carbon (one per residue) is joined to the next.

thereby creating a Cα plot (FIGURE 1.22). This gives a much less cluttered view of the shape and architecture, but to the inexperienced eye it contains limited information about secondary structural content. In the 'ribbons' diagram, each secondary structural element is distinguished by a different shape or motif. Thus β strands are shown as arrows and α-helices as tubes or helical ribbons. Segments of random-coil structures that link strands and helices together are represented as a 'worm' that may either rigorously follow the positions of the Cα atoms or, more commonly, trace an approximate and much smoother path determined by mathematical interpolation procedures.

Armed with an overall picture of the structure of a protein, together with the possibility of being able to illustrate more detailed aspects at the level of individual atomic arrangements, we may still wish to examine topographical and physico-chemical properties of the protein surface, i.e., the regions of the molecule that are directly in contact with bulk solvent (water, ions, and so forth), small-molecule ligands (cofactors, substrates, and so forth), and other proteins with which biologically interesting interactions are made. Surfaces are generated by computationally 'rolling' a sphere or probe with a given radius (usually 1.4 Å, which approximates that of a water molecule) over a hard-sphere model of the entire protein structure. If the path is taken as that defined by the center of the probe, the result is a 'solvent accessible' surface. A commonly used variation of this method generates surface points from the volume boundary of the probe, which tends to smooth out sharp grooves and crevices. This kind of representation is often called a 'molecular' or 'Connolly' surface (FIGURE 1.23).

Having generated the surface diagram, it is possible to 'map' various features of the molecule onto it. One of the most popular such uses is to display the electrostatic properties of the protein surface—that is, the regions of positive and negative charge that are most often associated with clusters of basic (Arg, Lys) or acidic (Asp, Glu) amino acids. This information can reveal likely binding sites for ligands, such as DNA (FIGURE 1.24). Similarly, algorithms have been developed to calculate the relative hydrophobicity of different parts of the molecular surface. Nonpolar interactions generally contribute greatly to the formation of stable protein–

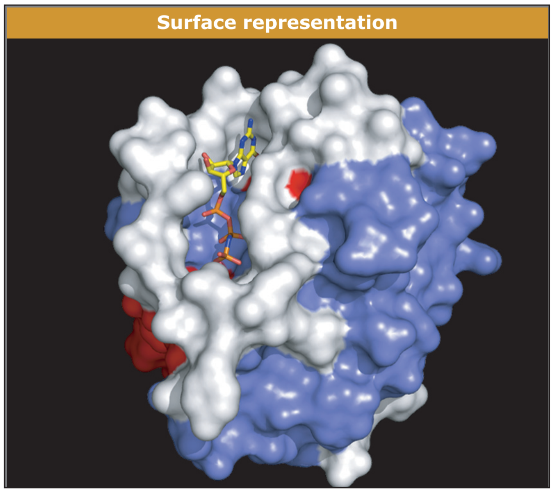

FIGURE 1.23 Surface representation. External, solvent-accessible features such as ligand-binding grooves and clefts may be revealed, and various properties (secondary structure in this case) of the underlying atoms can be mapped onto the surface.

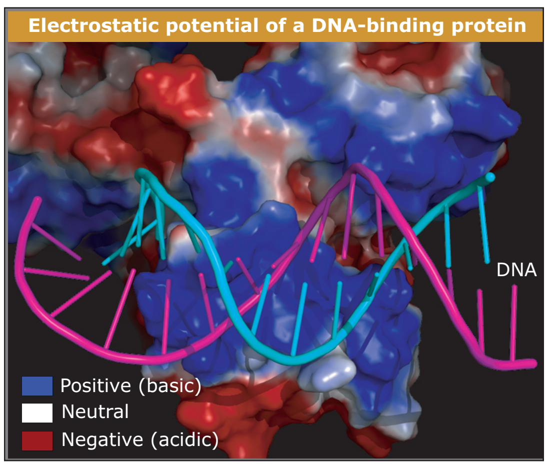

FIGURE 1.24 Electrostatic potential of a DNA-binding protein mapped onto the molecular surface. The highly negatively charged DNA molecule interacts predominantly with regions of positive charge (blue) on the protein.

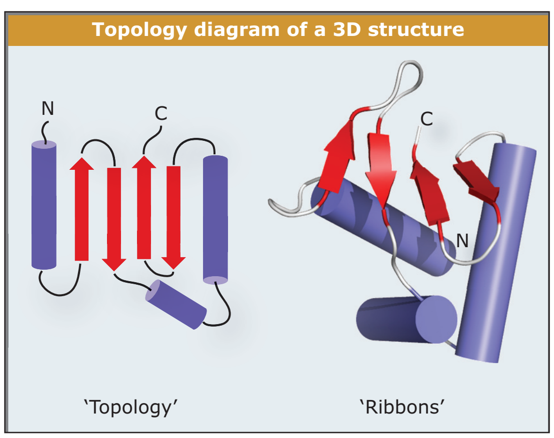

FIGURE 1.25 A topology diagram (left) can be used to simplify a three-dimensional structure (right) for comparison or other purposes.

protein interactions; as a result, this approach can also reveal potential binding sites and may even be used to estimate binding affinities. Additional indicators of functionally important regions can often be derived from mapping primary sequence homology within a family of related proteins onto the molecular surface of one of its members, given that the most highly conserved amino acids that are accessible to solvent are likely to be involved in an evolutionarily conserved function.

Finally, it is often convenient to be able to schematically represent the secondary structural elements of proteins in terms of their relative position and the order in which they occur in the polypeptide chain. This is often called a topology diagram, and such diagrams are used extensively in the classification and comparison of different families of protein folds (*Section 1.8, Tertiary structure and the universe of protein folds*). Unfortunately, there are almost as many variations on this theme as protein structures themselves; an example of one that will be used in this chapter is shown in **FIGURE 1.25**.

1.6 Proteins are linear chains of amino acids—primary structure

Key concepts

- Proteins are composed of linear chains formed by condensation reactions between amino and carboxyl groups of amino acids
- Only 20 amino acids are commonly found in proteins; all are L-enantiomers with a configuration.
- Differences in physico-chemical properties of amino acids are fundamental to the diversity that we observe in protein structure and function.

Proteins are linear and unbranched polymers of amino acid building blocks. (There are a small number of exceptions, such as cyclic peptides, which are formed posttranslationally.) This basic structural property is a consequence of the fact that the sequence of amino acids in proteins is encoded by triplets of bases in DNA—itself a linear, unbranched polymer chain of nucleotides. Unlike in common organic polymers such as polythene (polyethylene), the basic monomer units of proteins comprise not a single species, but 20 chemically distinct amino acids.

Remarkably, the same set of 20 amino acids is found in proteins from all living organisms. Nineteen share a basic structure, NH_2-(CH-R)-

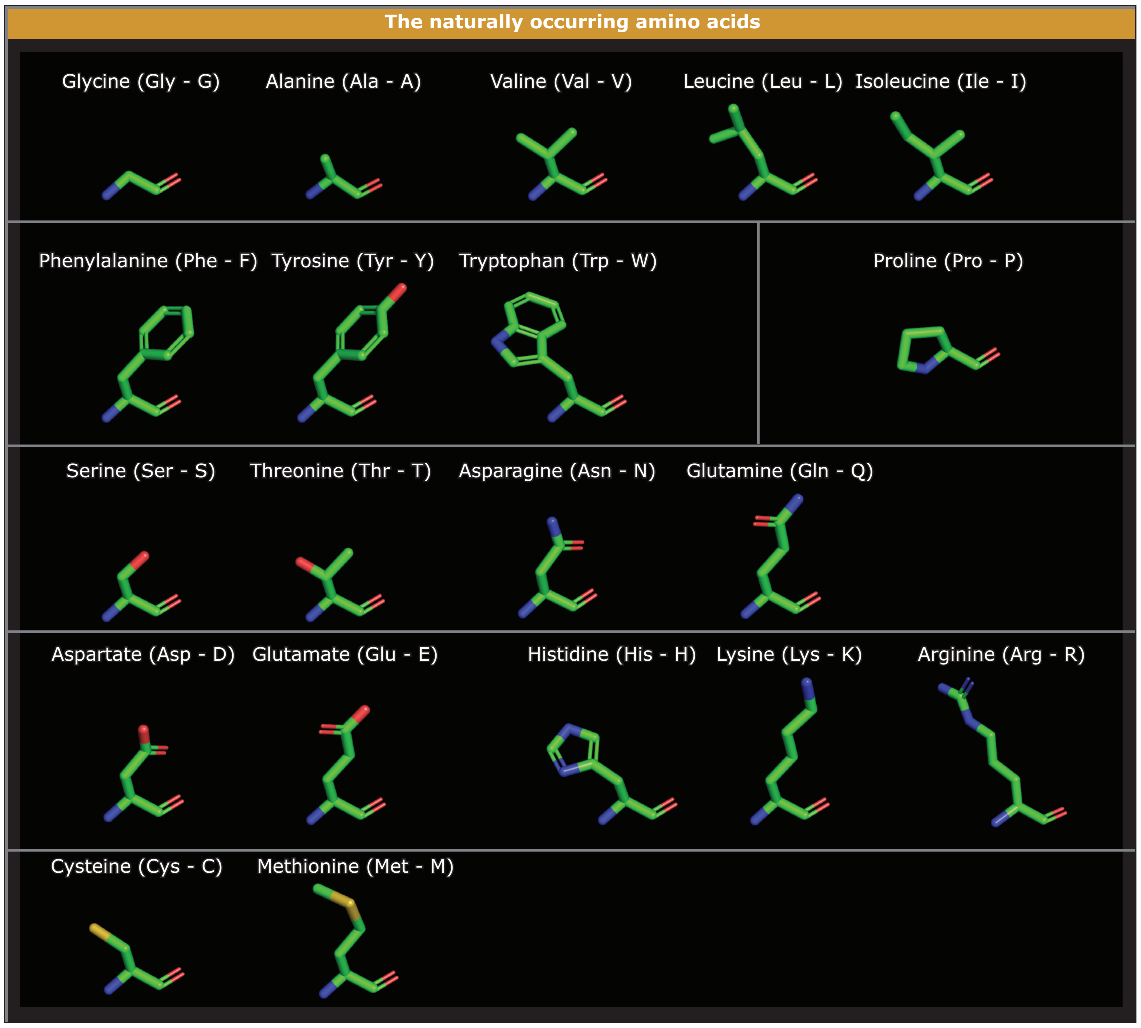

FIGURE 1.26 The 20 naturally occurring amino acids grouped roughly according to chemical properties, with three-letter and single-letter codes indicated. Carbon = green, oxygen = red, nitrogen = blue, and sulfur = yellow.

COOH, and each is distinguished by the composition of the R-group (**FIGURE 1.26**). The exception is the imino-acid proline, in which the nitrogen atom of the backbone is locked into a five-membered pyrrolidine ring. The conformational rigidity imposed by this arrangement plays a number of important roles in protein folding. Amino acids are generally referred to by their three-letter or single-letter abbreviation (Figure 1.26). For the most part, the three-letter form will be used here.

A wide variety of 'nonstandard' amino acids, such as selenocysteine, hydroxyproline, hydroxylysine, ornithine, and γ-carboxy-glu, occur as by-products of metabolic reactions, through post-translational modification or by the activity of specific biosynthetic pathways. These amino acids play highly specialized roles, some of which will be mentioned in forthcoming sections.

Amino acids generally found in proteins all have the α configuration (**FIGURE 1.27**), meaning that the amino group is attached to the α carbon. Thus α amino acids contain only a single carbon atom (excluding the carbonyl carbon) in the backbone. Alternative forms, such as β amino acids that have two backbone carbon atoms, are not found in proteins, but β alanine is found in some naturally occurring peptides such as carnosine. Presumably, β amino acids were selected against early in evolution due to the increased degree of rotational freedom around the C-C bond that would prevent higher-order folding, although some oligopeptides made from β amino acids are known to adopt novel 'secondary structures' in solution.

With the exception of glycine, which has a single hydrogen atom as its 'side-chain' and is thus symmetric, amino acids all possess a chiral center at the Cα atom; that is, they can adopt a right- (D) or left-handed (L) form, so named

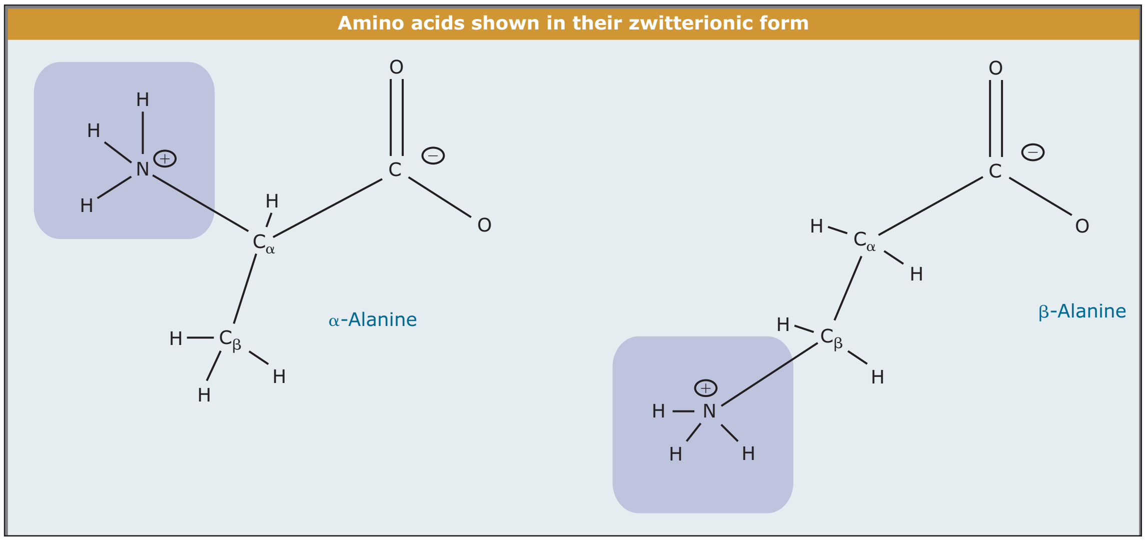

FIGURE 1.27 α (left) and β (right) amino acids shown in their zwitterionic form.

because of the way in which these forms rotate polarized light (FIGURE 1.28). While D-amino acids do occur in some specialized circumstances, all proteins are built from L-amino acids, although there appears to be no fundamental reason why evolution could not have selected the D-form.

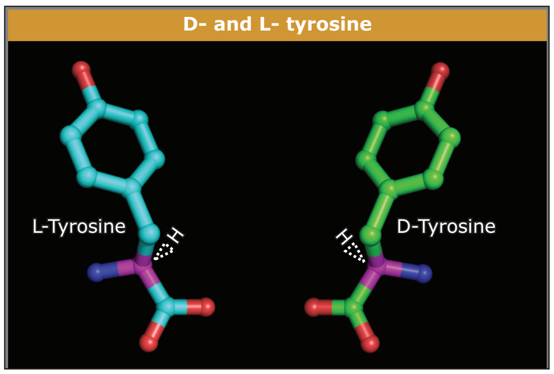

FIGURE 1.28 D- and L-tyrosine.

In general, amino acid residues in proteins adopt the *trans* configuration (FIGURE 1.29). In a *cis* configuration, the Cα atoms (and thus the side-chains) of adjacent residues are brought into close proximity; this configuration is, therefore, sterically disfavored. The major exception is the amino acid proaline, for which the energy barrier between the *cis* and *trans* forms is much lower, and which has been observed in the *cis* configuration more often than any other amino acid. Indeed, the utility of *cis-trans* proline isomerization is exploited in a number of signaling systems, and can be 'catalyzed' by a family of proteins called *cis-trans* prolyl isomerases that have evolved specifically for this purpose.

An extremely important chemical feature of amino acids is that they are amphoteric. At physiological pH the α amino and α carboxylic acid groups are essentially completely ionized, and amino acids (at least those with nonioniz-

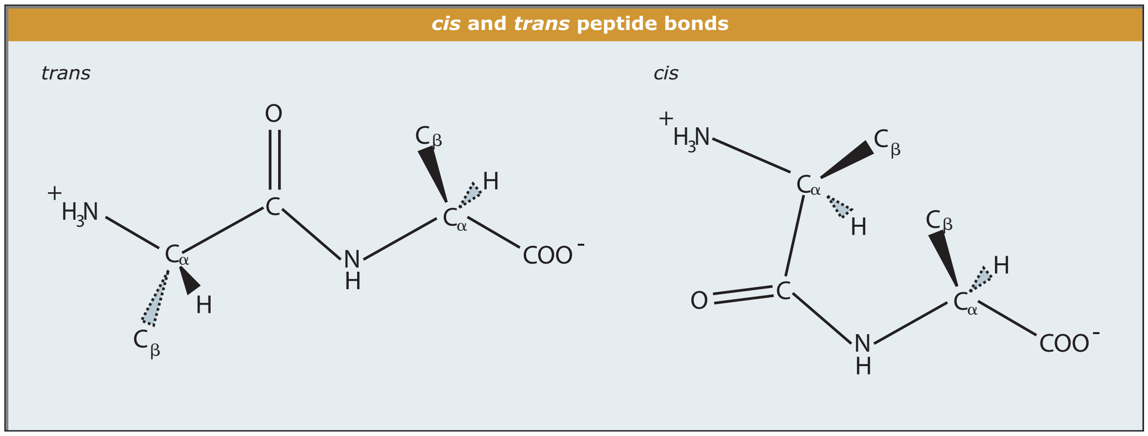

FIGURE 1.29 *cis* and *trans* peptide bonds. Most amino acids adopt the *trans* configuration.

able R-groups) are therefore also zwitterions with no net charge. As we will see, some amino acids have side-chains that contain additional ionizable groups. Together, these properties are important for the solubility of amino acids, the ability to form polypeptide chains, and the overall charge characteristics of the folded protein molecule that, as we will see, may be important for biological function in many contexts.

The amino acid sequence encoded in a DNA sequence is translated into a protein polymer by the 'decoding' of the mRNA template nucleotide sequence by the ribosome. These large, protein–RNA complexes catalyze the formation of 'peptide' or amide bonds between the amino (NH^2) and carboxyl (COOH) terminal groups in a **condensation** reaction, proceeding in the N- to C-terminal direction. The chemist Linus Pauling used available crystal structures of small molecules to show that the peptide bond is, in fact, a resonance hybrid of two forms. This results in a partially double-bonded character and confers rigidity (**FIGURE 1.30**). If this were not the case, the additional degree of rotational freedom would prevent amino acid chains from folding into the stable protein structures we see in biological systems.

The atoms of the peptide bond are essentially planar, and only a small degree of rotation about the ω dihedral angle is possible (Figure 1.30). Thus the backbone of a polypeptide chain has two degrees of rotational freedom: one around the N-Cα bond (ϕ), and the other around the Cα-C=O bond (ψ). Steric effects result in a rather limited range of ϕ-ψ dihedral angles that are energetically favorable in a polypeptide, and this accessible conformational space is classically represented in graphical form as the familiar ϕ-ψ or Ramachandran plot (**FIGURE 1.31**). Glycine is an exception, however, and the absence of a side-chain results in a much more extensive range of accessible ϕ-ψ combinations. The Ramachandran plot was originally derived from empirical considerations based on van der Waals contact distances (see below), observed in the limited database of small-molecule structures available at the time. Nevertheless, it has been largely confirmed by the protein structures determined since its introduction, and is one of the commonly used means of assessing the reliability of newly determined structures.

Based on the type of R-group or side-chain,

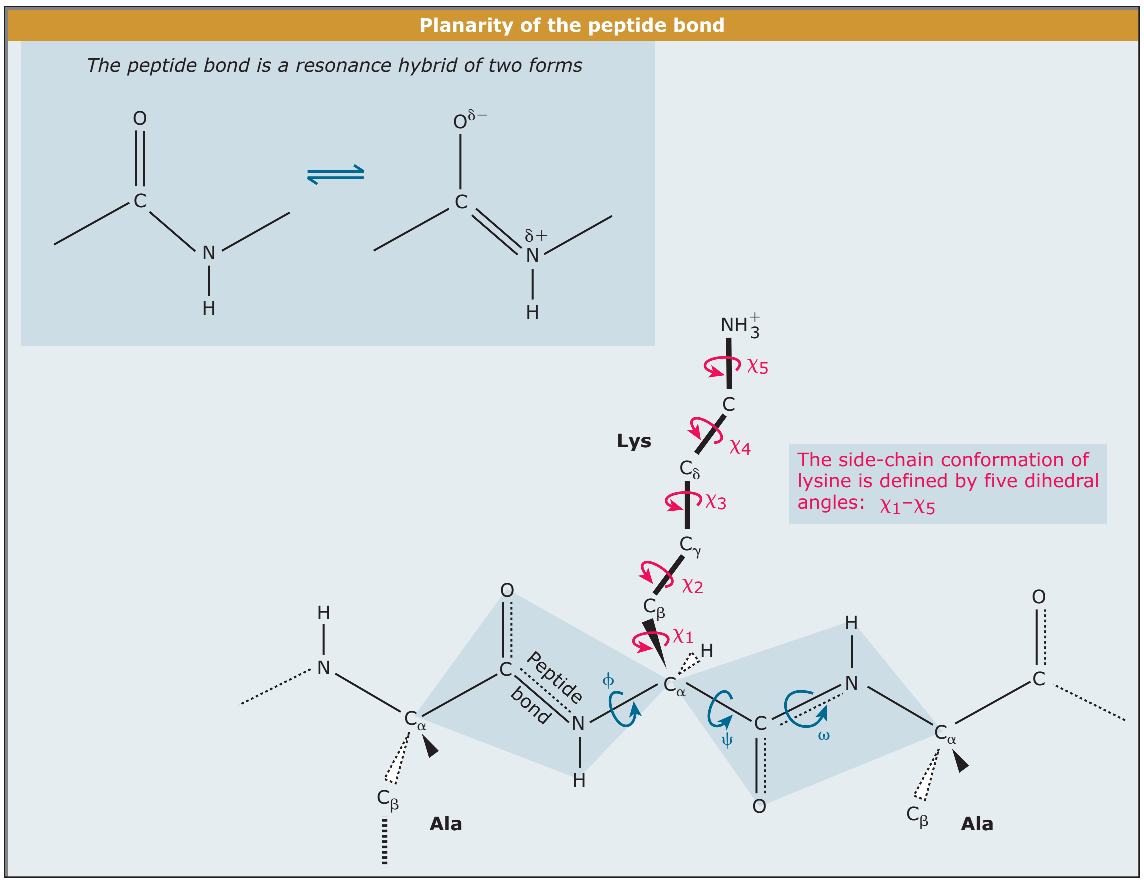

FIGURE 1.30 Planarity of the peptide bond. In addition, the five dihedral angles for lysine about which essentially free rotation can occur are shown.

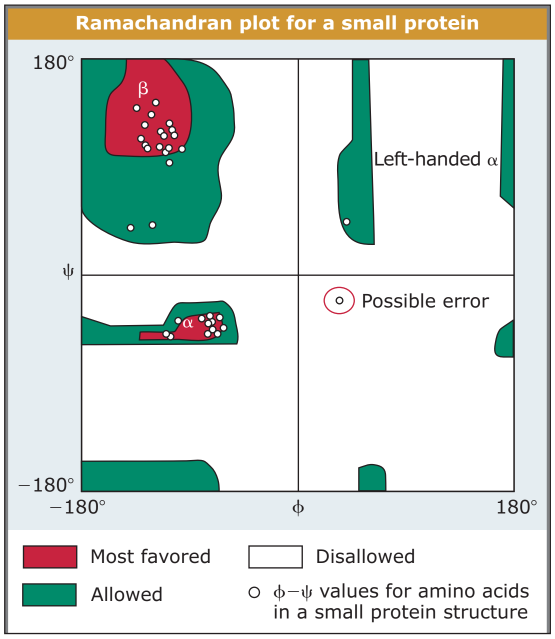

FIGURE 1.31 The Ramachandran plot for a small protein structure.

and with a few exceptions, the amino acids can be broadly classified based on their physico-chemical properties: nonpolar, positively charged-polar, negatively charged-polar, and neutral-polar, although many other classifications are possible (Figure 1.26).

The nonpolar amino acids can be subdivided into those with aliphatic side-chains (Ala, Val, Leu, and Ile) and those with aromatic side-chains (Phe, Tyr, and Trp). It is the presence of the aromatic residues that confers the characteristic absorption spectra of polypeptides in the UV-visible range of wavelengths. In general, they are rather **hydrophobic**, and do not favor interactions with polar solvents such as water. Thus they are most often, but not exclusively, located in the core of globular protein molecules. The hydroxyl group of Tyr, however, and the pyrrole nitrogen atom within the large indole side-chain of Trp have significant polar character. In addition, the pyrrolidone side-chain of proline contains three carbon atoms in a ring, and therefore has a significant hydrophobic nature.

Of the more **hydrophilic** amino acids, the four neutral-polar residues are distinguishable by the presence of either hydoxyl (Ser, Thr) or amide (Asn, Gln) groups within their side-chains, and these are uncharged at physiological pH. Carboxylic acid groups distinguish the two negatively charged (acidic)-polar amino acids, aspartate and glutamate. Lys, Arg, and His are the basic or positively charged amino acids, and are characterized by primary amine, guanidine, and imidazole groups, respectively. The imidazole group of His is highly versatile and is often employed in enzyme active sites due to its basicity at physiological pH and its nucleophilicity. These important properties will be revisited later.

Two amino acids contain sulfur in their side-chains: The sulfydryl of Cys is extremely reactive and is able to form disulfide bridges within or, less often, between polypeptide chains. This property is important for the stability and folding of some proteins, particularly secreted proteins or those that are otherwise exposed to the harsh conditions of the extracellular milieu. Methionine is rather nonpolar, and also contains a sulfur atom in its side-chain. As we have seen from the use of its selenium-substituted cousin in modern methods of crystallographic phase determination, it occupies a rather special place in the hearts of X-ray crystallographers!

As should be clear from the structures of the amino acids, many of their side-chains have rotational freedom around single bonds (Figure 1.30). These dihedral angles are referred to by the Greek letter χ such that χ_1 describes rotation around the Cα-Cβ bond, χ_2 around Cβ-Cγ, and so forth. Clearly, not all χ rotations are possible. For example, the χ rotations of Pro are restricted to very small angular increments associated with different puckers of the pyrrolidine ring. In addition, χ_5 rotations around the Arg Nϵ-Cζ bond are highly restricted at physiological pH because, when protonated, the N-C bond of the guanidinium has significant double-bonded character due to resonance. Examination of the structural database has revealed that additional restrictions imposed by steric and other effects result in favored conformations or 'rotamers' for many side-chains. This information has been usefully incorporated into a number of commonly used computer graphics programs to aid in the interpretation of electron density maps during the structure determination process.

The characteristics of the amino acids and the aqueous environment in which proteins exist determine the type of interactions that are observed within and between proteins and their ligands. Most of these are relatively weak and, with the exception of the disulfide linkage, do not involve the formation of chemical bonds.

By far the strongest and the major driving force in protein folding and the interactions between proteins is the hydrophobic effect. As mentioned earlier, 'hydrophobic' literally means 'water hating' and describes the property of certain atoms, such as carbon, that prevents them from interacting with water in a thermodynamically favorable way. Protein folding itself can be thought of as arising from the requirement that nonpolar atoms are buried in the interior or 'hydrophobic core' or a protein in its folded state. In a similar way, the hydrophobic effect contributes greatly to the inter molecular interactions between proteins and other proteins or ligands (*see Section 1.13, Protein–protein and protein–nucleic acid interactions*).

The hydrogen bond occurs most commonly between a hydrogen atom covalently bonded to an electronegative atom and possessing a partial positive charge, and a second electronegative acceptor atom. In proteins, the most common hydrogen bond (or H-bond) donors are NH groups (main-chain peptide NH), NH_2 groups (amino termini; asparagine/glutamine side-chains), and OH groups (serine and threonine side-chains). Many potential H-bond acceptors exist in proteins, including C=O (main-chain carbonyl oxygen; asparagine/glutamine side-chains), -N= (histidine side-chain), -O- (serine/threonine side-chains), and occasionally, the SH groups (cysteine side-chains). The strength of hydrogen bonds varies greatly depending on the identity of the donor/acceptor groups and geometry, but is generally in the range of 2 to 3 kcal/mol. A number of observations from high-resolution crystal structures also suggest that CH groups can act as H-bond donors, although such interactions are estimated to be much weaker. Some examples of H-bond interactions are shown in **FIGURE 1.32**.

Interactions between positively and negatively charged atoms in proteins most often occur between the side-chains of the basic amino acids lysine and arginine (and in some circumstances histidine) and those of glutamic/aspartic acids. These interactions are often classified as salt bridges and are most usefully thought of as H-bonds that also involve a significant contribution from electrostatic effects. Electrostatic interactions are complex and depend on a variety of factors, such as local **dielectric constant** and the ionization state of the atoms involved. The latter is described by the pK_a value and is dependent on environmental and chemical factors such as the pH, solvent polarity, and local electrostatic effects. Clustering of similarly charged groups on the surface of a protein can create regions or patches of positive or negative electrostatic potential that often define in-

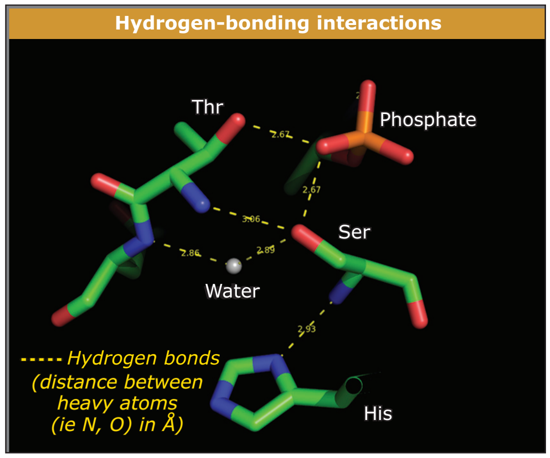

FIGURE 1.32 Hydrogen-bonding interactions.

teracting surfaces for cationic or anionic ligands. This is best exemplified by the interactions between positively charged surfaces on nucleic acid-binding proteins with the highly negatively charged polyanions, RNA and DNA.

Disulfide bonds or linkages are, except for some unusual examples described later (such as GFP), the only covalent interaction that takes place between protein side-chains. They are formed by oxidation of the sulfydryl side-chains of cysteine residues to form a sulfur–sulfur (disufide) bond. Although uncommon in cellular proteins due to the highly reducing conditions of the cytoplasm, they are often found in secreted proteins (such as hormones) or proteins anchored to the plasma membrane but exposed to the extracellular milieu, where the additional structural stabilization provides rigidity and resistance to proteolytic degradation (**FIGURE 1.33**).

1.7 Secondary structure—the fundamental unit of protein architecture

Key concepts

- The amino acid sequence (primary structure) can form essentially two secondary structures: α-helices and β strands.
- Secondary structural elements are often arranged into a few commonly occurring supersecondary structural units.
- Secondary structures are often connected through flexible regions or 'linkers' that may play important functional roles.

Protein structures are not usually composed of extended chains of amino acids, but rather are formed from the association of one or more segments of a rather limited number of regular

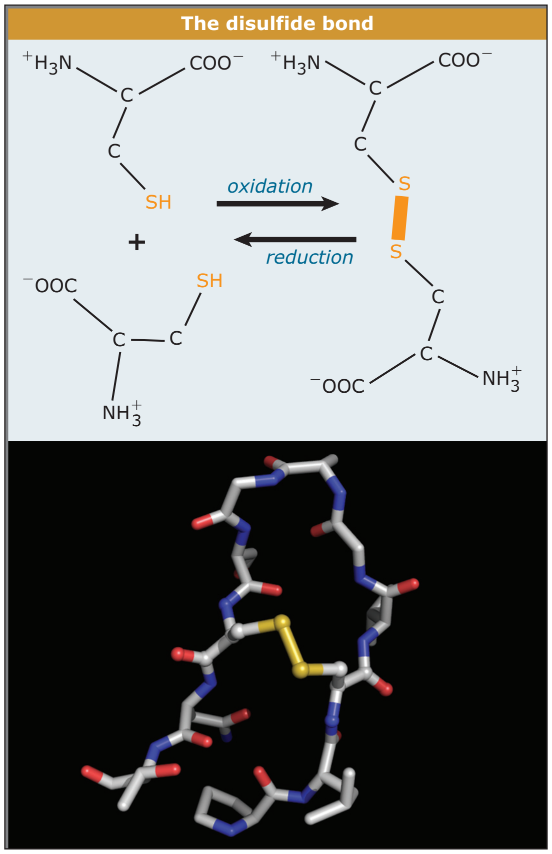

FIGURE 1.33 The disulfide bond.

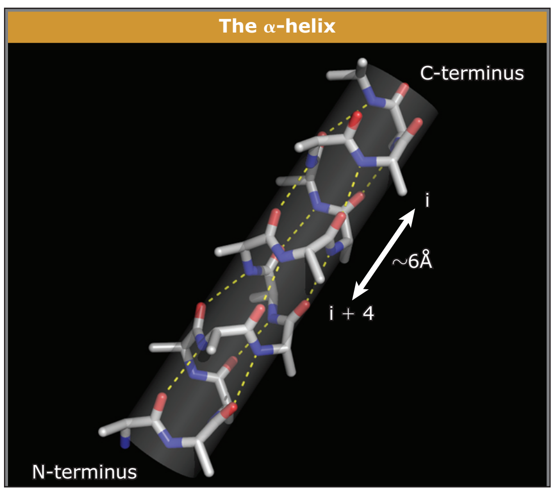

FIGURE 1.34 The α-helix.

structural elements. These elements adopt characteristic backbone ϕ-ψ angles (*Section 1.6, Proteins are linear chains of amino acids—primary structure*), and the majority of proteins utilize two major **secondary structures**, the α-helix and the β strand. This is exemplified by the fact that α and β main-chain configurations are the most highly populated regions of the Ramachandran plot. In the absence of crystal or NMR structures, the secondary structural content of a protein molecule can be estimated using spectroscopic methods such as **circular dichroism**.

The structure of the α-helix was first proposed by Pauling based largely on model building and profound chemical intuition. In fact, α-helices were the first secondary structural element revealed by crystallography. Initial 5 Å electron density maps of sperm whale myoglobin showed eight rodlike structures, and subsequent higher-resolution studies essentially confirmed the most important aspects of the Pauling model. The α-helix is formed and stabilized by a characteristic hydrogen-bonding pattern formed between the main-chain NH group of residue **i** with the main-chain carbonyl oxygen of residue **i+4** (**FIGURE 1.34**). The peptide bonds have a small dipole. As a result, and because they all point in the same direction in an α-helix, the α-helix behaves as a 'macro-dipole' with small positive and negative charges at the N- and C-terminal ends, respectively. This charge characteristic may, in turn, be exploited in protein–protein interactions or ligand binding .

Amino acid **chirality** dictates that the helical twist thus formed is always right-handed, and no examples of left-handed α-helices have been observed in any biological structure yet determined, nor are they likely to be. In a geometrically ideal helix, the direction of the interresidue hydrogen bonds is exactly parallel to the helix axis. This situation is only rarely observed in protein structures, and some nonlinearity of the hydrogen-bond geometry is most common.

A number of variants of the α-helix are, of course, possible. Of these, the so-called 3_{10} helix, in which residue **i** is hydrogen-bonded to **i+3**, is regularly seen to form short turns in an otherwise extended structure. A helix in which residue **i** is hydrogen-bonded to **i+5** is known as the π helix, but is predicted to be rather unstable and, presumably for this reason, does not occur in any known protein structure.

The β strand constitutes the second commonly observed secondary structural element in proteins. β strands were first observed in early crystal structures of lysozyme, and consist of an essentially fully extended polypeptide chain. Alone, the β strand is unstable and is almost always found in combination with others to form β sheet structures.

The conformation of a single strand is such that the peptide NH and C=O groups project in opposite directions, and these directions are reversed in successive residues within the strand. Due to the fundamental structure of α amino

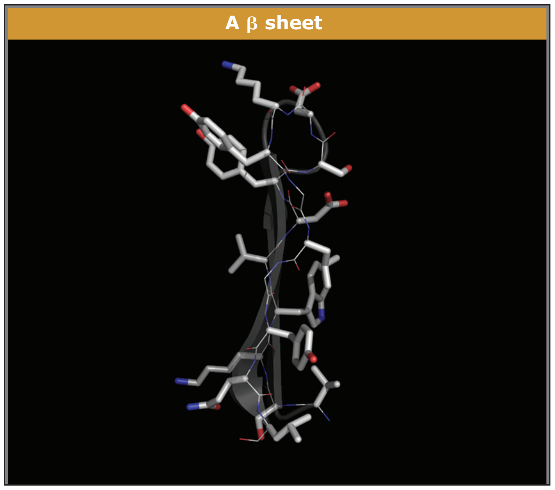

FIGURE 1.35 A β sheet viewed from the side. Amino acid side-chains project away from the plane of the sheet.

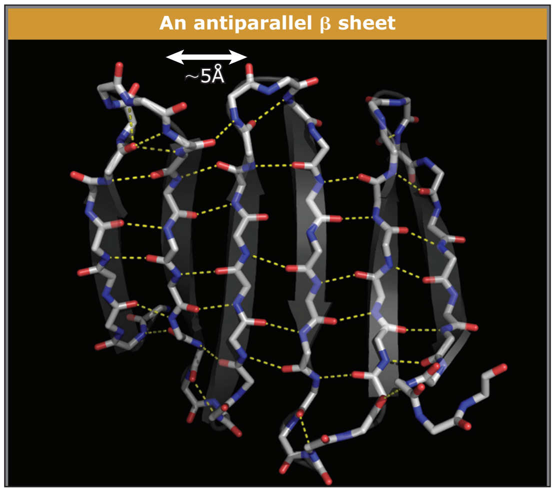

FIGURE 1.36 An antiparallel β sheet showing the characteristic interstrand hydrogen-bonding pattern.

FIGURE 1.37 The β barrel. Image generated from Protein Data Bank file 1BXW.

acids described earlier, the side-chains of each adjacent residue then project in opposite directions, but in a plane that is essentially orthogonal to that defined by the main-chain groups (FIGURE 1.35). Thus, free of potential steric clashes between side-chain atoms, strands can associate through NH—-O=C hydrogen bonds (FIGURE 1.36).

β sheets are not flat, but show a left-handed twist that may be more or less pronounced in different structural contexts (FIGURE 1.37). As may be already clear, association of β strands can occur in either a parallel or an antiparallel orientation, and β sheets are often composed of both (FIGURE 1.38).

In considering the association of individual secondary structural elements to form a folded tertiary structure (*Section 1.8, Tertiary structure and the universe of protein folds*), it is clear that some motifs are highly represented in the structural database. These are often referred to as elements of supersecondary structure, and examples include certain types of β hairpins that connect consecutive β strands in a sheet, certain combination of β strands (e.g., β meander, Greek key motif), and turns that link together α-helices to form the well-known helix-loop-helix and helix-turn-helix motifs (FIGURE 1.39).

Among the most common supersecondary structures is the coiled coil that arises from the interaction of two, three, four, or even more bundles of α-helices with a characteristic pattern of hydrophobic and charged amino acids that repeats every seven residues—the so-called heptad repeat (FIGURE 1.40). Coiled-coil motifs were first predicted by Francis Crick in a theoretical analysis of how α-helices pack together. We now know of many thousands of examples of proteins with a diversity of functions that are observed or predicted to contain coiled-coil regions. The heptad repeat results in a pronounced amphipathic nature where each helix has polar and nonpolar faces. The nonpolar side-chains pack together at the interface, leaving the more hydrophilic side-chains exposed to solvent. Often, basic and acidic residues are found juxtaposed in the coiled coil, where they form salt-bridging interactions and thus provide additional

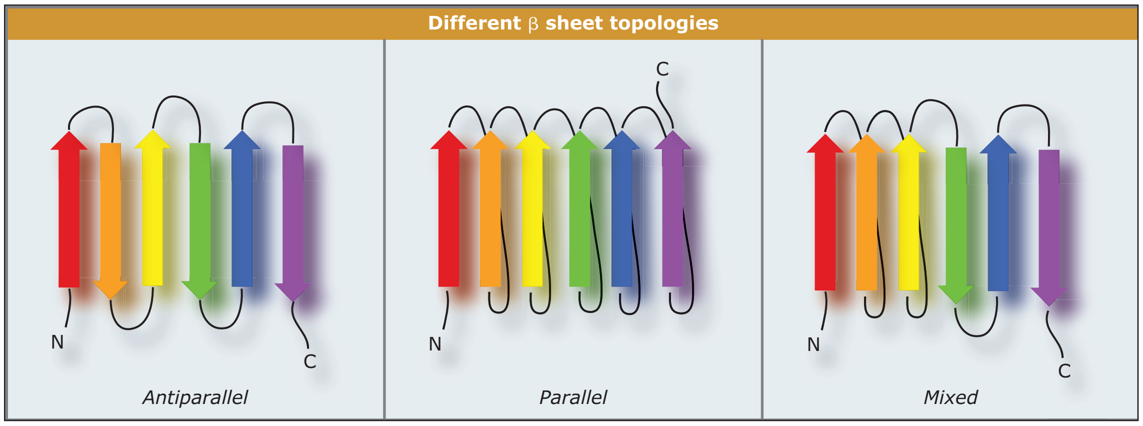

FIGURE 1.38 Different β sheet topologies: parallel, antiparallel, and mixed.

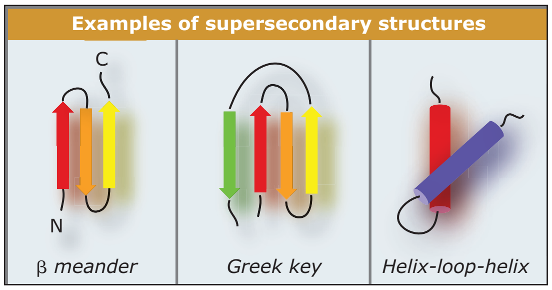

FIGURE 1.39 Examples of supersecondary structures.

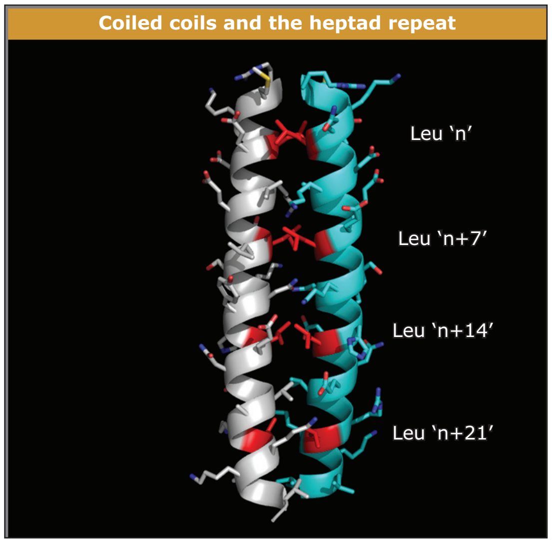

FIGURE 1.40 Coiled coils and the heptad repeat. Image generated from Protein Data Bank file 2ZTA.

structural stabilization. Although parallel coiled coils with a left-handed superhelical twist are most common, antiparallel configurations, and some with a right-handed twist, are also known. Coiled coils in which leucine dominates the 'a' position of the heptad repeat (-abcdefg-abcdefg-etc) are often referred to as 'leucine zippers.'

Although not generally considered to be secondary structures as such, most proteins contain regions of extended conformations that serve to join together the α-helices and β strands. These regions are often referred to as 'linkers' for this reason. In their most extreme form they can be rather long, comprising tens or even hundreds of amino acids. Alternatively, they may consist of only two or three residues forming tight turns that, themselves, have been observed to fall into a number of common supersecondary structural classes described earlier. In many cases, linker regions tend to be poorly defined in X-ray and NMR structures but, as will be seen later, they may become ordered, or even adopt a canonical secondary structure, upon interaction with specific ligands or partner proteins.

1.8 Tertiary structure and the universe of protein folds

Key concepts

- In spite of the complexity of protein sequences, it appears that the number of ways in which polypeptides fold into their final tertiary structures is limited; structure and function are more highly conserved than in an amino acid sequence.
- Some protein folds are seen to carry out numerous biological functions, whereas others appear to have evolved to perform specialized activities.

The folded state results from the formation and association of secondary structural elements described earlier, together with intervening linker regions and turns to form the tertiary structure of the molecule. On the basis of his studies of the reversible folding–unfolding of ribonuclease, Anfinsen suggested that the information that determined the final folded state of a protein was encoded entirely within its amino acid sequence. Since then, there have been considerable efforts to unravel the complex processes involved in the formation of a folded structure from a string of amino acids. Although we still do not fully understand how this occurs, progress has been made through a combination of experimental and theoretical approaches.

The magnitude of the 'folding problem' was noted by Levinthal in the 1960s. Levinthal presented the problem, now known as 'Levinthal's paradox,' in the following way: If we consider a protein of a 100 residues and assume that the main-chain ϕ-φ angles can take one of three possible values (side-chain rotamers are completely ignored), then each peptide can adopt nine (3×3) possible conformations. Therefore, our 100-residue protein can adopt 9^{100} possible three-dimensional structures. Making a further assumption that changes in peptide conformation can occur on the femtosecond (10^{-15} s) time scale, finding the correct folded conformation would take, on average, ~10^{70} years, rather than the few seconds or minutes that might be actually observed in the test tube.

As we have seen, the number of possible three-dimensional structures for even a small protein is astronomical. Nonetheless, it is clear that proteins do only seem to adopt a relatively small number of tertiary structures. Broadly speaking, two classes of tertiary structure are discernable: fibrous and globular. Fibrous proteins, as their name suggests, are characterized by rather elongated architectures and are exemplified by 'structural' proteins such as keratin and collagen. In contrast, most tertiary structures fall into the second broad class and are generally referred to as 'globular,' reflecting their more spherical shape.

A question that fascinates many structural biologists is, how many different structures of globular proteins are there? The structural database, at present, contains some 40,000 structures (some of which may be, for example, mutant forms of the same protein). But how many of these represent different protein folds, and how many such folds remain to be determined? The importance of these questions has resulted in an increasing number of 'structural genomics' programs (*Section 1.16, What's next? Structural biology in the postgenomic era*) whose aims are, in part, to determine the structures of all possible **protein folds** in biology. This is not merely a stamp-collecting exercise but, if successful, will play an important role in our efforts to understand how a protein sequence determines its final three-dimensional structure.

The overall structure of large (>50 kDa) proteins can almost always be subdivided into combinations of smaller, compact entities that are generally referred to as **domains**, and are often thought of as segments that are capable of independent folding. Indeed, this notion has been extensively used to experimentally define domain boundaries within larger molecules by limited proteolytic digestion. In this method, small quantities of proteases are mixed with varying molar excesses of the protein of interest with the hope that flexible, exposed linker regions, or unfolded N- and/or C-terminal sequences, will be more easily cleaved than the compact globular domains (FIGURE 1.41).

If we assume that all globular proteins are formed from one or more 'domains,' then our 'fold' problem becomes one of deciding how many topologically discrete combinations of α-helices and β strands are likely to be represented in the structural proteome. A number of approaches have been taken to this problem, including the reduction of 'fold-space' into a periodic tablelike classification of likely topological arrangements of beta and alpha secondary structural elements. Despite the exponential increase in structure determinations, and the fact that large-scale structural genomics efforts often choose target proteins on the basis that they may contain a novel fold, the rate at which new folds are being revealed does not follow the same exponential behavior. This is, in part, related to an overarching question of how relationships between different structures are measured, defined, and classified. Regardless, it seems likely that as more and more variations in **topology** are revealed, a picture may emerge

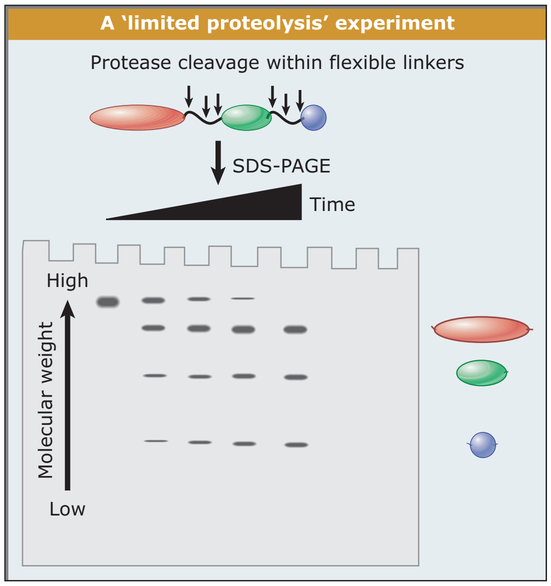

FIGURE 1.41 A 'limited proteolysis' experiment.

that describes 'fold-space' as a continuum of structures, each subtly different from the next, rather than the somewhat 'quantized' view that has been favored for so long.

To date, around 800 protein folds have been revealed and classified. It is clear, however, that a subset of these must have arisen early in evolution and occur in many hundreds of known protein structures. In general, these 'common' folds fall into two distinct classes: those that seem to possess functional versatility and appear in proteins with extraordinarily diverse biological activities, and those that appear especially well suited for a specific role. For example, the 'TIM barrel' that is made up of eight copies of one β strand and one α-helix (**FIGURE 1.42**) was first observed in an early structure of the metabolic enzyme triose phosphate isomerase, but has subsequently been seen in over 100 different structures of molecules with a diversity of largely enzymatic functions. Similarly, the all-beta immunoglobulin or 'Ig' fold, first seen in structures of the Fab fragment IgG, is now known to occur in molecules with diverse functions, such as chaperones and transcription factors. In contrast, the α-helical globin fold is exquisitely tailored to bind macrocyclic cofactors called metalloporphyrins (see Figure 1.62). More than 100 structures of globin fold proteins have been determined yet, remarkably, this appears to be their major function.

Given that only ~800 distinct protein folds have been seen in the structures determined to date, and that considerably greater diversity is seen in the vast database of primary sequence data, it is clear that their structure and function are much more strictly conserved than in an amino acid sequence. This situation arises through two distinct mechanisms of convergent and divergent evolution.

Convergent evolution was noted early in the history of structural biology. The X-ray structures of two proteolytic enzymes, trypsin and chymotrypsin, revealed a close relationship in both sequence and overall structure. In particular, the precise three-dimensional arrangement of a histidine-aspartate-serine triad of catalytic residues was seen to be almost identical at the active sites of each protein. It was with some surprise that the later structure of a completely unrelated protease, subtilisin, showed a completely different fold but retained the identical stereochemical arrangement of catalytic triad residues that, nonetheless, occur in a different relative order in the sequences of the two enzymes (**FIGURE 1.43**). Thus, in these enzymes, the same enzymatic function appears to have been 'invented' more than once in evolutionary history.

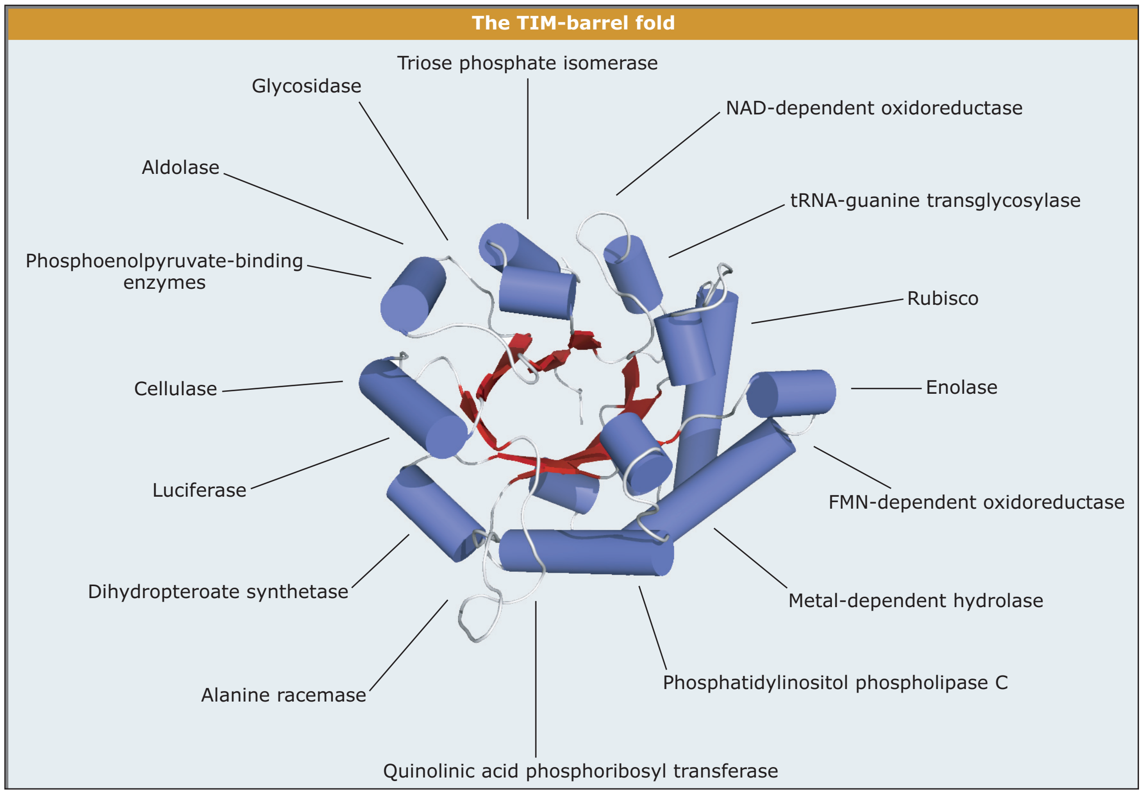

FIGURE 1.42 The TIM-barrel fold occurs in a variety of functional contexts. Image generated from Protein Data Bank file 8TIM.

Although this represents convergent evolution of the structural features of an active site, it would appear that convergence of overall protein fold may also have taken place. Proteins have been shown to have near-identical tertiary structures, yet share absolutely no detectable homology other than that expected for alignment of a pair of random sequences. This is something of a gray area, where it may be difficult or impossible to decide whether two structures have arisen by convergent or divergent processes.

Divergent evolution is most commonly observed and is manifested by conservation in sequence and overall structure, or the structure of a core segment, along with the position of functionally important residues (where function is also maintained). In its most generally accepted sense, divergent evolution implies that multiple, related protein sequences share a common evolutionary origin. Observed differences then arise from selective pressure to evolve, for example, differences in substrate specificity between members of a family of enzymes. Again, the serine proteases represent a good example, where trypsin and chymotrypsin share close structural homology, but differ in their preference for different classes of amino acids, C-terminal to the substrate's scissile bond.

In the pantheon of structural folds/motifs observed to date, a number of interesting 'outliers' have been observed. We started this section with the tentative assertion that related primary structures (i.e., amino acid sequences) must give rise to similar tertiary structures in the folded state. Although this is generally true, it has occasionally been seen that two sequences that are apparently closely related can fold in similar but topologically distinct ways, often with important functional consequences. For example, members of the KH family of RNA-binding domains have similar overall structures that, nevertheless, fall into several classes differing in the order of their alpha and beta secondary structural units. Nowhere is this phenomenon more obvious than in prion-related diseases that are caused by the aggregation of a small protein PrP, following a dramatic conversion from a predominantly helical form to a form containing a preponderance of β structure. Prions are discussed in more detail in *Section 1.15, Structure and medicine*.

A related phenomenon called cyclic permutation occurs when one or more secondary structural elements at one end of the molecule are transposed to the other end, where they form identical or near-identical interactions with remaining secondary structural elements (**FIGURE 1.44**). This is only possible because of the fact that, in many structures of diverse proteins, the N- and C-termini are observed to be located close together in space, despite being distant at the level of the linear amino acid sequence.

A number of examples of structures have now been reported that show exchange of one or more secondary structural elements between two, or rarely three or four, identical domains, an effect that is variously referred to as strand exchange, segment swap, or domain swap (Figure 1.44). This arrangement obviously im-

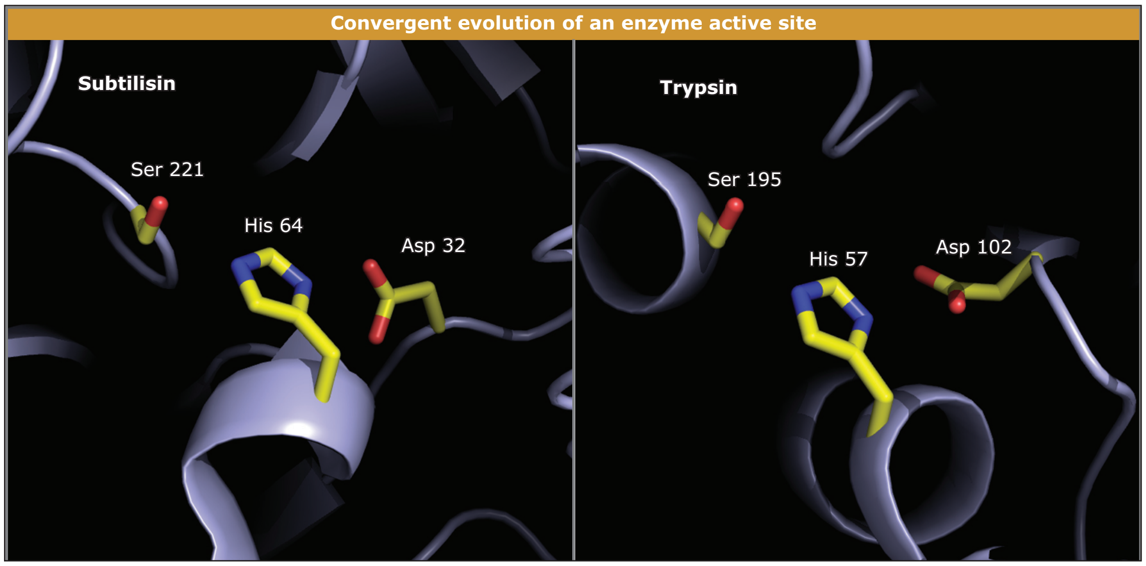

FIGURE 1.43 Convergent evolution of an enzyme active site. Images generated from Protein Data Bank files 1N65, 1GNS.

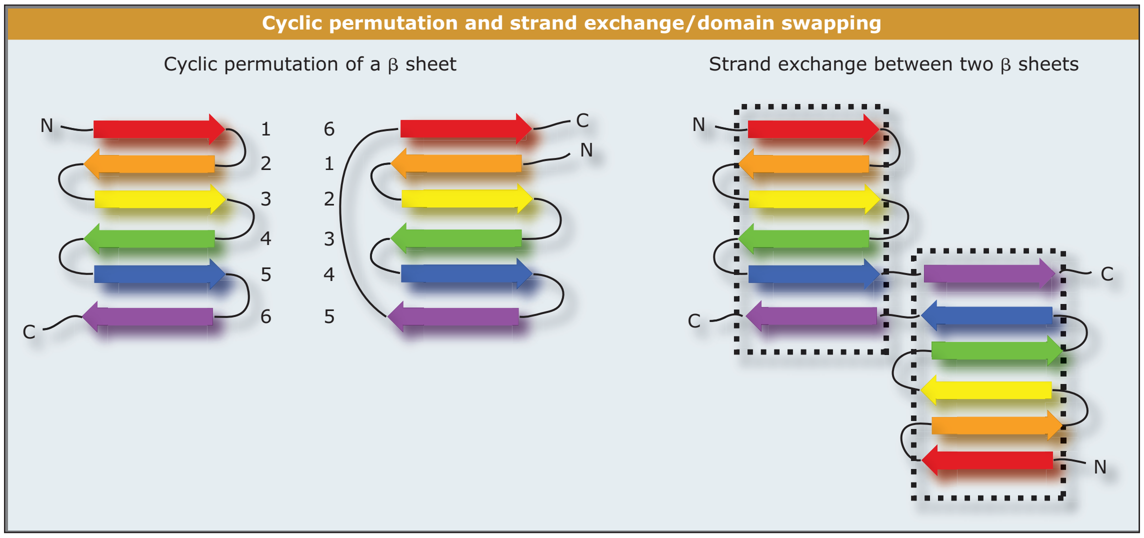

FIGURE 1.44 Schematic representation of cyclic permutation and strand exchange/domain swapping.

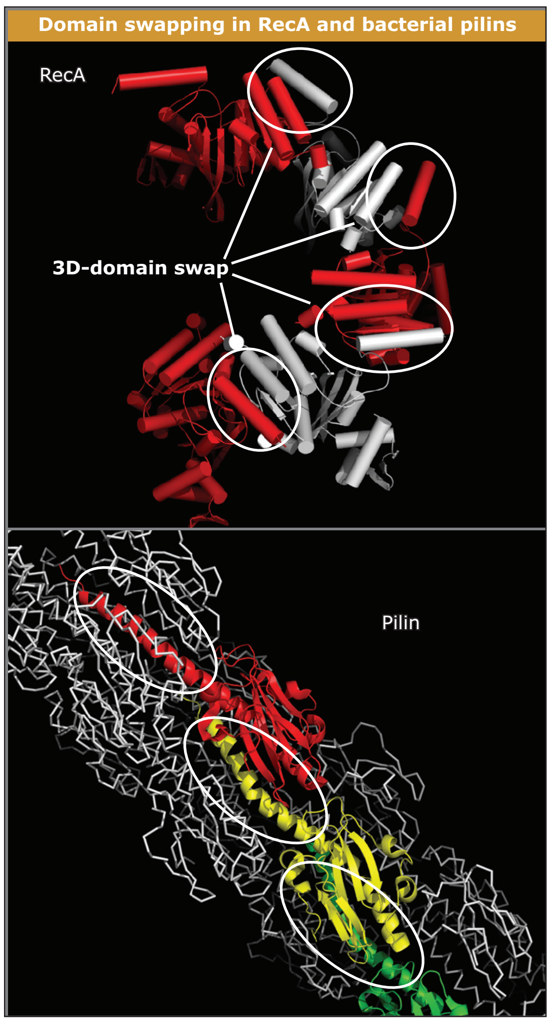

FIGURE 1.45 Domain swapping in two example structures, RecA and bacterial pilins. Images generated from Protein Data Bank files 2REB, 2HIL.

plies that self-association (dimer-, trimer-, tetramerization, or even polymerization) may be functionally important. The biological relevance of segment swapping is not always clear, however, and because of its predominance in crystal rather than NMR structures, it has sparked several arguments about the relative merits of the two approaches! In many cases, this phenomenon has been shown to be artefactual, but nonetheless it does play well-characterized roles in biological processes such as RecA filament formation and the assembly of bacterial pili (FIGURE 1.45).

As should be clear from the structures presented so far, if one conceptually held the N- and C-termini of most proteins in either hand and pulled, the structure would unravel back to the original linear chain of amino acids. That is, unless different parts of a protein chain are covalently crosslinked—for example, by disulfide bond formation, as observed in so-called cysteine-knot structures—proteins don't form knots! Until relatively recently, this was the widely held view and for good reason: a knotted structure would have considerable implications for folding pathways. In 2000, however, bioinformatics analysis detected a disulfide-independent knotted structure in a plant protein, acetohydroxy acid isomeroreductase (FIGURE 1.46). Although a few additional examples have since been discovered, these 'deep-knotted' structures remain uncommon and a real structural curiosity!

A deep-knotted protein

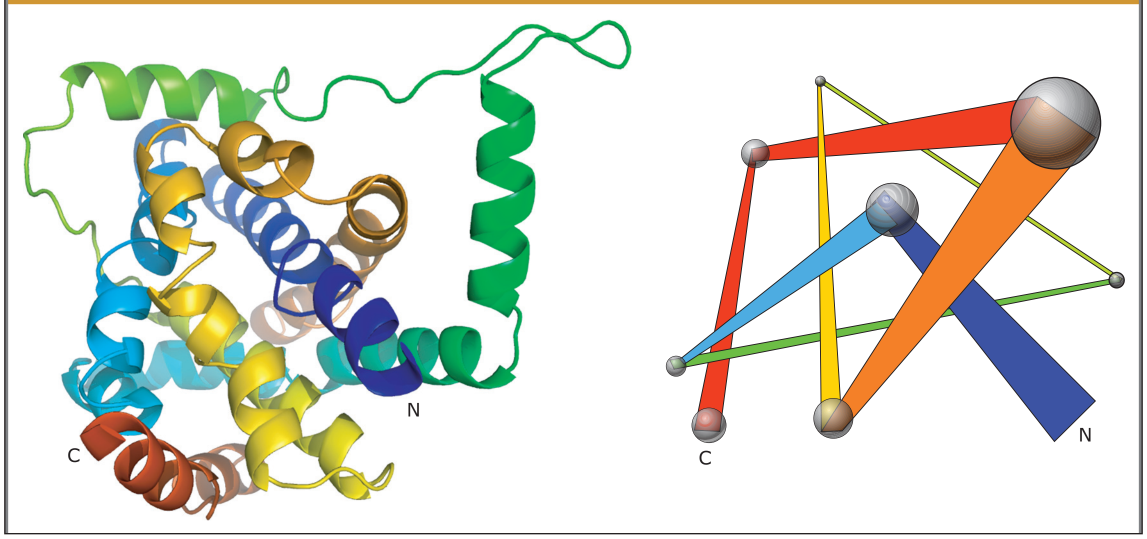

FIGURE 1.46 A deep-knotted protein (left) with a schematic representation (right) to show how the knot is formed. Image generated from Protein Data Bank file 1YEV.

1.9 Modular architectures and repeat motifs

Key concepts

- Many eukaryotic proteins do not consist of a single globular domain with a single biological activity. Instead, they consist of multiple domains or modules, each with a specific function, connected together like beads on a string.
- Modular architectures provide evolutionary flexibility, where different functions may be added or removed from a protein by the insertion or deletion of modules with specific activities.

Any complex biological process, such as the biosynthesis of an amino acid, may require the activity of many proteins. In prokaryotes, genes are most often organized in operons, whereby they are transcribed as a single, polycistronic mRNA. This elegant mechanism ensures that the expression of a particular set of genes can easily be temporally coordinated in response to a single biological stimulus. In contrast, eukaryotic genes are generally isolated, and expression of each protein occurs from an individual monocistronic mRNA.

Genome sequencing has revealed that eukaryotic proteins are often constructed as a linear array of protein '**modules**' joined together, much like beads threaded onto a string. Here, each module has a specific function that contributes to the overall activity of the 'polyprotein' within which it resides. Individual modules may play specific roles that are otherwise structurally or functionally independent of those of their associated partner domains. In contrast, the activities of individual domains within a modular protein may be interdependent and may be closely associated through intramolecular domain–domain interactions.

Analysis of available genome sequence data has revealed the existence of several hundred protein modules with diverse biological functions. These modules have been classified and curated in several databases (e.g., SMART, PFAM, and ProDom). Although many have been characterized with respect to structure and biological activity, many more remain to be investigated, and yet more to be discovered.

The degree of modularity is, in some cases, breathtaking. For example, the Vav proto-oncogene is a regulator of Rho family small GTPases and contains a total of seven modules that variously mediate binding to lipids, phosphotyrosine, proline-rich motifs, and actin, in addition to a domain that mediates guanine-nucleotide exchange on small GTPases of the Rho family (FIGURE 1.47). This architecture is evolutionarily very flexible, affording the possibility of facile generation of proteins with novel functions though domain/exon shuffling. Structures of many of these isolated domains are now available and have generally shown that the N- and C-termini are often located close together in space. Thus additional domains may be added by insertion into the linking regions between preexisting modules without any structural disruption.

Perhaps the most extreme example of the use of modular architecture is seen in the sarcomeric protein titin. The human protein contains around 37,000 amino acid residues and

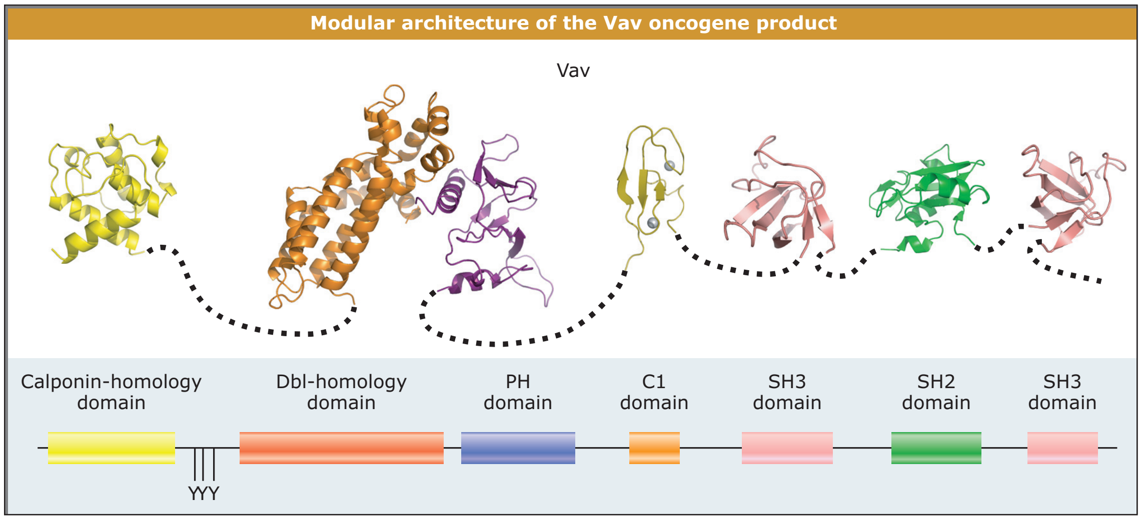

FIGURE 1.47 Modular architecture of the Vav oncogene product. This representation is deceptive and a number of physiologically important intramolecular interactions between domains are known to occur. Images generated from Protein Data Bank files 1AAZ, 1KBE, 1GCP, 1DBH.

has a molecular mass of ~4 MDa. Sequence and structural analysis has identified ~300 immunoglobulin and fibronectin domains, along with a kinase domain at the C-terminal end and stretches of sequence unique to titin.

In many cases, the precise biochemical roles of a given domain can only be inferred from known activities of well-characterized homologues. In a few examples, one or more domains within a protein may exhibit dual functionality through both intramolecular and intermolecular interactions. This is true of Vav (Figure 1.47), but is perhaps best illustrated by the Src-family protein kinases (**FIGURE 1.48**). These molecules all contain three distinct functional domains: an N-terminal SH3 is followed by a short proline-rich linker, which connects to an SH2 domain and a tyrosine kinase domain. Regulation of the activity of Src kinases critically depends on the phosphorylation status of a highly conserved tyrosine residue at the extreme C-terminus. When phosphorylated, the phosphotyrosine (pTyr) binds to the central SH2 domain, and additional interactions between a proline-rich sequence and the N-terminal SH3 domain serve to maintain the kinase in an inactive conformation. Dephosphorylation of the tyrosine and/or competition for binding by phosphotyrosine residues on Src-interacting proteins disrupts these intramolecular contacts and results in activation of the kinase domain.

An additional level of modularity can be seen in protein domains that in turn consist of many tandem copies of short repeating sequences (ANK, LRR, TPR, WD40, HEAT/ARM, pumilio, and so forth). These molecules are functionally diverse; they may play purely structural roles, mediate protein–protein or protein–ligand interactions, or a combination of all of these. As is true of protein structures in general, modular-repeat proteins can be loosely classified on the basis of their secondary structural content. For simplicity of presentation we will consider two classes: those that have predominantly α-helical structure (although some β content may be evident) and those in which β structure dominates.

A major characteristic of the helical-repeat

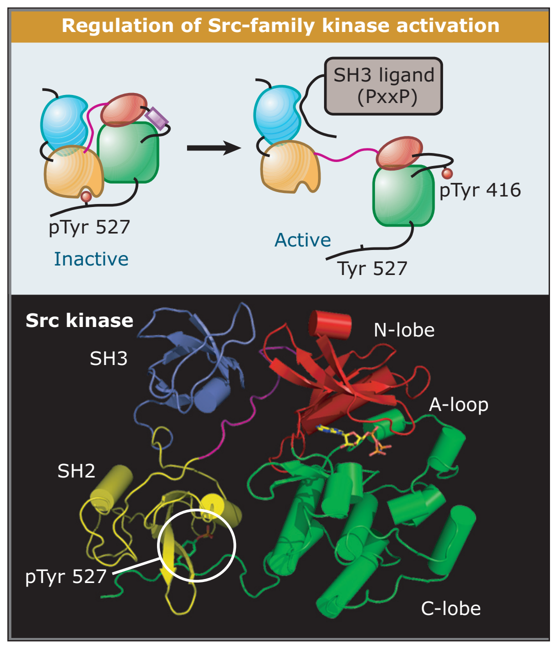

FIGURE 1.48 Intra- and intermolecular interactions regulate Src-family kinase activation. Images generated from Protein Data Bank files 2CRH, 2SRC.

proteins is that the number of repeats within any one molecule is highly variable, a feature presumably generated through recursive gene duplication. Evolutionarily, this type of tertiary architecture is attractive, given that it results in molecules with extended interaction surfaces that can be tailored in size by the simple addition or subtraction of repeating units.

Helical repeating motifs were first noted in comparisons of budding and fission yeast transcription factors, and were initially known as Swi6/Cdc10 repeats after the molecules in question. Subsequently, they were noted in the ankyrin cytoskeletal proteins, which may contain up to 50 repeats of the characteristic 30- to 35- residue motif and are now most commonly known as ankyrin (ANK) repeats. At present, over 3500 ANK repeat proteins containing more than 15,000 individual motifs are known. Structural studies have revealed the ANK repeat to consist of a short, tight β-hairpin-like turn followed by a pair of antiparallel α-helices and a partially conserved linker region. The helices of each repeat pack against those of the neighboring motif through conserved hydrophobic residues. This generates an elongated structure that is both curved and twisted around the long axis of the stack (**FIGURE 1.49**). The hairpin loops extend away from the helical bundle at an angle of approximately 90°, forming a groove that may be used for interaction with a variety of protein targets and other ligands (Figure 1.49). Although this is the most common mode of ANK domain binding, it is clear that other regions of the surface may also be utilized for inter- and/or intramolecular interactions in specific contexts.

Other repetitive helical-repeat motifs include leucine-rich repeats (LRR), tetratricopeptide (TPR) repeats, Pumilio repeats that are in turn related to the HEAT/ARM family of motifs, and others. Association of multiple copies of these motifs can give rise to a variety of tertiary structures, and in the majority of cases where the activities have been well characterized, helical-repeat architectures assemble to produce extended surfaces for interaction with protein partners or other ligands. LRRs form a characteristic horseshoe shape (**FIGURE 1.50**), which provides a highly concave binding surface. TPR and HEAT/ARM repeats associate to form a superhelical array of helical motifs, which results in a binding groove that spirals along the length of the molecule. In the case of TPR repeats, the twist on the helical stack is rather small. In contrast, HEAT/ARM proteins tend to show a superhelical twist that can be quite spectacular and of a size that can wrap around entire proteins (Figure 1.50).

In contrast to the diversity seen in helical-repeat architectures, β-repeat motifs are less common and essentially fall into two classes known as β-propellors and β helix structures. The WD40 domain is named after a conserved Trp-Asp dipeptide that is embedded in a repeating sequence of ~40 residues. Although the number of repeats is quite variable, seven are most commonly observed. The first structure of a WD40 repeat protein was that of the β subunit of heterotrimeric G-proteins, which revealed that each motif folds into a four-stranded β sheet. These sheets then pack together to form an extended barrel-like structure that resem-

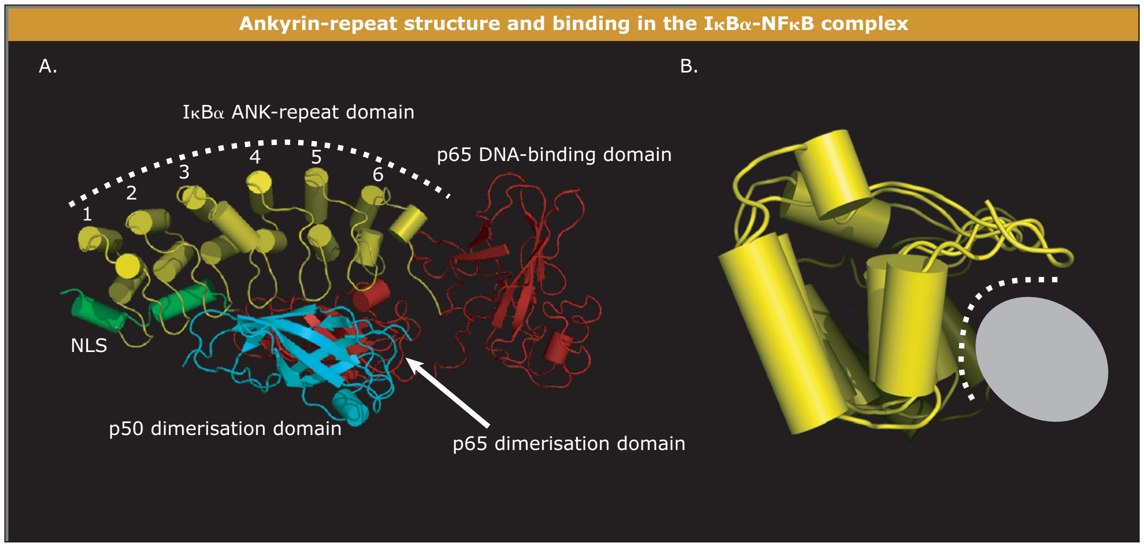

FIGURE 1.49 Ankyrin-repeat structure and binding in the IκBα-NFκB complex. The NLS binds into the ankyrin groove, but binding occurs in other regions of the ankyrin repeat stack. Image generated from Protein Data Bank file 1NF1.

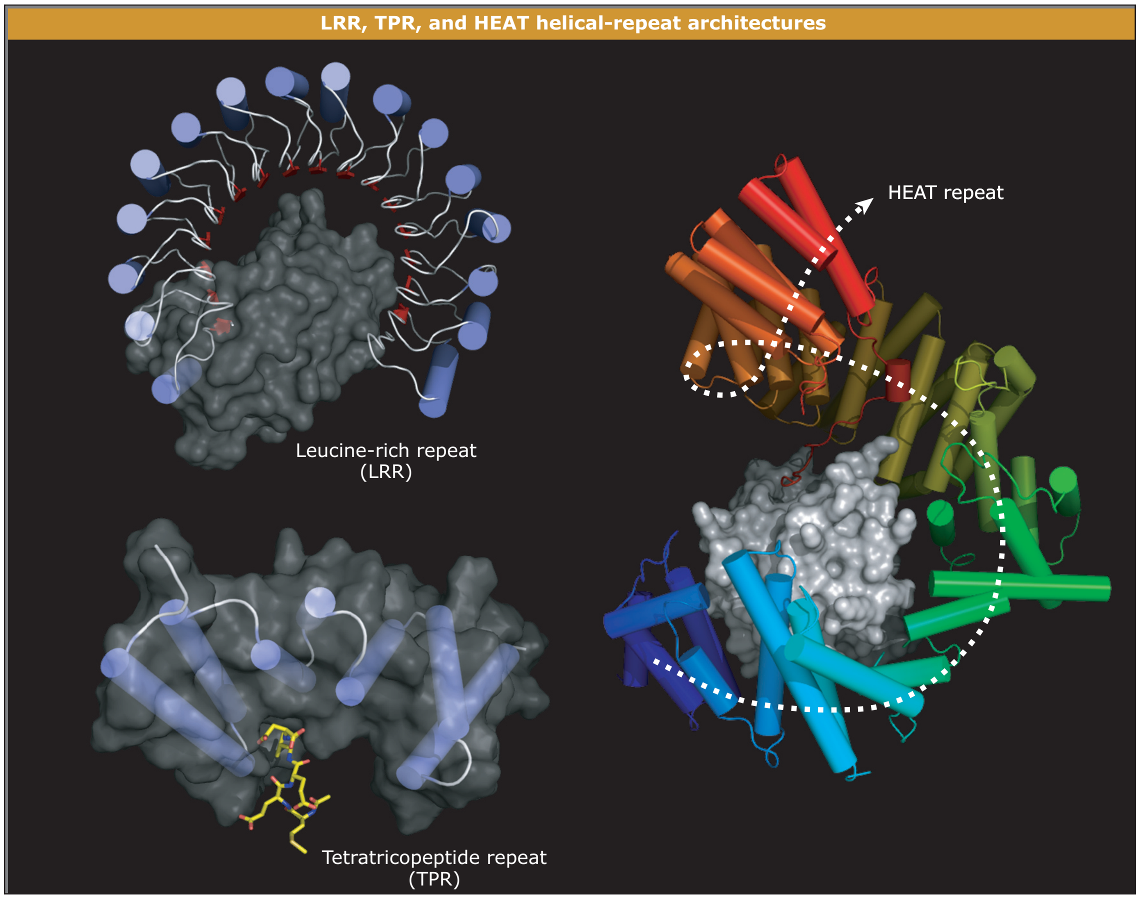

FIGURE 1.50 LRR, TPR, and HEAT helical-repeat architectures. Images generated from Protein Data Bank files 1DFJ, 1WA5, 2BUG.

bles a propeller in which each repeat forms an individual 'blade.' Functionally, WD40 domains play a variety of roles in mediating protein-protein interactions. Most recently, it has become clear that they constitute a family of modules that may bind to peptide motifs that are specifically modified posttranslationally (*Section 1.11, Posttranslational modifications and cofactors*). **FIGURE 1.51** shows an example of this phenomenon where the WD40-repeat domain of WDR5 recruits lysine methytransferases to dimethylated lysine 4 of histone H3 by binding the modified lysine side-chain through a recognition pocket formed at the center of the β-propeller domain.

RCC1 (regulator of chromatin condensation 1) is a multifunctional signaling molecule that localizes to the nuclear compartment. Its major role appears to be as a nucleotide exchange factor for a small GTPase Ran that regulates the trafficking of cargo into and out of the nucleus. Structurally, RCC1 resembles WD40-repeat domains in that it adopts a seven-bladed β-propeller-like fold. There is, nonetheless, no detectable sequence homology between WD40 and RCC1 repeat motifs and, as can be seen from the complex of RCC1 with nucleotide-free Ran (represented as a transparent surface in Figure 1.51), they can be distinguished by differences in the topological arrangement of β strands, and the inclusion of a short region of α-helix that packs at the outer edge of each of the blades.

As we have seen, WD40 and RCC1 repeats tend to form closed, circular arrays. β-repeat structures, however, are also able to form the kinds of highly extended structures that are generated in helical-repeat molecules such as karyopherins (Figure 1.50). Such an example is the β-helix architecture where repeating pairs or triplets of β strands associate to form 'helical' arrays that can be highly elongated. The β-helix fold has been observed in many different proteins with a plethora of functions. For example, pertactin, a virulence factor that mediates adhesion of the pathogenic *Bordetella pertussis* bacterium with host cells (Figure 1.51), is constructed from consecutive strands that, conceptually, form triplets and pack around a right-handed spiral (Figure 1.51 inset). Conversely, other examples such as insect antifreeze proteins have similar structures, but with a left-handed twist.

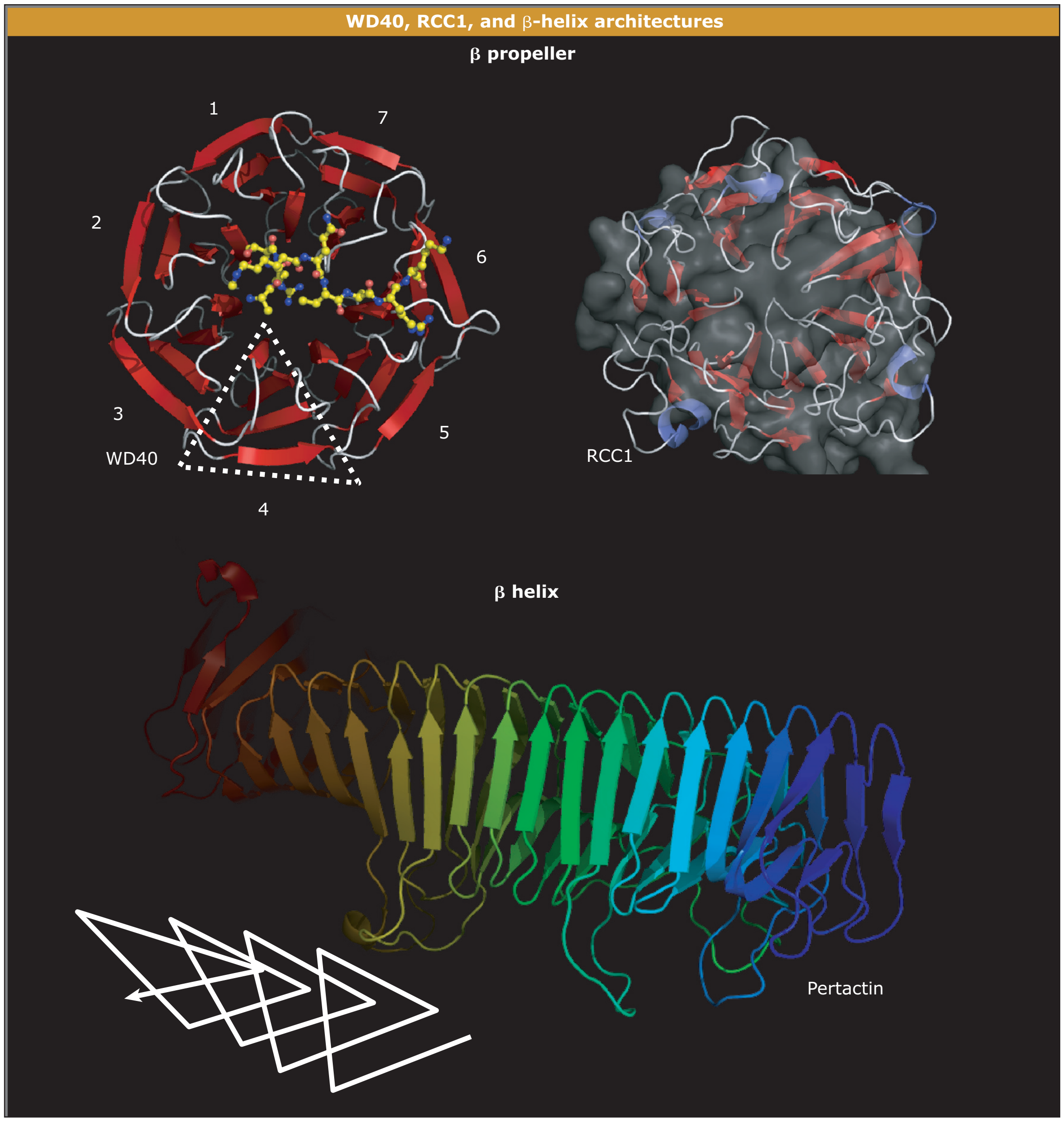

FIGURE 1.51 WD40, RCC1, and β-helix architectures. Images generated from Protein Data Bank files 2HK6, 1DAB, 1I2M.

1.10 Quaternary structure and higher-order assemblies

Key concepts

- The quaternary association of individual proteins is extremely common. The simplest quaternary structures are dimers of two identical monomers, but considerably complexity is observed and assemblies such as viral capsids may contain hundreds of protein chains.
- Protein assemblies facilitate allosteric and cooperative effects that most often arise from structural changes within the assembly and provide for exquisitely precise regulation of activity.
- Changes in quaternary structure may involve the reversible binding of regulatory subunits to a stable core assembly.

Quaternary structure describes the fact that many proteins do not exist or function as monomers, but associate to form oligomers. The simplest form of oligomer arises from the **homotypic** association of two identical subunits to form a dimer (more precisely a homodimer). Alternatively, two different proteins may bind to form a heterodimer (a **heterotypic** association). Obviously, any number of combinations is possible and, as we will see, many variations are observed to occur in biological systems.

Most dimers arise from the interaction between identical (or near-identical) surfaces on each **protomer**. Such interactions are known as **isologous** and are the most commonly observed in structures of protein oligomers. The interaction surfaces are buried in the dimer, and as a result they are not available for interaction with another subunit. As shown in FIGURE 1.52, isologous interactions are not restricted to dimers, but can and do occur in higher-order homo-oligomers (trimers, tetramers, pentamers, and so forth). In contrast, **heterologous** interactions between protomers involve distinct surfaces that do not overlap. In its simplest form, this situation is less commonly observed because it implies that the protomer will become infinitely polymerized. Additional rotational symmetry, however, can result in 'closing' of the system, and this closed heterologous oligomerization is much more prevalent. Nonetheless, 'nonclosed' heterologous oligomerization (effectively polymerization) of proteins such as actin and tubulin is extremely important in muscle contraction and formation of the actin cytoskeleton and microtubules.

The question arises: Why be oligomeric? In some cases this is a difficult question to answer, because a functional relevance may not be obvious from the structure or from known biological function. It is clear, however, that oligomerization can confer a degree of structural or chemical stability. From an evolutionary point of view, formation of oligomers, and particularly hetero-oligomers, affords a great deal of functional and architectural flexibility, in a similar manner to the modular architectures that were discussed in the preceding section.

The combination of subunits that constitute a particular protein complex may provide different structural characteristics that are tailored to specific biological activities. In addition, activity and specificity can be modulated through regulated association and dissociation of acces-

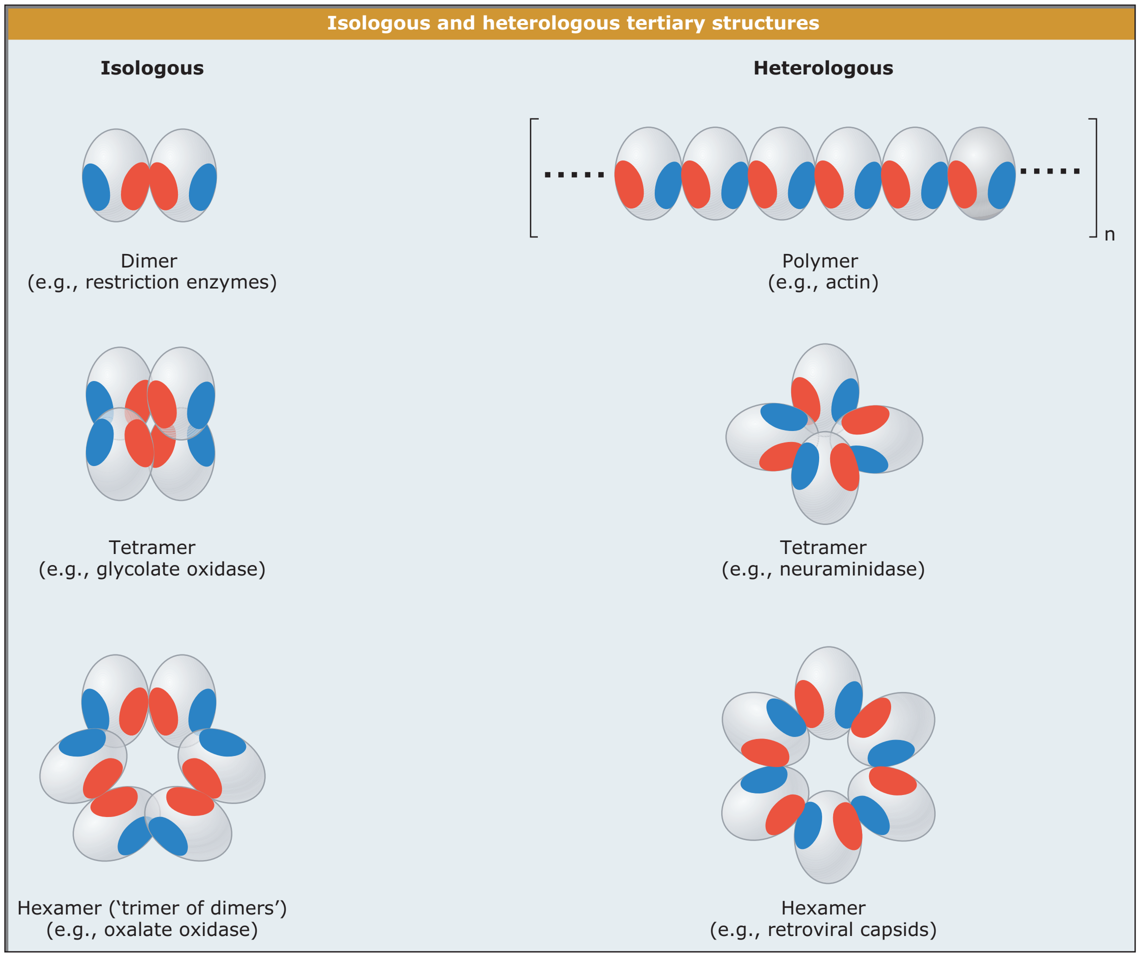

FIGURE 1.52 Isologous and heterologous tertiary structures.

sory subunits. An excellent example of such a system is provided by RNA polymerases, high-molecular-weight complexes that carry out the fundamental process of gene transcription into mRNA. Prokaryotic RNA polymerases have a basic structure composed of five subunits ($\alpha_2\beta\beta'\omega$; the subscript denotes that the α subunit is, itself, a dimer) that carry out the core activities of DNA binding, RNA synthesis, and translocation along the DNA template. The stable, core assembly binds nonspecifically to DNA. Binding of an accessory subunit, however (σ, which directly interacts with DNA in a sequence-specific manner), is required to allow the RNA polymerase **holoenzyme** to be able to recognize sites on chromosomal DNA at which transcription should begin (called promoter regions). Through the binding of different σ subunits, specificity for different classes of promoters (with different DNA sequences) can occur, allowing precise temporal regulation of transcription of particular classes of genes (**FIGURE 1.53**). RNA polymerase is just one example of an asymmetric quaternary arrangement of multiple protein subunits that shows both hetero- and homotypic character. As we will now see, however, symmetrical quaternary arrangement has been exploited by evolution in a number of ways.

Pyruvate dehydrogenase is a three-enzyme complex comprising E1 (pyruvate dehydrogenase), E2 (dihydrolipoyl transacetylase), and E3 (dihydrolipoyl dehydrogenase), which catalyze five distinct reactions in a pathway that leads to oxidative decarboxylation of pyruvate to acetyl CoA. The bacterial enzyme consists of a core of eight E2 trimers that interact to form the corners of a cube of approximate dimensions $80 \times 80 \times 80$ Å^3 (**FIGURE 1.54**). The cube is further decorated with twelve E1 dimers and six E3 dimers to form a complex of around 4.5 MDa. The size and complexity are truly staggering, but considerable advantages are achieved in these large multienzyme aggregates. Structural intimacy of different catalytic domains provides opportunities for tight and coordinated regulation through, for example, binding of allosteric effector molecules, or by posttranslational modifications such as phosphorylation. Furthermore, the multiprotein ‘lattice’ generates a cagelike environment that protects the products of one reaction from the unwanted attention of other enzymes and cytoplasmic components, and maintains substrates in close proximity to active sites. Alone or in combination, these effects can provide for considerable enhancement of catalytic rate/efficiency. As mentioned earlier, the oxygen-carrying molecule hemoglobin was among the first protein structures to be determined. Hemoglobin forms an $\alpha_2\beta_2$ heterotetramer of subunits, each of which carries a single heme (Fe-protoporphyrin IX) prosthetic group as the site of binding of diatomic ligands such as oxygen and carbon monoxide. As we will see in *Section 1.12, Dynamics, flexibility, and conformational changes*, the α and β subunits are not functionally isolated, but are, in contrast, highly coupled. Loading of oxygen onto successive subunits progressively increases the affinity of the unfilled sites such that the oxygen affinity for the fourth and final subunit is increased several hundred times compared to the uncharged (deoxy) hemoglobin tetramer molecule.

In some cases, the symmetry of a protein in a homomeric complex is correlated with structural characteristics of its binding partner. Restriction endonucleases are the workhorses of molecular biological techniques due to their absolute specificity for particular sequence mo-

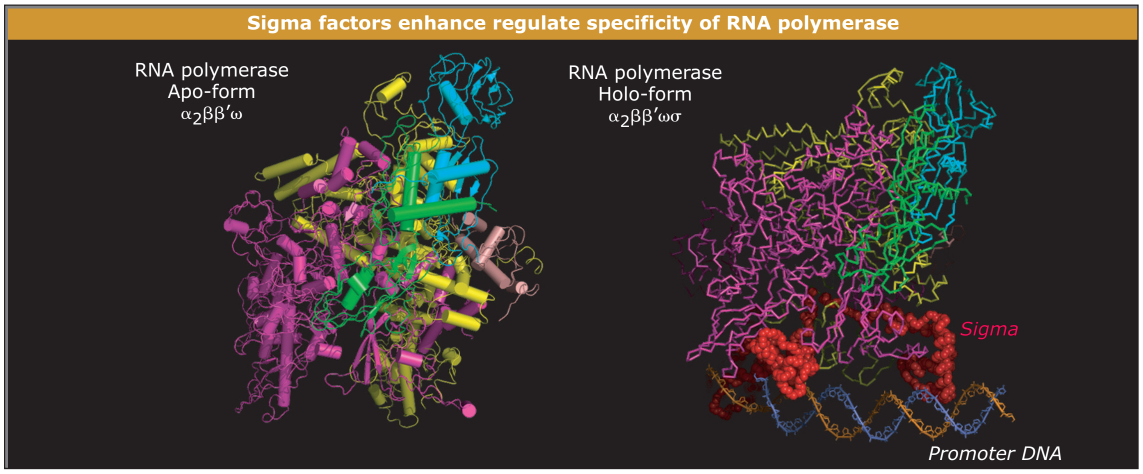

FIGURE 1.53 Promoter specificity of RNA polymerase may be regulated by binding of sigma factors that supply additional, sequence-specific DNA interactions. Images generated from Protein Data Bank files 1HQM, 1L9Z.

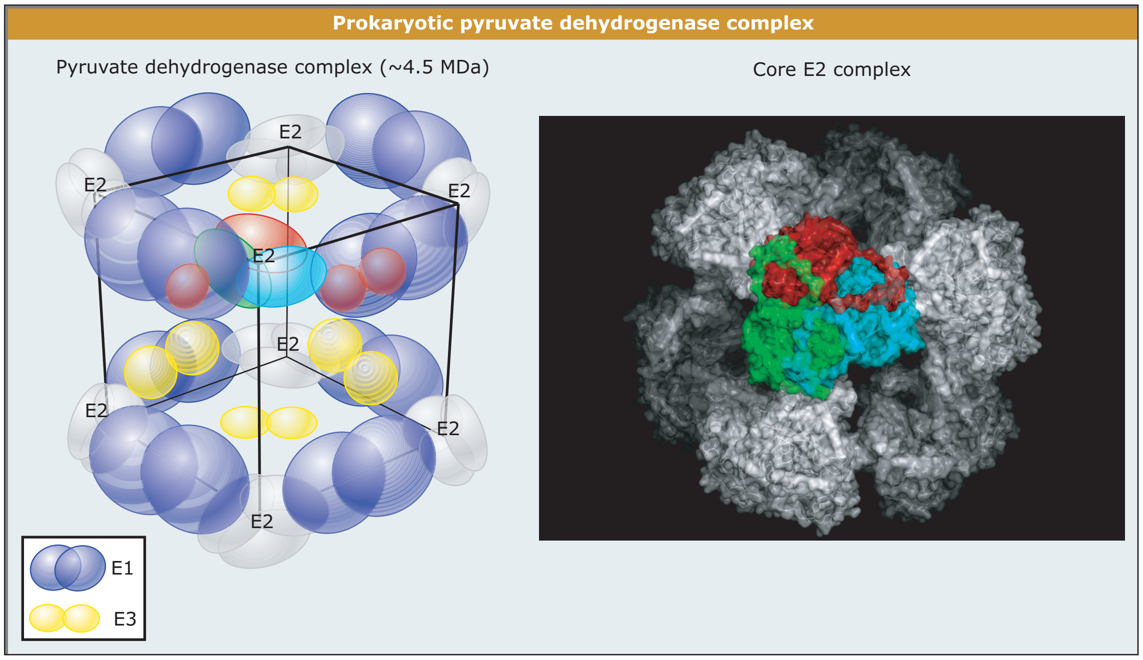

FIGURE 1.54 The prokaryotic pyruvate dehydrogenase complex. The locations of the E1, E2, and E3 subunits are shown on the left, along with the X-ray structure of the core, cubic E2 complex. Image generated from Protein Data Bank file 1DFM.

tifs in DNA. The major (and arguably the most useful) class of restriction enzymes includes those that recognize 'palindromic' motifs for which the sequence of bases of one strand of double-stranded DNA read in the standard 5′ to 3′ direction is exactly the same when read on the opposite strand (FIGURE 1.55). The astute observer will note that this arrangement generates a twofold axis of symmetry at the center of the double-stranded sequence motif. The remarkable ability of restriction endonucleases to cut the phosphodiester backbones of both strands at identical positions in the sequence is achieved straightforwardly by dimerization of the catalytic domains of the enzyme. This generates a twofold axis of symmetry that coincides with that of the DNA substrate in the specific and catalytically competent protein–DNA complex.

An impressive use of symmetry in large protein assemblies is seen in the structures of viral capsids. In all cases, many copies of one or a small number of protein chains are used to build a viral shell that may be 1000 Å in diameter or more, which is necessitated by the obvious problem that viral genomes must be relatively small in order to be accommodated within the capsids that they ultimately encode.

In broad terms, viruses can be classified in terms of the architecture of their shells. Enveloped viruses, such as human immunodeficiency virus (HIV) and influenza, are coated with a lipid bilayer, ultimately derived from the plasma membrane of the infected cell within which the virus was replicated and shed. Viral proteins involved in binding to receptors displayed on the surface of target cells are embedded within this bilayer. These viruses will not be considered further.

Of the nonenveloped viruses, two major architectures are seen, which are generally referred to as helical, and icosahedral or spherical. Tobacco mosaic virus is the archetypal helical virus, and its capsid is formed from ~2100 copies of a single protein subunit of 154 amino acids. These are arranged to form a helical rod with 16 copies per turn and a total length of ~3000 Å.

Icosahedral viruses have been by far the most intensively studied by X-ray crystallography and cryo-EM methods. The icosahedron is one of the few ways of symmetrically assembling identical subunits into a roughly spherical shape. It has 20 triangular faces and contains fivefold, threefold, and twofold axes of symmetry (532 symmetry). Each face has threefold symmetry and can thus contain a minimum of three identical subunits (FIGURE 1.56). Therefore the simplest icosahedral virus (such as plant satellite viruses) must contain 60 (20×3) subunits with each located in an identical environment to the other 59. In order to construct larger shells, the number of subunits must increase and, clearly, this number must be a multiple of 60. In fact,

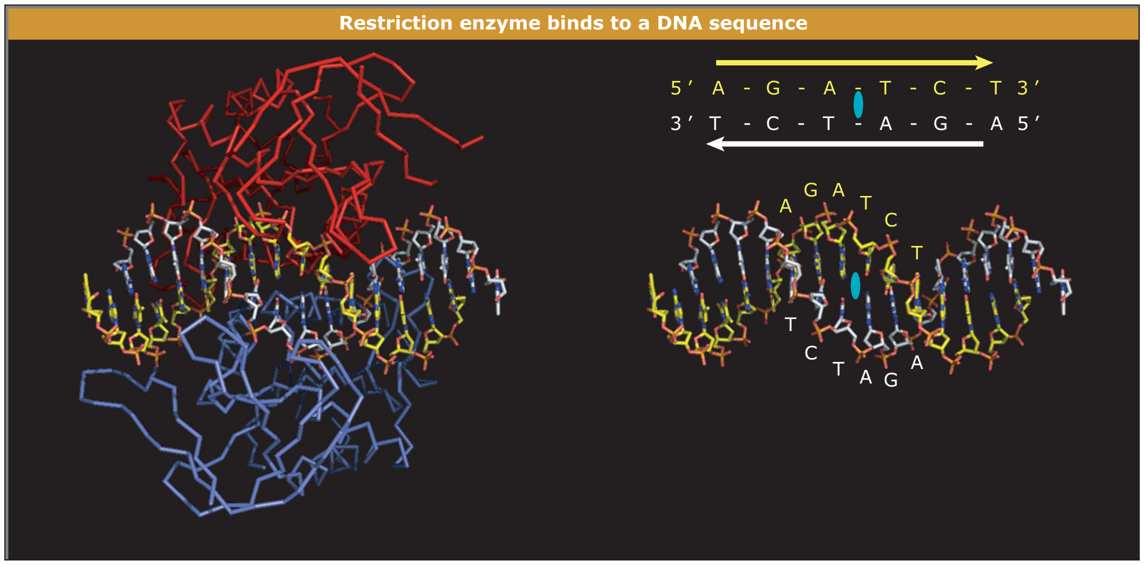

FIGURE 1.55 A twofold symmetric restriction enzyme binds to a twofold pseudosymmetric DNA sequence. Image generated from Protein Data Bank file 1DFM.

it was suggested by Donald Caspar and Aaron Klug in the 1960s that only certain multiples of 60 are consistent with formation of closed spherical shells (1, 3, 4, 7, 13, ...). Caspar and Klug termed these multiples triangulation or 'T' numbers. Thus a T = 3 virus (e.g., tomato bushy stunt virus) contains 180 subunits, whereas a T = 13 virus such as reovirus contains 780! Caspar and Klug also invoked the notion of 'quasi-equivalence' to explain how the same protein must be able to form more than one type of contact in order to form an icosahedral shell in spherical viruses with $T > 1$. It was not, however, until the structure of a small T = 3 plant virus—tomato bushy stunt virus—was solved by Stephen Harrison and coworkers that the structural basis of quasi-equivalent packing was revealed.

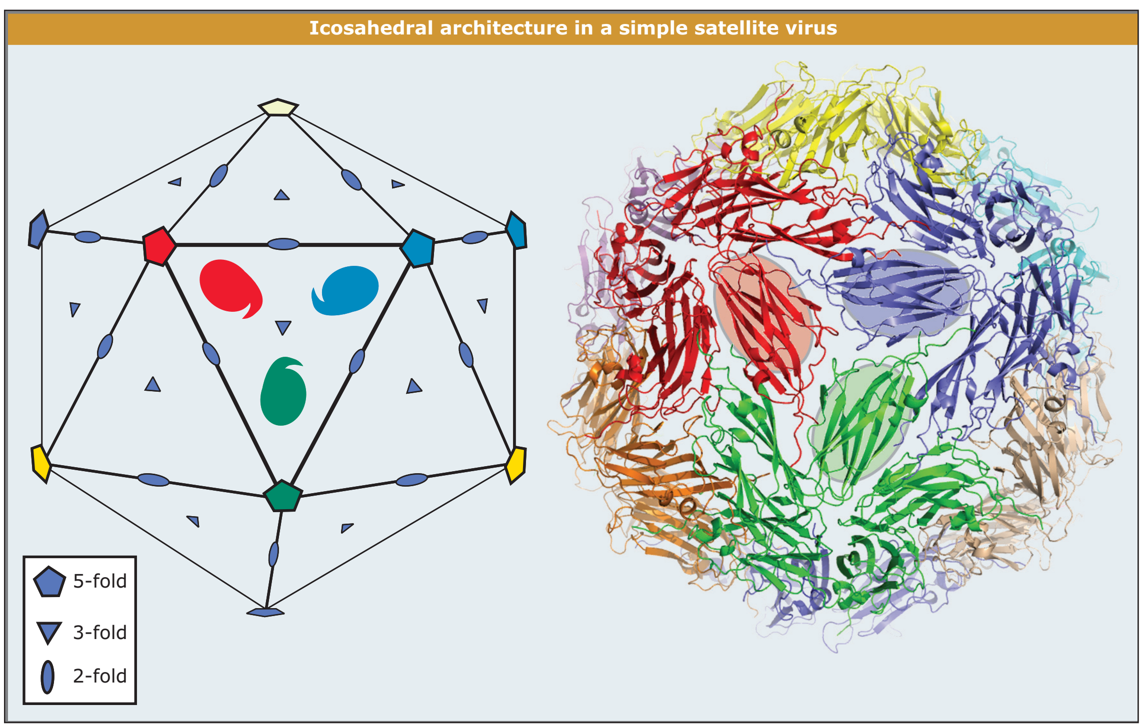

FIGURE 1.56 Icosahedral architecture in a simple (T = 1) satellite virus. Image generated from Protein Data Bank file 1STM.

1.11 Posttranslational modifications and cofactors

Key concepts

- Posttranslational modifications are often the final step in the production of an active protein or enzyme.
- Many different modifications are known, ranging from proteolytic cleavage to the addition of chemical groups.
- Modifications exert a variety of effects and may induce conformational changes or generate signals for the formation of protein–protein complexes.
- Posttranslational modifications may either activate or inhibit biological activity and are often reversible.

So far, we have seen how proteins are produced as amino acid chains that fold into a myriad of different tertiary and quaternary structures. In some cases, further processing of the polypeptide chain must occur, which can happen in several ways. For most proteins, posttranslational chemical modifications are the rule rather than the exception. Posttranslational modifications can have a variety of different effects on the structure of proteins and many aspects of their behavior. Indeed, a single modification may exert all of these effects on certain proteins, or a number of modifications of a single protein may occur. Furthermore, these need not occur at the same time and may take place at different stages of a protein's lifetime. An exhaustive description of all modifications known at present is unrealistic here. Instead, this section will focus on some specific examples in order to give a flavor of the biological and structural versatility and flexibility that they provide.

Proteolysis is one of the most dramatic posttranslational modifications, and is seen in the activation of some precursors of proteases, hormones, viral polyproteins, and other molecules (**FIGURE 1.57**). Functionally, the necessity for proteolytic activation most often is related to a need for precise and timely regulation. This is no better demonstrated than by proteases of the blood coagulation cascade, where inappropriate activation can lead to thrombosis that may ultimately be fatal. In general, the protein is initially translated and folds into a defined tertiary structure that is, nonetheless, biologically inactive. This precursor may be denoted with the prefix 'pro' to distinguish it from the mature active form. In the case of molecules with enzymatic activity, the inactive precursor may be referred to as a **zymogen**. Examples would include proinsulin and procaspase or the zymogen forms of trypsin and chymotrypsin (trypsinogen and chymotrypsinogen). Similarly, a family of cysteine-aspartyl proteases, the caspases, are activated by proteolysis to release the enzymatically proficient form in response to a variety of proapoptotic signals. Once activated, the so-called effector caspases are able to cleave, and inactivate, a variety of downstream proteins to initiate cell death.

Many known modifications have interesting and important effects on solubility, localization, and biological/chemical stability. For example, proteins that are released into the extracellular milieu, or are bound to the external surface of the plasma membrane, are often glycosylated on asparagine (N-linked) or serine (O-linked) residues (**FIGURE 1.58**). This may directly aid in protein folding, protect against proteolytic attack, provide immune surveillance, and generate binding sites for interacting partners. This diversity of effects arises, in large part, from the complexity of glycosylation patterns that may involve a number of different sugars and glycosidic linkages. The extent of the modification may be so great as to represent up to 40% of the overall mass of the glycoprotein. This, plus the heterogeneity of the attached sugar chains, provides an enormous barrier to high-resolution structural studies. In most cases, crystallization of highly glycosylated molecules is only possible after extensive enzymatic deglycosylation, removal of known glycosylation sites by mutagenesis, inhibition of glycosylation during recombinant protein expression, or a combination of all three.

Of greatest interest here are the variety of modifications that directly or indirectly influence protein structure and activity. Phosphorylation most often occurs on the hydroxylated amino acids, tyrosine, threonine, and serine. It can, however, occur on histidine and aspartate residues, but in these cases is highly unstable. It is the most prevalent posttranslational modification that occurs in human cells, and it has been estimated that ~30% of all proteins in the human proteome are phosphorylated at some stage during their lifetime! The role of phosphorylation in driving conformational change is best exemplified by protein kinases themselves, and we have already seen an example of how this occurs in the Src kinase family (Figure 1.48). This example also highlights the fact that phosphorylation can create binding sites for interaction with

Proteolytic processing of zymogen to form α-chymotrypsin

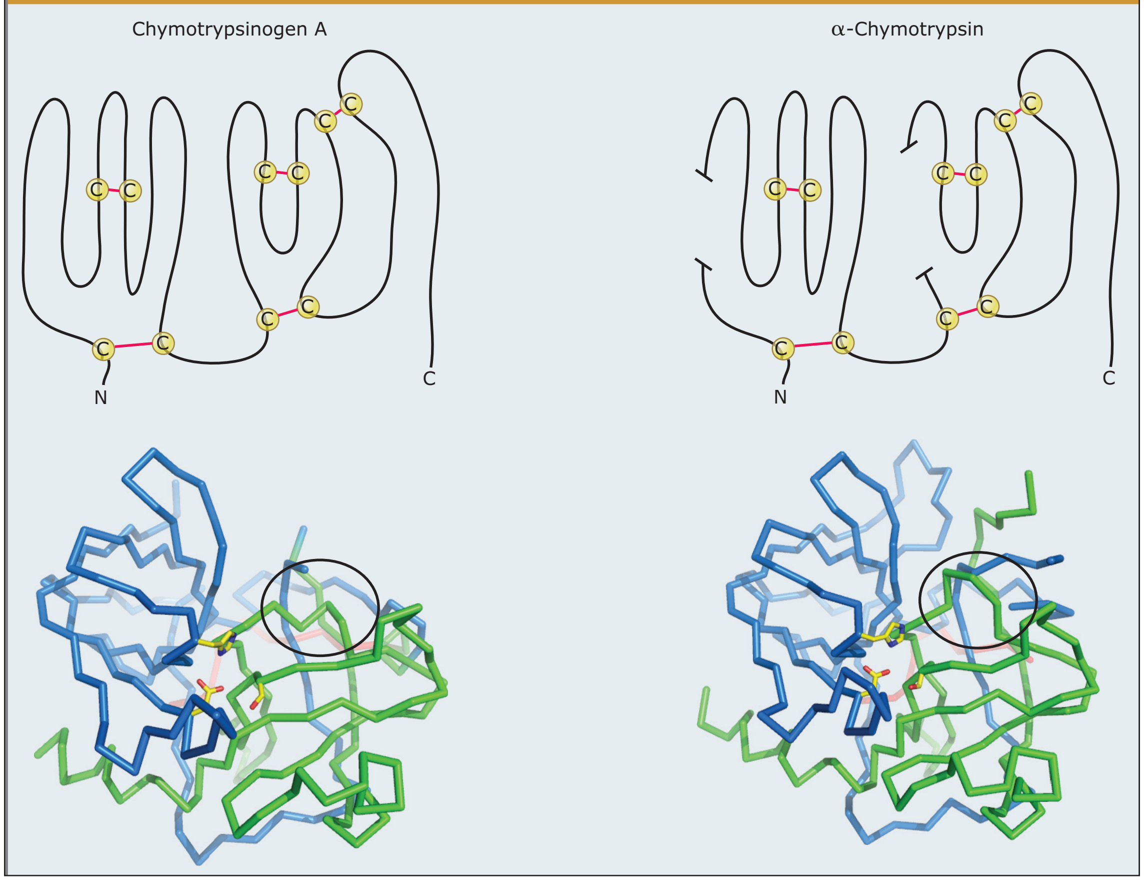

FIGURE 1.57 Proteolytic processing of an inactive zymogen (chymotrypsinogen) to form the active enzyme, α-chymotrypsin. Images generated from Protein Data Bank files 1CHG, 4CHA.

other proteins/domains capable of specifically recognizing phosphorylated serine, threonine, or tyrosine (*Section 1.13, Protein–protein and protein–nucleic acid interactions*). The paradigm for phosphorylation-driven complex formation is undoubtedly the SH2 domain (Src-homology-2), which features prominently in receptor tyrosine kinase signaling pathways through its ability to specifically bind to phosphotyrosine motifs. It is now clear, however, that a diversity of proteins and domains function in all aspects of protein kinase signaling. This is best exemplified by the proliferation of protein modules now known to function as phosphodependent binding domains in serine/threonine kinase signaling pathways. Additionally remarkable is the diversity in architecture seen in these molecules, which ranges from the all-helical 14-3-3 family through to the all-beta Forkhead-associated domains.

Phosphorylation is not unique in its ability to mediate protein–protein interactions and several other modifications, notably acetylation and methylation of the basic amino acids lysine and arginine, and its ability to stimulate interactions with a number of domains such as Tudor, PHD, and Bromo domains, most notably in the context of modification of histone tails in epigenetic regulation of chromatin structure (*Section 1.13, Protein–protein and protein–nucleic acid interactions*). Indeed, the combinatorial effects of specific histone acetylation, methylation, phosphorylation, and ubiquitinylation produce highly specific patterns of modifications, which collectively have become known as the 'histone code.'

Although the covalent attachment of small organic or inorganic molecules to proteins is by far the most commonly observed posttranslational modification, one of the most important regulatory modifications that occurs in eukaryotic cells is the addition of the small protein ubiquitin (FIGURE 1.59). Formation of polyubiquitin chains that are conjugated via a specific lysine (Lys48) flags the target protein for degradation by the 26S proteasome. Ubiquitin modification is achieved in a sequential cascade of reactions catalyzed by ubiquitin-activating enzyme (E1). Modification of ubiquitin-conjugating enzyme (E2) and ubiquitin ligase (E3) results in the formation of an isopeptide bond between the acti-

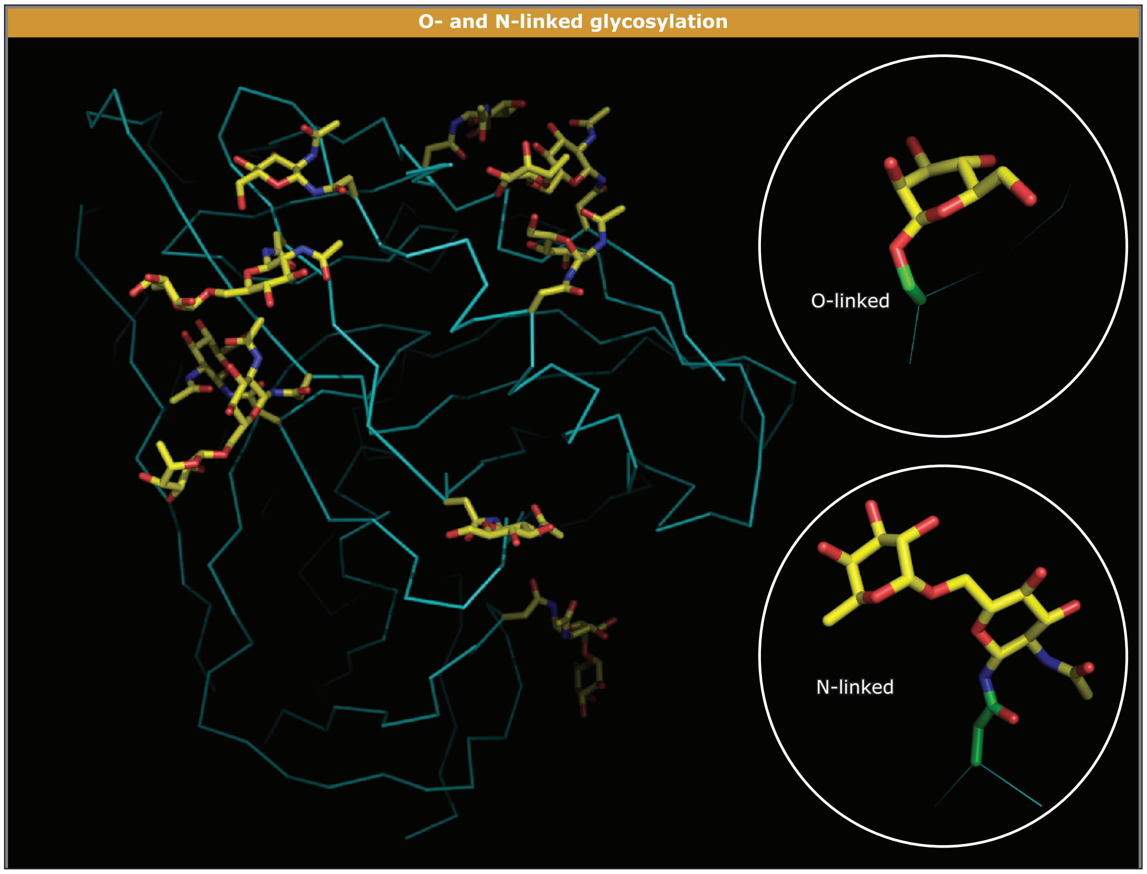

FIGURE 1.58 O- and N-linked glycosylation. Image generated from Protein Data Bank file 1GC1.

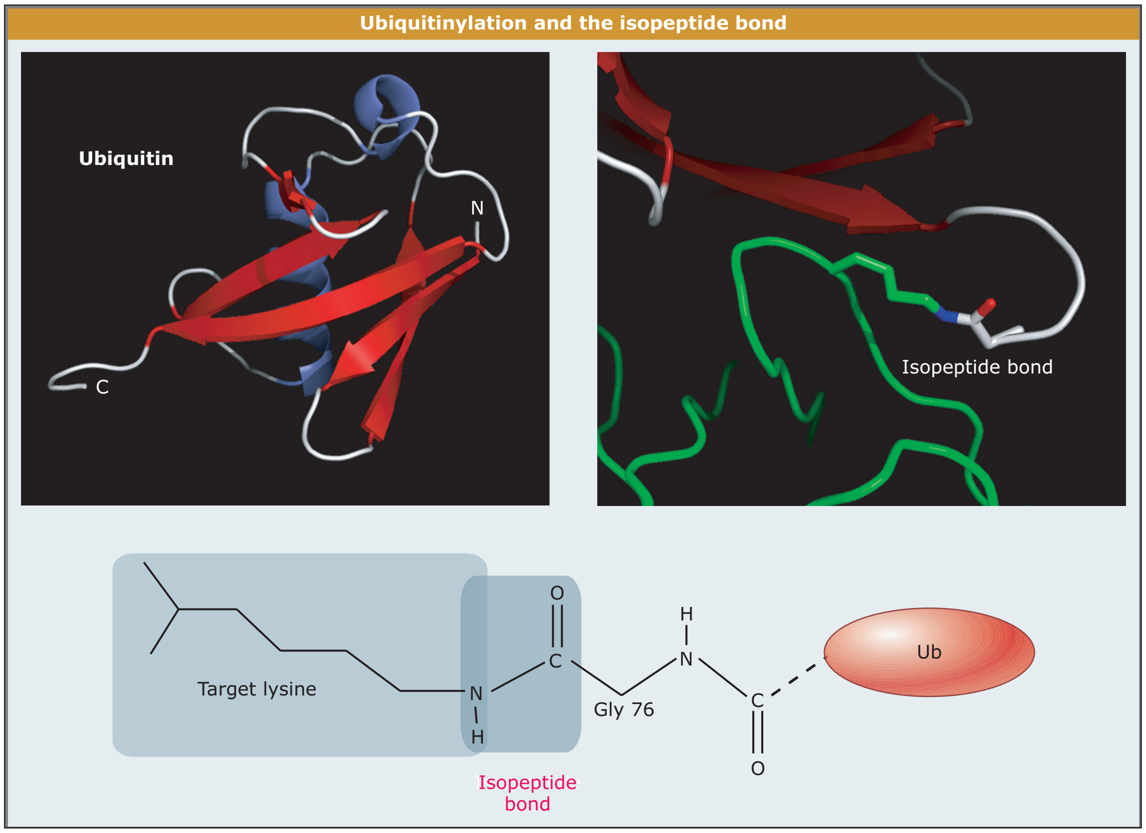

FIGURE 1.59 Ubiquitinylation and the isopeptide bond. Image generated from Protein Data Bank file 1AAR.

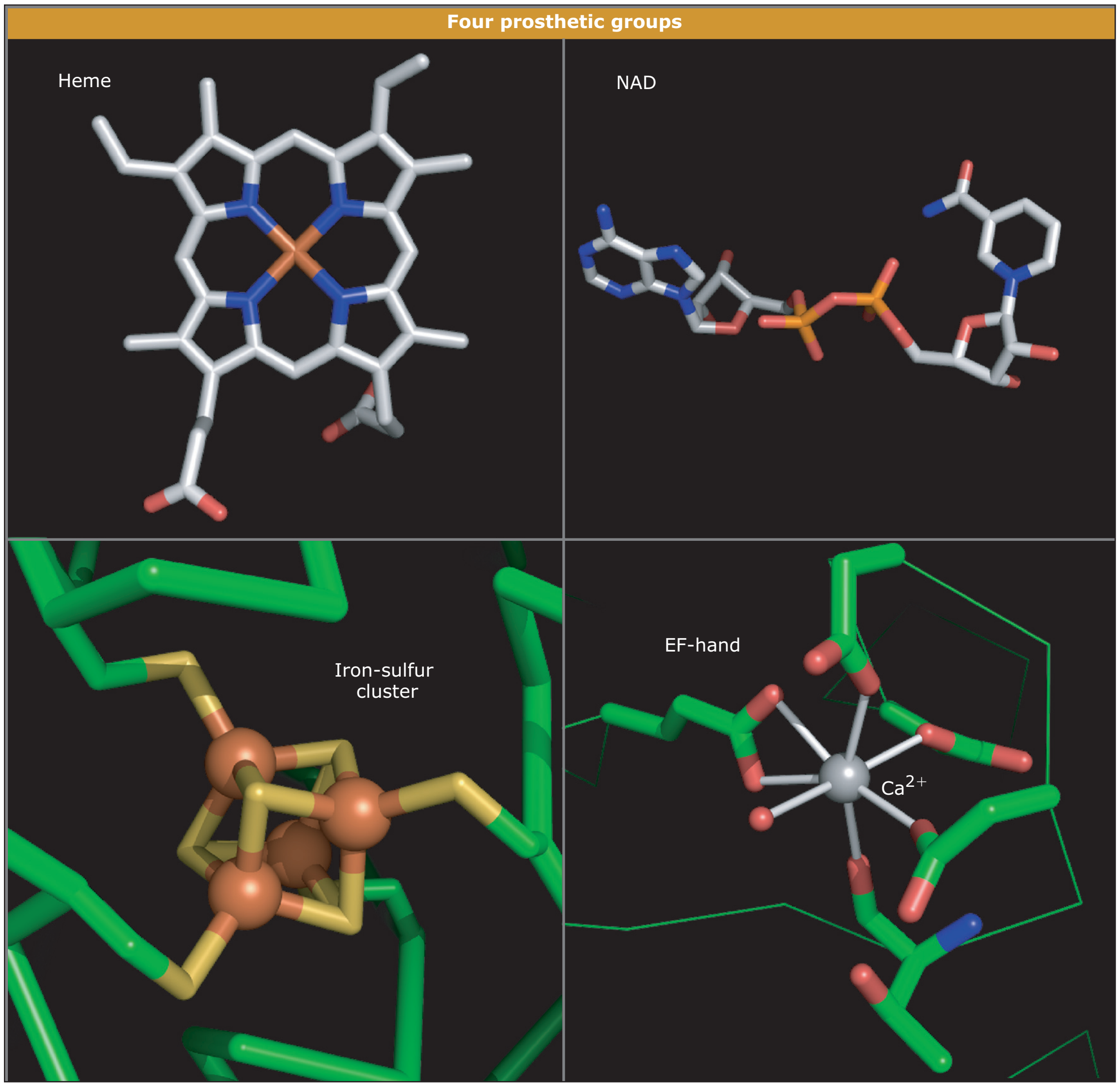

FIGURE 1.60 Examples of four prosthetic groups: heme (iron-protophorphyrin IX), NAD, an iron-sulfur cluster, and a calcium-binding 'EF' hand. Images generated from Protein Data Bank files 1MBN, 1OG3, 1CP2, 1CLL.

vated C-terminal carboxylate group of ubiquitin and the terminal amino group of one or more lysines in the target protein. More recently, it has become apparent that monoubiquitination, or polyubiquitination through Lys63, may play a role in cell signaling. This notion is supported by the fact that deubiquitinating enzymes exist and ensure that this modification is reversible in a manner akin to protein phosphorylation.

Among the most common modifications encountered is the binding of cofactors or **prosthetic groups** that play central roles in biological function (**FIGURE 1.60**). The diversity of prosthetic groups is considerable, ranging from single metal ions to large organic macrocycles. In the case of the hemoglobins and myglobins, the heme prosthetic group itself contains a centrally coordinated iron atom that constitutes the binding site for oxygen. Indeed, the binding of various metal ions is observed to occur in a variety of different contexts, where they may play direct catalytic roles, as in phosphoryl transfer (such as $Mg^{2+}$and Mn^{2+}) and electron transfer processes (metalloporphyrins/iron-sulfur clusters), or may contribute to structural integrity as observed in Zn-finger (Zn^{2+}) and EF-hand motifs (Ca^{2+}). Other common prosthetic groups include NAD/NADH (nicotinamide adenine dinucleotide and its reduced form), NADP/NADPH (nicotinamide adenine dinucleotide phosphate and its reduced form), FAD (flavin adenine dinucleotide), and FMN (flavin mononucleotide), which are present in many enzymes involved in intermediary metabolism and other pathways.

Finally, one of the most remarkable posttranslational modifications occurs in a family of fluorescent proteins well known to modern cell biologists. These are single-domain molecules that form a β-barrel tertiary fold. Green fluorescent protein (GFP) is the archetypal member of this family, and was originally purified from a species of jellyfish, *Aequoria victoria*. Its fluorescent property derives from the formation of a fluorophore within the hydrophobic core, by means of a series of reactions involving a triplet of conserved Ser-Tyr-Gly amino acids that result in rapid cyclization of the main-chain between the serine and glycine residues, along with a slow oxidation of the tyrosine side-chain (**FIGURE 1.61**). The structural stability of GFP has allowed extensive modification of its fluorescent properties by site-directed mutagenesis through alteration of the local environment of the fluorophore and even changing its structure. These mutational variants, along with fluorescent proteins identified and cloned from other organisms, have provided a veritable arsenal of molecules with different excitation and emission spectra. These, in turn, constitute powerful tools for investigation of the subcellular localization and interactions of proteins to which GFP and its multicolored offspring have been fused.

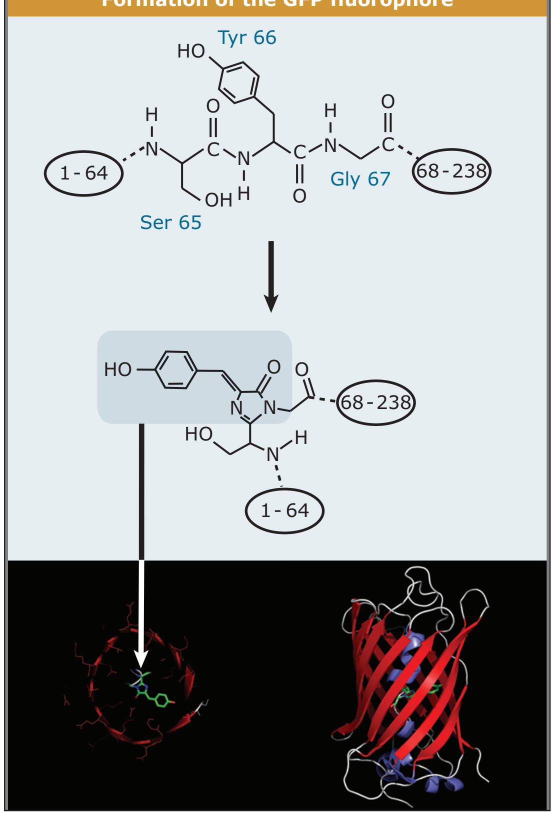

FIGURE 1.61 The GFP fluorophore forms upon folding of the protein itself and remains buried in the hydrophobic core of the β barrel. Image generated from Protein Data Bank file 1GFL.

1.12 Dynamics, flexibility, and conformational changes

Key concepts

- Proteins are often mistakenly construed to be rather static and rigid structures. Structural information derived from NMR, other solution spectroscopy, and even X-ray crystallography, however, has shown that proteins are highly dynamic.
- A wide variety of atomic motions have been observed in proteins, and may occur on a broad range of time scales.
- The dynamics of a protein can be, and most often are, intimately linked to their biological function.
- Conformational changes within a single protein or a multiprotein complex can involve fluctuations in chemical bonds, amino acid side-chain motions, or large movements of domains or entire proteins within a complex.

The small oxygen-storage protein, myoglobin, has been called the 'hydrogen atom' of molecular biology, reflecting its prominence in some of the major developments over the last 50 years. Mammalian myoglobins are extremely highly conserved proteins, containing 153 amino acids together with a macrocyclic iron-binding cofactor, heme. Myoglobin's importance in our understanding of the role of protein dynamics in biological function arises from the early observation that the site of oxygen binding on the distal side of the heme (**FIGURE 1.62**) is inaccessible to bulk solvent. Nevertheless, oxygen binding to myoglobin in solution occurs with association kinetics only marginally more slowly than the diffusion limit. This implies that significant conformational displacements of atoms from their positions observed in crystal structures of myoglobin must take place in order that even small, diatomic ligands such as oxygen can bind to the heme iron. Through a variety of computational, biochemical, biophysical, and structural studies on myoglobin and a host of other systems, we now view proteins as existing not as a single rigid structure, but rather as an ensemble of rapidly interconverting conformations. In this way, the motions experienced by atoms within myoglobin are sufficient to open up channels in the structure, allowing access of oxygen and carbon monoxide to the protein interior.

The structural motions that occur in proteins can be crudely classified as 'local' or 'global.' Local changes involve, for the most part, extremely small movements resulting from thermal fluctuations in covalent bonds that take

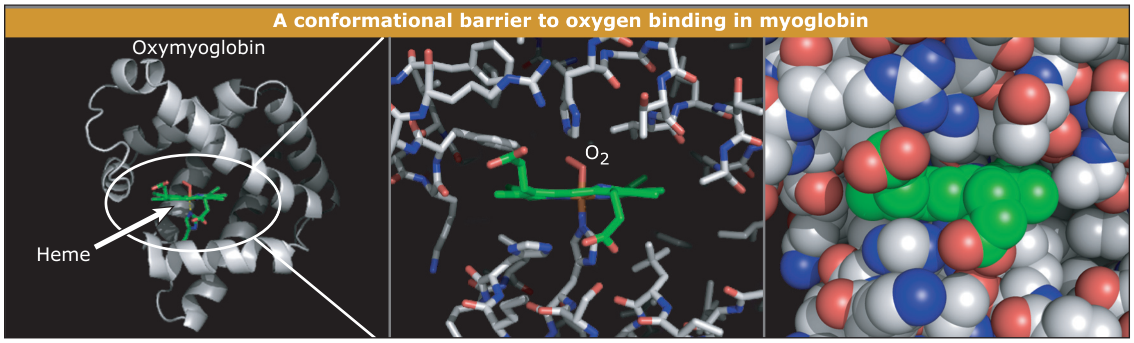

FIGURE 1.62 A conformational barrier to oxygen binding in myoglobin. Image generated from Protein Data Bank file 1MBO.

place on the femto (10^{-15}) to pico-second (10^{-12}) time scale and reflect positional displacements of less than 1 Å. Conformational changes in amino acid side-chains (such as flipping of the aromatic rings of phenylalanine and tyrosine) or the aromatic rings of phenylalanine and tyrosine occur on the millisecond time scale, whereas more extensive 'global' motions of secondary structural elements or whole domains can take place on time scales ranging from pico- to milliseconds, and sometimes even longer.

The idea that conformational change and flexibility are important in protein function has been around for many years. The effects of mutational disruption in hemoglobin as a cause of sickle-cell anemia were recognized as a conformational defect long before hemoglobin's structure was finally determined. Indeed, we now recognize that mutation is a major cause of structural change in proteins, which we consider in more detail later (*Section 1.15, Structure and medicine*).

Some of the major insights into the biological significance of conformational changes have been from studies of enzyme catalysis. In 1958, Koshland suggested the notion of **induced-fit** in order to explain both the high specificity and the catalytic activity of enzymes. The model proposed that binding of the correct substrate to an active site would be accompanied by conformational changes in that active site (or even the substrate itself). In this way, non- or pseudo-substrate binding would not induce an enzyme conformation that was catalytically proficient, because it would not be reconfigured to bind most tightly to the transition state of the reaction. Many examples of induced fit have now been observed structurally, the first being the structures of hexokinase when free and bound to its substrate glucose, which show large changes in the relative orientation of two domains as the enzyme closes around the substrate (**FIGURE 1.63**).

We have already addressed the concept of **allosteric** conformational changes in the context of the oxygen carrier hemoglobin. Here binding of oxygen to the ferrous heme-iron atom of one subunit results in a shortening of the coordination bond that connects the heme group to the so-called proximal histidine residue. This change in bond length exerts a pull on the F-helix within which the proximal histidine resides, and the ensuing conformational change is transmitted through the breaking of salt bridges to the other subunits, raising their oxygen affinity. As such, hemoglobin is a model of allostery, where binding of a ligand or 'allosteric effector' causes a conformational change that either positively or negatively affects interactions at a second site. The effect of allostery on the affinity of successive binding sites is known as **cooperativity**.

Allostery is one of the most common regulatory mechanisms found in biological systems and is facilitated by quaternary association. The concept of allostery was originated by Monod, Wyman, and Changeux in a classic paper published in the *Journal of Molecular Biology*. The 'MWC' or 'concerted' model considers allostery as acting on only two symmetric and rigid quaternary structures called the 'tense' or 'T'-state, which has low ligand affinity, and the 'relaxed' or 'R'-state, which has higher affinity. These two states are in equilibrium, and binding of a ligand/effector molecule to successive subunits 'pulls' the conformational equilibrium toward the high-affinity R-state. Alternatively, Koshland and coworkers described a 'sequential' model, which allows for binding of a ligand to one subunit to directly influence the binding site conformation of other subunits in the oligomer. In fact, both models may be necessary to explain the overall behavior of hemoglobin in response to pH (the 'Bohr effect'), binding of diatomic ligands (oxygen, carbon monoxide), and al-

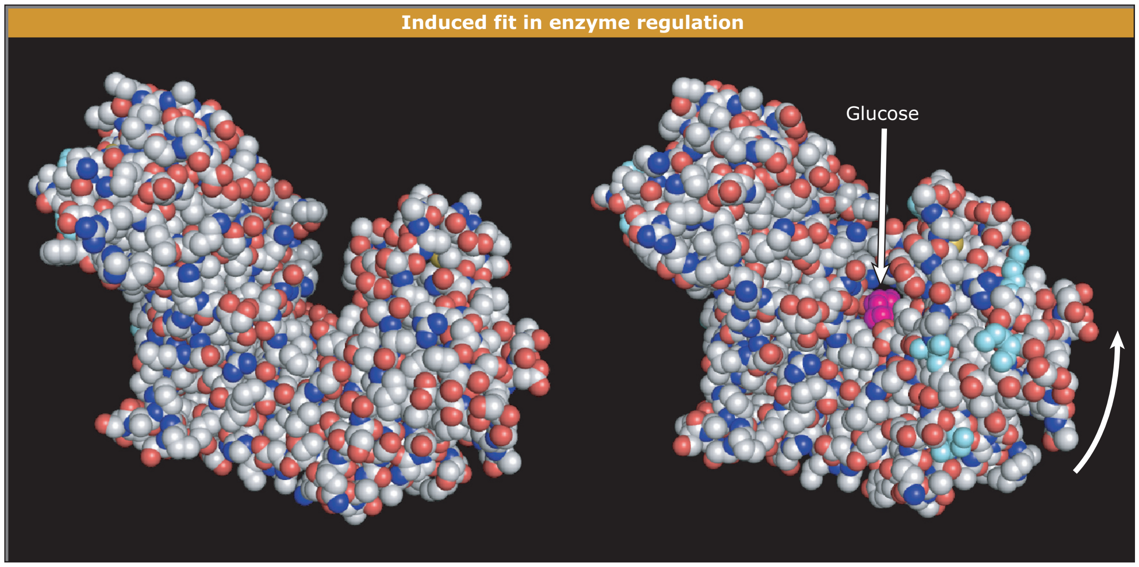

FIGURE 1.63 Induced fit in enzyme regulation. Images generated from Protein Data Bank files 2YHX, 1HKG.

losteric effectors such as diphosphoglycerate.

Allosteric effectors need not always be small molecules. They can be whole proteins, and there are many instances of conformational changes that result from binding of regulatory subunits to activate or even inhibit activity of the resulting complex. As an example, we will consider the cyclin-dependent kinases (CDKs) that function as master regulators of cell-cycle progression in eukaryotes. As such, their kinase activity must be highly controlled, and this occurs at a number of levels. Each CDK associates with different but specific activating subunits called cyclins at precise times in the cell cycle. The major effect of cyclin binding is a conformational change in the kinase subunit that pushes an α-helix containing a catalytically important glutamate residue into the active site (**FIGURE 1.64**). This is not the whole story, however, because additional events are required for full activation. First, an inhibitory tyrosine phosphorylation must be removed by protein phosphatase. Second, an activating phosphorylation on a specific threonine residue located in the 'activation' or 'T'-loop is required for it to adopt an ordered conformation required for substrate binding and catalysis. Indeed, the central role of phosphorylation-dependent conformational changes in CDK activation, and kinase activation in general, exemplifies how posttranslational chemical modifications can result in biologically important structural changes.

Possibly the most common driver of conformational changes in proteins is the hydrolysis of nucleoside triphosphates (NTPs), particularly ATP. NTPs are remarkable molecules because they are relatively stable in isolation in spite of the fact that the phosphate–phosphate bonds are referred to as 'high energy': hydrolysis of the bond between the β and γ phosphates yields around 12 kcal mol^{-1} of free energy. This occurs efficiently only when hydrolysis of ATP is catalyzed by enzymes called ATPases, an essential characteristic given that spontaneous hydrolysis by water would otherwise render ATP too unstable to be useful!

In earlier sections we have seen how ATP hydrolysis during phosphorylation by protein kinases can result in structural changes that arise from rearrangement of protein segments containing the phosphorylated residues, or through the interaction of phosphospecific binding proteins with phosphorylated regions. In *Section 1.15, Structure and medicine*, we will also see how small structural changes in small GTPases are driven by GTP binding and hydrolysis. ATP hydrolysis can, however, be coupled to large conformational changes, a phenomenon perhaps best exemplified by so-called motor proteins such as myosin, kinesin, and dynein. These molecules most generally produce mechanical movement through interaction with fibrous cellular substructures such as actin filaments and microtubules formed by polymerization of actin and tubulin, respectively.

The most complete structural picture of how the chemical energy of ATP hydrolysis is transformed into mechanical force has emerged from studies of a proteolytic fragment of myosin called S1, which contains the globular 'head' or 'motor' domain that binds to actin filaments

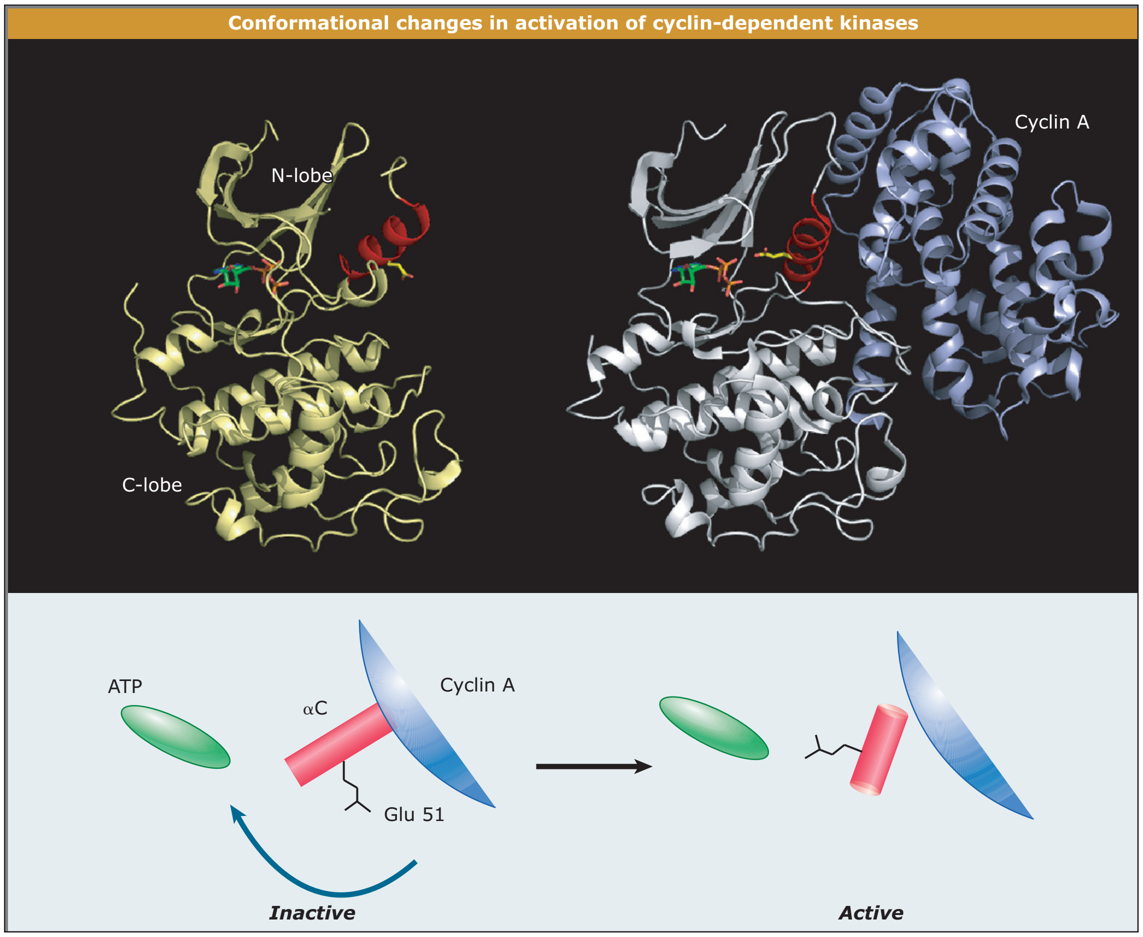

FIGURE 1.64 Protein–protein and phosphodependent conformational changes in the activation of cyclin-dependent kinases. Images generated from Protein Data Bank files 1HCL, 1QMZ.

and ATP itself, and a regulatory domain, which is often referred to as the lever arm. In full-length myosins, a third 'tail' domain is present that is responsible for interacting with other cellular proteins. Movement is produced by changes in affinity of the regulatory domain for the actin filament, which is, in turn, coupled to a cycle of ATP binding, hydrolysis to ADP plus inorganic phosphate (P_i), and finally, ADP and phosphate release (FIGURE 1.65). These changes in affinity result in the successive binding and dissociation of the myosin/actin complex during the cycle. The small conformational changes that occur during the ATP cycle are then transduced to and magnified by the lever arm, an extended coiled-coil structure that is stabilized through calcium-binding EF-hand proteins known as light-chains. The lever arms of different classes of myosins (of which some 20 are currently known) may differ substantially in length. This provides for different lengths of powerstroke that are adapted for specific biological functions.

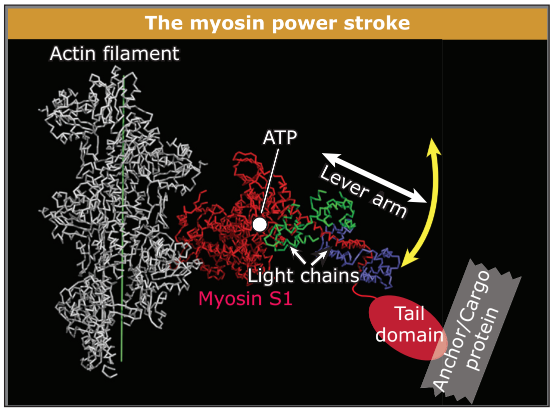

FIGURE 1.65 The myosin powerstroke.

1.13 Protein–protein and protein–nucleic acid interactions

Key concepts

- The interactions that proteins make define their biological function and are dictated by their structure, dynamics, and physico-chemical properties.
- Protein interfaces are formed by combinations of polar, nonpolar, and electrostatic interactions whose relative contributions define binding specificity and affinity.

The biological functions of proteins are exercised through the interactions that they make with other molecules, and the diversity of these interactions can be seen in specific contexts elsewhere in this chapter. This section will address the molecular basis of how protein interactions occur, the nature of the interfaces that have been observed, and the ways that these characteristics regulate binding affinity, stability, and specificity. Although the discussion will focus on protein–protein and protein–nucleic acid complexes, the principles described are equally applicable to interactions of proteins with other ligands, such as sugars, lipids, cofactors, and substrates.

Reversible binding of two molecules to form a bimolecular complex is most often described by the affinity of the interaction. The term **affinity** refers to the equilibrium constant that is defined by the relative concentrations of the bound and free species in a mixture at equilibrium. In thermodynamic terms, the equilibrium constant is related to the Gibbs free energy (ΔG) of binding (FIGURE 1.66), and this must always be negative for a spontaneous interaction. The association equilibrium constant, K_a, has units *mol*$^{-1}$, so we will refer to affinity in terms of dissociation constant K_d, which has the more intuitive units of mol. Thus tighter binding complexes have lower K_d values such that an affinity of 1 mM is weaker than 1 μM, and so on. The free energy, ΔG, can also be described as a sum of enthalpic (ΔH) and entropic (ΔS) terms. The relative enthalpic and entropic components of protein interactions can be measured in several ways, but the structural basis for the contributions of each of the two components is often difficult or impossible to delineate. Nonetheless, some general trends are discernable and the ability to predict ab initio how proteins form complexes is, along with the protein-folding problem, a major goal of computational and experimental biochemists.

The structure and physico-chemical properties of protein–ligand interfaces are necessarily dictated by the distribution and conformations of the amino acids at that interface and, importantly, by their interactions with surrounding solvent. For this reason, interfaces can be roughly described in terms of the relative contributions of hydrophobic interactions, hydrogen bonding, and electrostatic effects. Quantitatively, the overall size of the interface is often reported in terms of the solvent-accessible surface that is buried or rendered inaccessible to solvent upon complex formation (FIGURE 1.67). Surfaces actually observed in crystal structures show buried surface areas in the range of around 800 Å^2 at the lower end up to or exceeding 5000 Å^2, with a mean value around 1500 Å^2. A more detailed analysis of the binding surface can supply information about the relative extents of interfacial contacts mediated by polar and nonpolar atoms, with nonpolar contacts typically contributing around 60% of the interface in high-affinity, stable complexes.

In terms of biological activity, it is convenient to classify protein–protein or protein–ligand complexes as either *stable* (long-lived) or *transient* (short-lived). For example, the four subunits of hemoglobin are extremely stable, which makes biological sense given that the individual proteins could not fulfill their biological function. Conversely, many functionally significant complexes may only form for a short but precise time period, allowing rapid responses to changes in the cellular environment. Posttranslational modifications can 'switch' a weak and transient interaction into a tight and stable one. In the absence of such modifications, though, interfaces formed between weak, tran-

Binding, equilibria, and free energy

For binding of two molecules A and B at equilibrium

$$A + B \rightleftharpoons AB$$

$$K_a = \frac{[AB]}{[A][B]}\ M^{-1}$$

$$K_d = \frac{1}{K_a} = \frac{[A][B]}{[AB]}\ M$$

where [ ] denotes concentration

The association equilibrium constant K_a is related to the Gibbs free energy ΔG by

$$\Delta G = -RT\ \ln K_a$$

where T is the temperature (in K) and R is the gas constant

ΔG can be expressed as the sum of enthalpic and entropic components:

$$\Delta G = \Delta H - T\Delta S$$

FIGURE 1.66 Binding, equilibria, and free energy.

siently interacting proteins tend to be small and somewhat less hydrophobic in nature.

Although the numbers of protein–protein and protein–ligand complexes represented in the protein databank continues to increase, it remains difficult or even impossible to accurately predict the affinities of interactions from structure alone. This is because interaction energies that occur between large and complex interfaces consist of enthalpic contributions from van der Waals forces, hydrogen bonding, and electrostatic interactions, along with entropic effects derived from the formation of nonpolar interactions that displace surface water molecules (and therefore increase entropy). Often, a loss of entropy arises from a reduced conformational flexibility upon complex formation. Broadly speaking, the overall size of the interface is loosely correlated with affinity, which has the inevitable consequence that for the interaction of two globular proteins, an increase in affinity can only be achieved by an increase in the overall size of one or both partners. As we will see in *Section 1.14, Function without structure?*, some proteins have circumvented this evolutionary 'limitation' by utilizing tracts of unstructured sequence to maximize interaction surfaces.

Specificity is one of the most critical features of any interaction and can be usefully defined as the relative affinities of a protein for different binding partners. For this reason, high affinities are not necessarily an indicator of high specificity, and vice versa. As a general rule, it is important to realize that affinities and specificities of interactions in biological systems are optimized rather than maximized. Nonpolar interactions are thought to contribute most to overall affinity, whereas specificity is mostly derived from shape complementarity (which may have a nonpolar component) and hydrogen bonding. In this respect, the pioneering work of Alan Fersht in the early 1980s is particularly significant. Using the enzyme tyrosyl tRNA synthetase as a model system, Fersht and coworkers employed the newly developed technique of site-directed mutagenesis to introduce individual amino acid substitutions into the protein and examine the effects on catalysis and specificity. In particular, these studies showed that single uncharged hydrogen bonds contribute relatively little to binding and specificity, whereas those involving a charged donor or acceptor are much more significant.

As mentioned in the preceding discussion of nonpolar interactions, water molecules play an important, albeit indirect, role in the formation of protein-binding interfaces. Ordered water molecules are commonplace in protein–protein and protein–ligand interfaces, where they may mediate linking hydrogen-bonding contacts between side-chains across the binding surface (FIGURE 1.68), or fill 'holes' and thus increase surface complementarity. How and if they contribute to interaction specificity has been a controversial question that initially gained prominence in the context of protein–nucleic acid interactions. Most structures of complexes solved at high resolution contain interfacial waters, many of which make their full complement of four hydrogen bonds (two H-bond donors and two acceptors). These water molecules clearly play an important structural role, and in some cases appear to have been conserved through evolution.

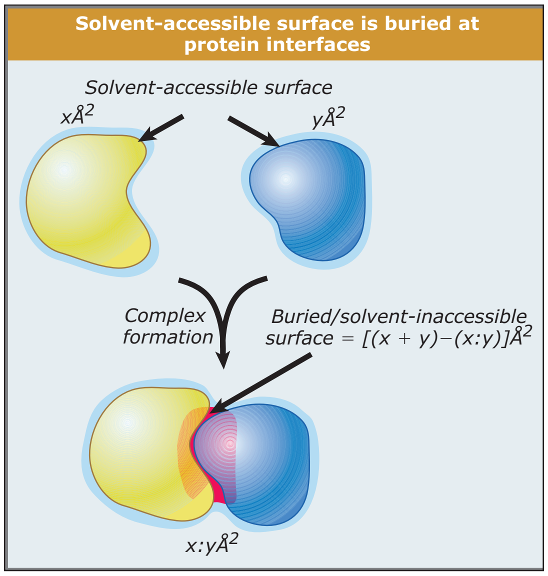

FIGURE 1.67 Solvent-accessible surface is buried at protein interfaces.

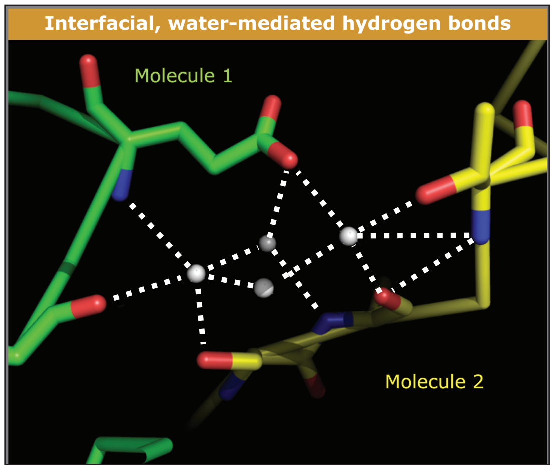

FIGURE 1.68 Interfacial, water-mediated hydrogen bonds.

Nucleic acids constitute a major class of binding partners for proteins, and these interactions are central to the regulation of the processes of transcription, translation, and DNA replication. The first protein–DNA complexes characterized by X-ray crystallography were those of prokaryotic transcriptional regulators, and these structures showed, for the first time, how proteins are able to exploit the characteristics of B-form DNA in generating specific and high-affinity interactions. The structure of double-stranded DNA, as suggested by Watson and Crick in the 1950s, has become something of an icon. To recap, the two strands of connected nucleotide bases intertwine in an antiparallel arrangement to form a right-handed double-helical structure with the phosphate groups of each nucleobase located on the periphery of the double-helix, linking to the 3′ ribose hydroxyl group of the next through a phosphodiester bond. The strands are held together through specific patterns of hydrogen bonds (the familiar base-pairing interactions) of the nucleobases (adenine with thymine, cytosine with guanine) at the center.

From the point of view of a DNA-binding protein, the major structural features of classical B-form double-stranded DNA (dsDNA) are two 'grooves' that differ in width and depth (FIGURE 1.69). The minor groove is rather narrow, but the major groove is much wider, allowing access of binding proteins to nonpolar and hydrogen-bonding groups on the edges of the base-pairs. In fact, the width of the major groove is ideally suited to accommodate a single α-helix, and many X-ray and NMR structures have shown how helices in the context of helix-turn-helix (HTH), helix-loop-helix (HLH), basic leucine zipper (bZIP), zinc-finger, and other motifs interact with the major groove (FIGURE 1.70). Major groove recognition, however, is not limited to α-helices, and the structure of another bacterial repressor, MetJ, showed that a pair of antiparallel β strands can function in a very similar and equally effective way (Figure 1.70). Indeed, it has subsequently

Structure of double-stranded DNA

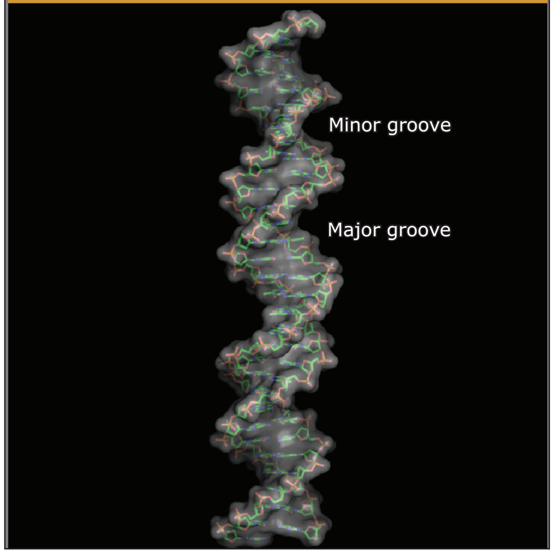

FIGURE 1.69 Structure of double-stranded DNA.

Interactions with the major groove of DNA by α-helices and β strands

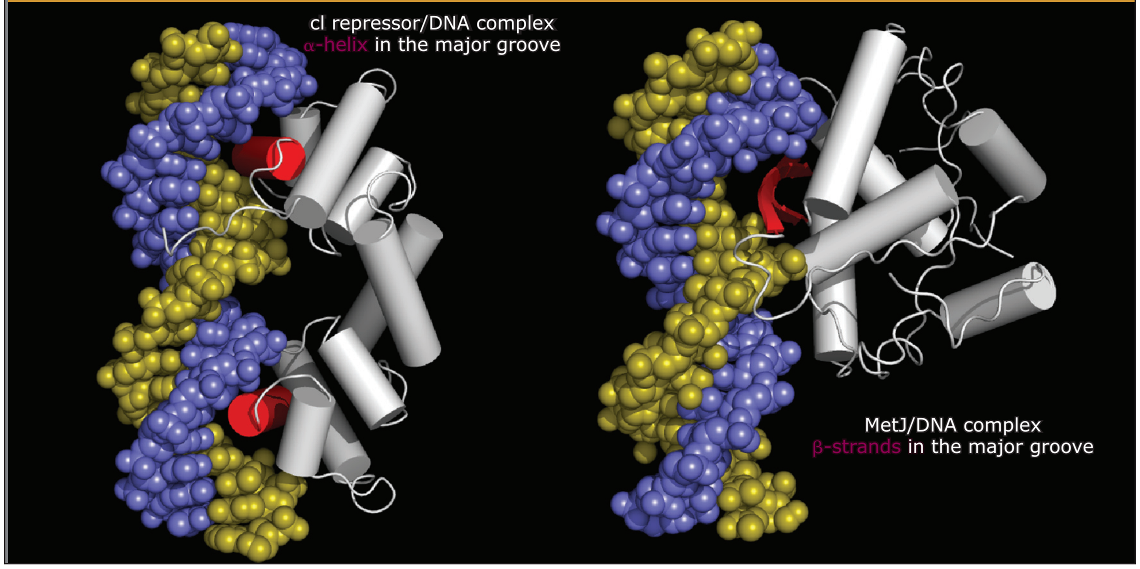

FIGURE 1.70 Interactions with the major groove of DNA by α-helices and β strands. Images generated from Protein Data Bank files 1LMB, 1MJ2.

become clear that combinations of α, β, and extended structure may be used in combination in protein–DNA recognition to provide specificity through interaction with the base edges, along with increased affinity through largely nonspecific electrostatic interactions with the negatively charged phosphate backbone. It has also emerged that proteins do not only bind to linear DNA duplexes, and major distortions in bound DNA have been observed that range from significant bends in the DNA helical axis to dramatic deformations in the phosphodiester backbone, allowing access to the minor groove itself.

In comparison to DNA, RNA presents a much more varied and complex spectrum of problems in molecular recognition by proteins. This arises from the fact that RNA is produced as a single-stranded molecule by transcription from a DNA template. As such, RNA can and does form a bewildering array of secondary and even tertiary structures that are, nevertheless, technically difficult to investigate by either X-ray or NMR methods. Double-stranded RNA regions form a so-called A-form structure in which the minor groove is wider than in B-form DNA, whereas the major grove is deeper but narrower, presenting different structural features to potential binding proteins (FIGURE 1.71).

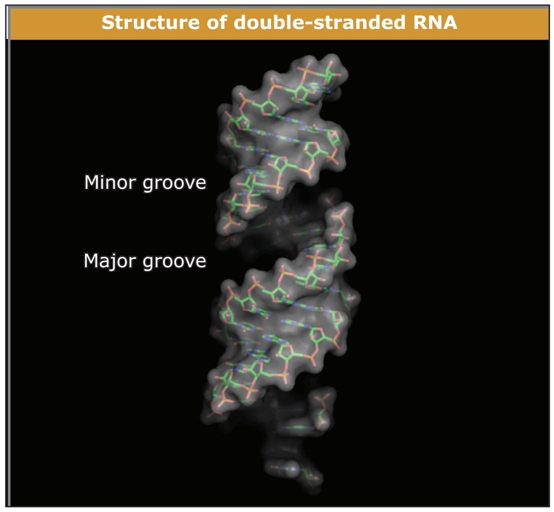

FIGURE 1.71 Structure of double-stranded RNA.

The diversity in RNA structure appears to be matched by the diversity of ways in which proteins interact with it. Single-stranded regions in RNA may allow direct 'reading' of the base sequence through interactions with the base-pairing hydrogen-bond donor/acceptors that are otherwise inaccessible in double-stranded RNA or DNA. In addition, proteins may recognize three-dimensional surfaces created by tertiary interactions within any given RNA molecule. Examples of many or all of these strategies are available (FIGURE 1.72), but the generalizations that enable us to broadly classify protein–DNA interaction mechanisms are less obvious for protein–RNA binding systems. Suffice it to say that the high-resolution structures of the 50S and 30S subunits of prokaryotic ribosomes revealed a host of new RNA structural motifs within the 23S, 16S, and 5S RNAs. At the same time, these remarkable structures have revealed a host of novel interactions of the ribosomal RNAs with the complement of

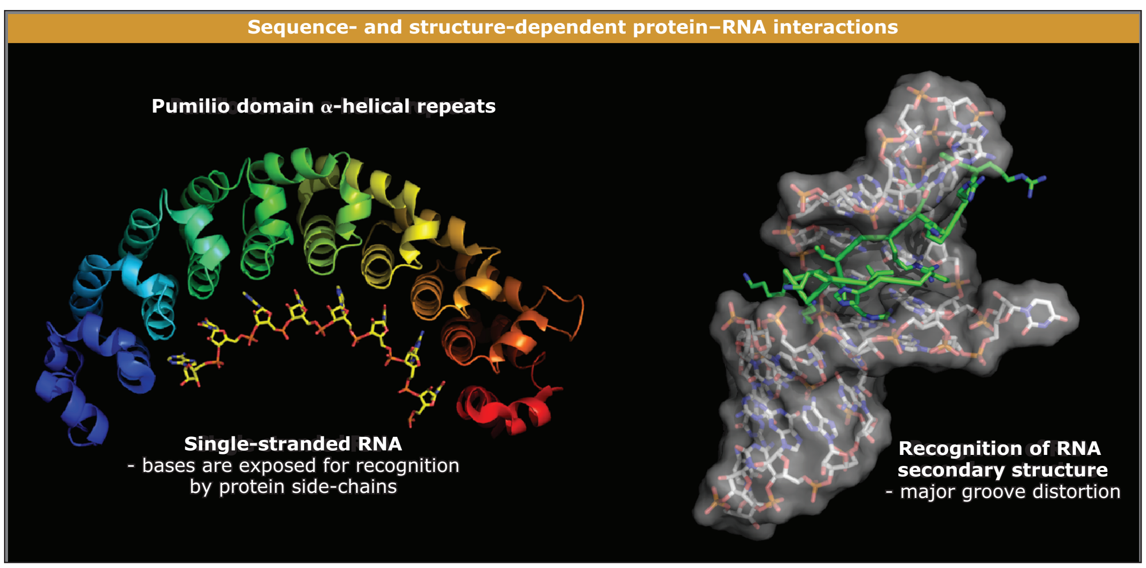

FIGURE 1.72 Sequence- and structure-dependent protein–RNA interactions. Images generated from Protein Data Bank files 1M8Y, 1ZBN.

RNA structure and RNA–protein interactions in the bacterial ribosome

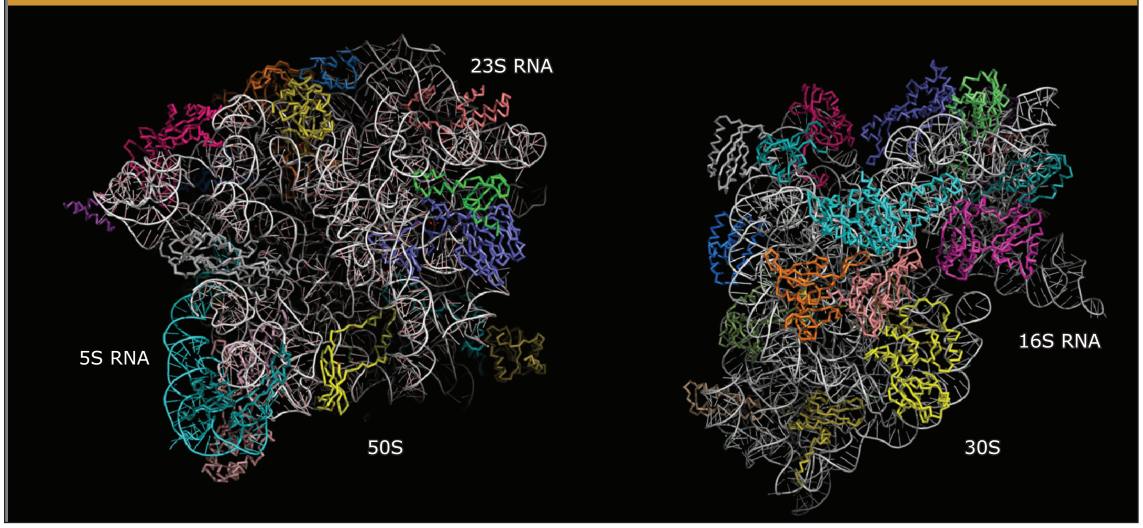

FIGURE 1.73 RNA structure and RNA–protein interactions in the bacterial ribosome. Images generated from Protein Data Bank files 1KC8, 1HNX.

ribosomal proteins (**FIGURE 1.73**). Nevertheless, even including the information from the structures of ribosomes and their components, protein–DNA complexes dominate the complement of nucleic acid structures available at present and there is much more to be learned about RNA structure and the ways in which proteins and RNA interact.

1.14 Function without structure?

Key concepts

- Many protein functions may be carried out by, or depend on, unstructured regions of amino acids.
- In the formation of complexes involving unstructured regions, complexes may remain completely or partially unfolded, or may adopt secondary or tertiary structures upon binding.

To this point we have seen many examples of how the three-dimensional structure of a variety of proteins and protein complexes is exquisitely related to function. It has, however, become increasingly apparent that some proteins with clearly defined and important biological activities appear to lack any obvious tertiary or even secondary structure. These have become known as ‘natively unfolded’ or ‘intrinsically unstructured’ proteins. These molecules are not as uncommon as one might expect. Although only a few examples have been found or predicted in prokaryotes, it appears that upward of 30% of proteins encoded in eukaryotic genomes may fall into this class. Although some proteins may consist entirely of unstructured regions, a hierarchy of organization exists, with some proteins containing some secondary structure and others containing unstructured regions within which globular domains may be embedded.

Unfolded proteins or regions within proteins are generally characterized by amino acid sequences that are of ‘low complexity.’ Such sequences may contain extended tracts of polar residues in many combinations, and such primary structures are not able to adopt a globular fold due to the absence of nonpolar groups that could form a hydrophobic core. Given the remarkable functional characteristics that are bestowed upon proteins by virtue of the secondary, tertiary, and quaternary structures, the advantages of unfolded conformations might appear to be somewhat obscure. Indeed, there would appear to be some obvious disadvantages, most notably sensitivity to proteolytic degradation. Several useful features of disordered regions are, however, discernable, and we will consider them in the context of a few of the structurally and biochemically characterized biological systems in which the function of natively unfolded proteins has been investigated.

One of the first structures of a complex of a natively unfolded protein (NUP) with a binding partner to be described was that of the the cyclin/CDK inhibitor p27 and its cognate kinase cyclinA/CDK2. In fact, this is one of a number of examples of the activity of NUPs in the general area of cell-cycle regulation that seems to be something of a focus for this class of molecules. The X-ray structure shows clearly how p27 wraps around the cyclin/CDK complex and inhibits the enzyme by rearranging the position of the kinase N-lobe and intruding into the ATP binding site (**FIGURE 1.74**). In forming the complex, a large solvent-accessible surface is buried by fewer than 70 residues of p27. This exemplifies the fact the extent of binding interfaces generated by NUPs is much greater than would be possible for a globular protein of the same number of amino acids.

The general importance of posttranslational modifications has been discussed (*Section 1.11, Posttranslational modifications and cofactors*) and the extended, unstructured nature of natively unfolded regions would be expected to facilitate access when modifying enzymes to target residues. This notion correlates well with the observation that NUPs are prevalent in cell-cycle regulatory mechanisms where a good deal of posttranslational modification, particularly phosphorylation, is employed. In addition to cell-cycle proteins, natively unfolded regions are characteristic of many proteins involved in nucleic acid recognition and transcriptional activation, many of which may also be subject to phosphorylation or other regulatory mechanisms. In terms of DNA binding, a globular domain may mediate sequence-specific recognition whereas an extended, basic 'tail' region binds nonspecifically to the phosphate backbone, contributing to the overall interaction affinity. Similarly, transcriptional activation regions such as the classical 'acid blob' segment of herpes simplex virus VP16 execute their biological function in a natively unfolded form. In other transcriptional activator proteins, however, natively unfolded regions may spontaneously fold into globular structures upon interaction with partner proteins or activation targets. For example, the kinase-inducible transcriptional activation domain (KID) of the cyclic AMP-response element binding protein (CREB) binds to the KIX

p27-inhibition of cyclin-CDK

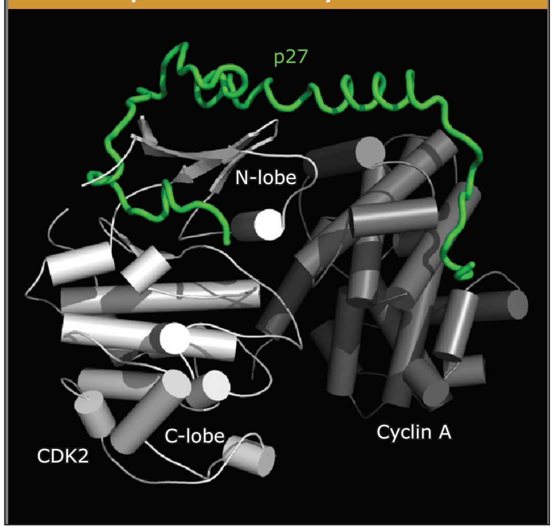

FIGURE 1.74 p27-inhibition of cyclin-CDK. Image generated from Protein Data Bank file 1JSU.

Phosphorylation-dependent folding-upon-binding in KIX/KID interactions

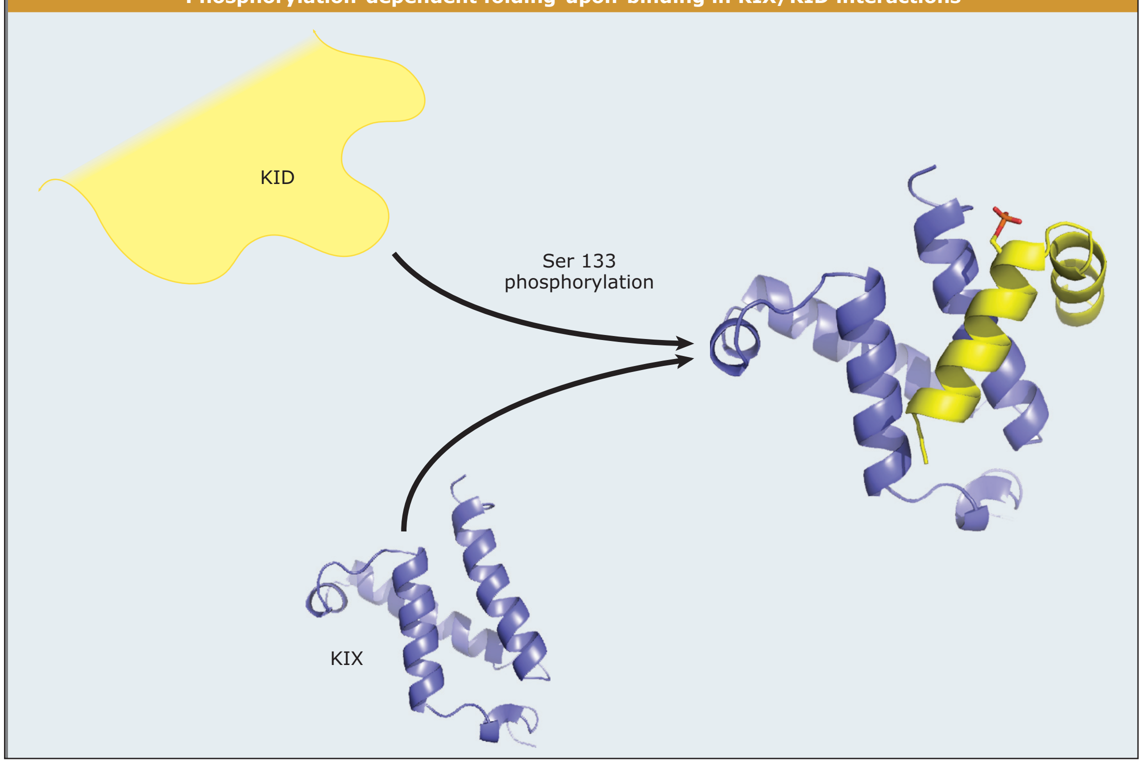

FIGURE 1.75 Phosphorylation-dependent folding-upon-binding in KIX/KID interactions. Image generated from Protein Data Bank file 1KDX.

domain of CREB-binding protein (CBP) upon phosphorylation of a specific serine residue, Ser 133. Both the phosphorylated and unphosphorylated forms of KID are natively unfolded in solution but fold into a roughly helix-turn-helix motif upon binding to KIX (FIGURE 1.75). As the name implies, KID refolding and binding to CBP is absolutely dependent on phosphorylation, and the extensive hydrogen-bonding interactions between phospho-Ser 133 and KIX residues in the complex presumably provide a favorable enthalpic contribution that allows formation of the folded structure. Similarly, folding-upon-binding has been observed to occur in a number of DNA-binding proteins following interaction with their DNA targets, and this effect can directly contribute to the specificity of the interaction.

1.15 Structure and medicine

Key concepts

- The accumulated knowledge of the detailed three-dimensional atomic structures of many thousands of proteins and their complexes has provided unparalleled insights into a great many aspects of the biological functions of proteins.
- Protein structure analysis has opened up new vistas of opportunity for understanding the molecular basis of disease and the design and development of new therapeutic approaches.

Mutation as a cause of disease has long been a focus for molecular and structural biology. Indeed, the existence of naturally occurring mutations, and the ability to induce mutations chemically, underpinned much of the early research into the nature of the gene and the genetic code itself. As we have seen, the precise sequence of amino acids in any given protein has evolved over millions of years to provide a precise architecture tailored to biological function. Clearly, not all mutations are necessarily deleterious or the process of natural selection could not work, and we now know of thousands of sequence polymorphisms within the human genome that are functionally (or phenotypically) silent. Equally, however, we also know of many point mutations, frameshifts, deletions, insertions, and genetic rearrangements that have devastating medical consequences.

The concept that disease may arise as a result of genetic aberration has a long history and stems from the observations of Archibald Garrod in the late nineteenth century, who coined the term 'inborn errors of metabolism' to describe a variety of congenital metabolic diseases such as phenylketonuria. Garrod insightfully attributed these to defects in enzymes long before it was known that enzymes were proteins! In the 1950s, Max Perutz's studies on sickle-cell hemoglobin were the first to apply structural methods—in this case, X-ray crystallography—to attack a human disease at the molecular level, showing how a single mutation of glutamate to valine in the β subunits of adult human hemoglobin causes pronounced changes in structure and solubility of the affected $\alpha_2\beta_2$ Hb tetramer, causing it to aggregate as fibrillar structures in red blood cells and induce the sickle shape long known to microscopists.

Sickle-cell anemia is one example of how the disruption of structural integrity can lead to disease. Among these are the 'diseases of aggregation' such as Alzheimer's, Huntington's, and prion-related disorders that are often associated with neurodegeneration. Huntington's disease is representative of a class of disorders caused by expansion of cytosine-adenine-guanine nucleotide triplets that encode the amino acid glutamine. The number of consecutive glutamines within the expanded glutamine tract is linked to the onset of disease with a threshold of 36 or more. The physical basis of this observation remains unclear, but as the expansion exceeds this number, affected molecules become deposited in insoluble aggregates known as inclusion bodies. In fact, disorders resulting from conformational disruption may arise not only by mutation, but also through other physico-chemical effects, as appears to be the case with prion-related diseases. The native prion protein (PrPc) is a largely α-helical molecule (FIGURE 1.76) that undergoes a dramatic structural transition to form fibrillar aggregates with a characteristic β-sheet structure. Again, the cause of this conformational change is still under debate and has been variously proposed to involve posttranslational modifications, metal binding, and other events such as the three-dimensional domain swapping that we described earlier. It appears, however, that the extended beta structures may be a feature of many, if not all, diseases of aggregation, suggesting a common, but still poorly understood, aggregation mechanism.

Among the best-studied diseases of mutation is cancer, and thousands of pro-oncogenic genetic lesions have been identified and mapped within a large array of cell-cycle regulators, signaling molecules/complexes, and others. The first oncogene identified and characterized was derived not from human cells, but from a 'transforming' retrovirus, Rous sarcoma virus. The oncogene product was shown to be nearly identical in sequence to cellular *ras*, a small GTPase. The crucial difference between the virally encoded protein (v-Ras) and the cellular homologue (H-Ras) is a point mutation that results in substitution of a glycine at position 12 to valine (G12V). Ras is the archetypal member of a large superfamily of GTP-binding proteins that function in many different signaling pathways. They generally have a low intrinsic GTP-hydrolysis activity (GTPase) and exist in either GTP-bound or GDP-bound forms that differ in the structural arrangement of two 'switch' regions of the protein, switches I and II (**FIGURE 1.77**). They are in turn controlled by two classes of regulatory molecules that inactivate the GTP-bound form through stimulation of the intrinsic GTPase activity (GTPase-activating proteins or GAPs) and guanine-nucleotide exchange factors (GEFs) that catalyze the replacement of GDP with GTP. In the GTP-bound form, small GTPases are able to bind to and regulate the activity of a wide variety of downstream effector molecules such as protein kinases, whereas the GDP-bound state is inactive for effector interaction. The oncogenic G12V mutation is located in the phosphate-binding loop (P-loop) that forms a structural cradle for the β and γ phosphates of GTP. The mutation has two major effects. First, it lowers the intrinsic rate of GTP hydrolysis, maintaining the active conformation and thus continuously providing *Ras*-dependent growth and proliferation signals. Second, the G12V mutation blocks the productive association of GTP-bound Ras with RasGAP, effectively protecting the GTP-bound state even further (**FIGURE 1.78**). We now know that somatic G12V mutation of normal cellular *ras* occurs in a high proportion of human cancers, classifying the Ras gene as a proto-oncogene.

A second class of molecules that are intimately involved in protecting cells against the effects of cancer-promoting mutations are the so-called tumor suppressors. Of these, one of the best characterized is a tetrameric protein called p53 (53 kDa is its apparent molecular weight on SDS-PAGE gels). p53 has been called the 'gatekeeper' of the cell cycle. It is a modular protein comprising an N-terminal regulatory region, a central DNA-binding domain, and a C-terminal tetramerization motif, and it functions primarily as an activator of the transcription of an inhibitor of cyclin-dependent kinases. Mutations that directly interfere with biochemical activity, or result in reduced expression of p53, are found in the majority of tumors, and occurrence of p53 mutations in the germline

Human prion protein fragment

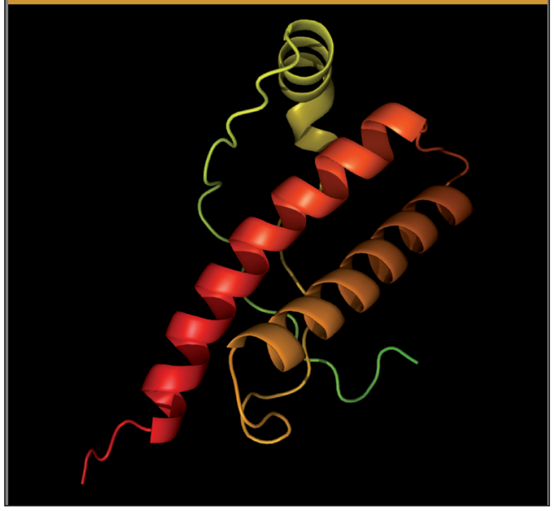

FIGURE 1.76 Sheep prion protein fragment. Image generated from Protein Data Bank file 1UW3.

The GTPase 'switch'

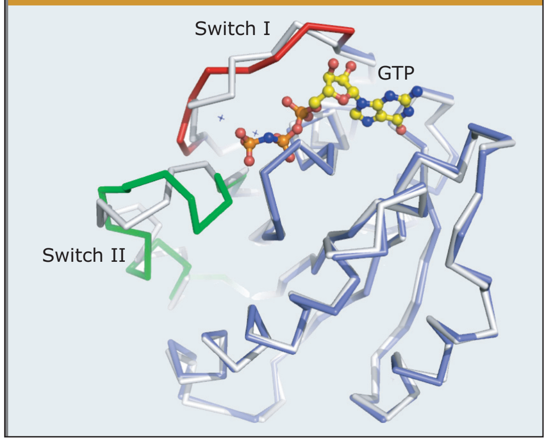

FIGURE 1.77 The GTPase 'switch.' Images generated from Protein Data Bank files 2CL0, 4Q21.

result in a familial predisposition to cancer. Although different classes of mutations may be associated with different cancer types, many occur within the central DNA-binding domain of p53. Here, mutations have a variety of structural effects, including overall destabilization of the domain, but the X-ray structure shows clearly that many mutations occur at residues that are intimately involved in nucleic acid binding (**FIGURE 1.79**).

One of the greatest problems facing modern medicine is that of drug resistance, and structural biology continues to play an important role in understanding and combating it. The extent of the problem is exemplified by the fact that a number of highly pathogenic microorganisms are resistant to virtually every antibiotic in clinical use. Resistance to the action of drug molecules can occur in a number of ways. First,

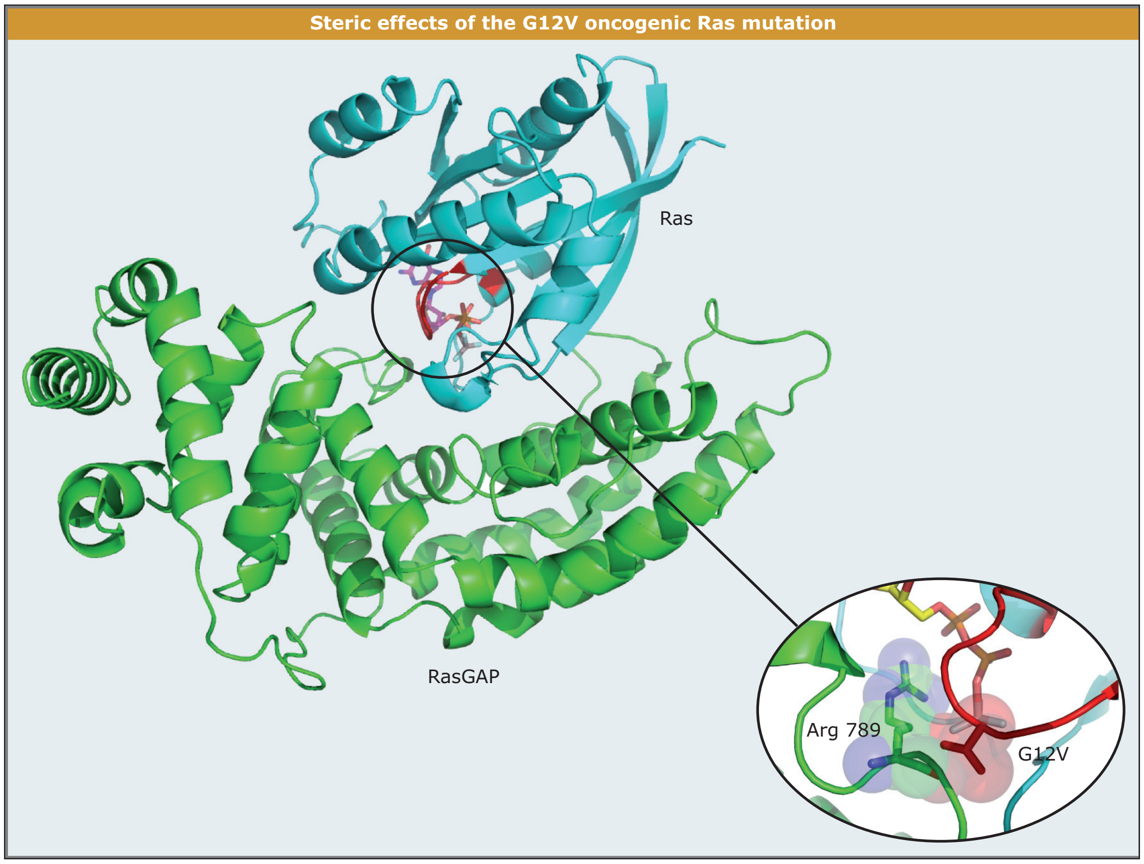

FIGURE 1.78 Steric effects of the G12V oncogenic Ras mutation. Image generated from Protein Data Bank file 1WQ1.

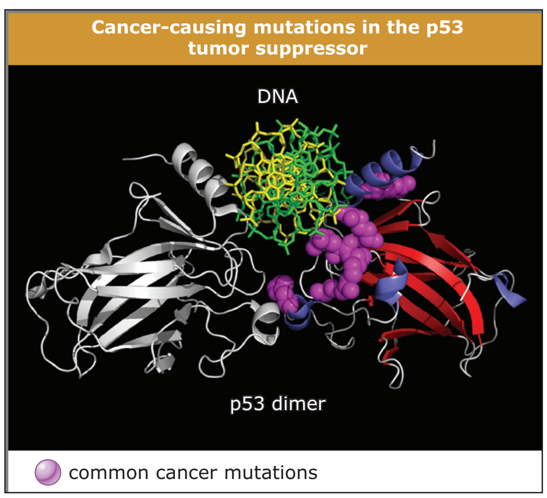

FIGURE 1.79 Cancer-causing mutations in the p53 tumor suppressor. Image generated from Protein Data Bank file 2GEQ.

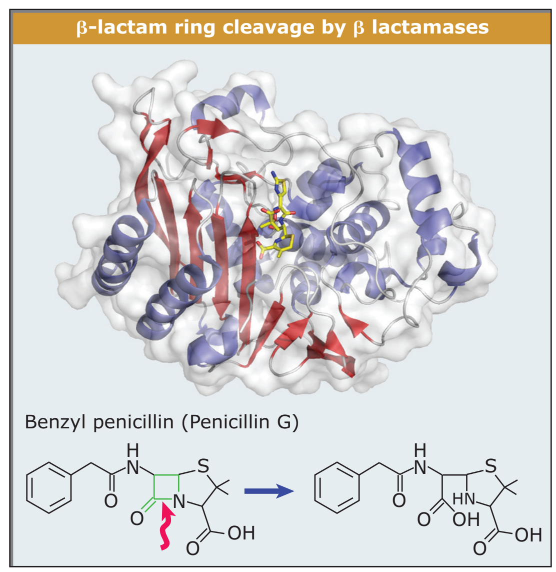

FIGURE 1.80 β-lactam ring cleavage by β lactamases. Image generated from Protein Data Bank file 1IEL.

enzymes that are able to use the drug molecules as substrates and chemically inactivate them are common. For example, penicillin, the first antibiotic identified by Alexander Fleming nearly a century ago, is a member of a large family of β-lactam antibiotics in common clinical use. Resistance to β lactams, however, is common and often mediated by a group of enzymes called β lactamases that are able to cleave the β-lactam ring (FIGURE 1.80). Second, membrane-bound efflux pumps such as the *Escherichia coli* TolC/AcrA/AcrB complex are able to efficiently export a broad spectrum of drug molecules from target cells (FIGURE 1.81). Finally, resistance may be mediated by mutations in the protein targets

A model of the TolC/AcrA/AcrB complex model

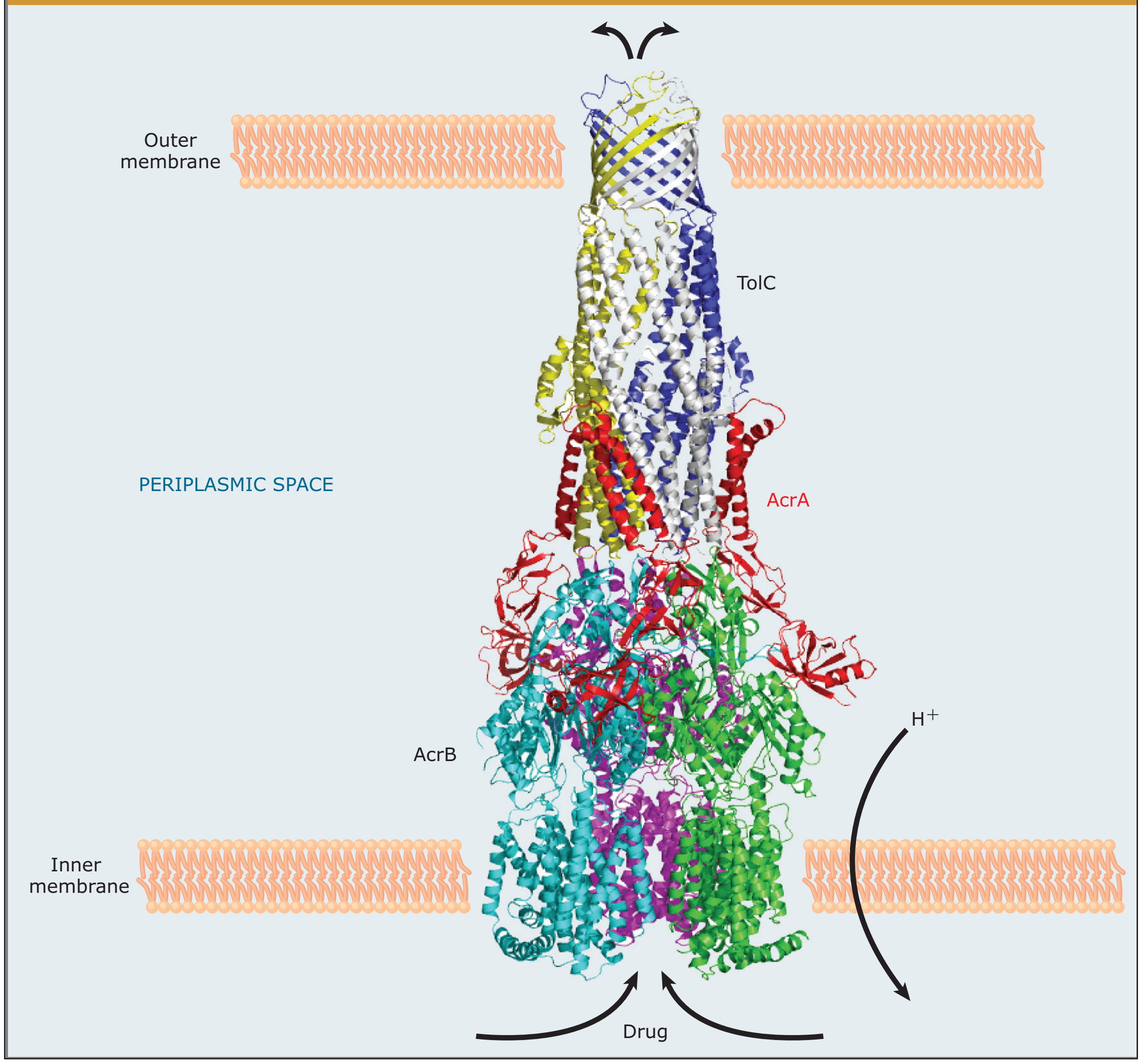

FIGURE 1.81 A model of the TolC/AcrA/AcrB complex model. (Coordinates kindly supplied by Dr. Ben Luisi, Cambridge, UK.)

of small-molecule inhibitors that prevent or reduce the affinity of interaction. For example, the anticancer drugs Iressa and Gleevec/Imatinib target several receptor (e.g., epidermal growth factor receptor) and nonreceptor (e.g., c-Abl) tyrosine kinases as competitive inhibitors of ATP binding. Unfortunately, it appears that cancer cells treated with these compounds can rapidly accumulate mutations that reduce the efficacy of these drugs through a number of effects, including direct steric interference with drug binding (**FIGURE 1.82**). Knowledge of the location and structural effects of these mutations, however, can assist in the design of new compounds that circumvent these problems.

Drug-resistance mutation in c-Abl kinase

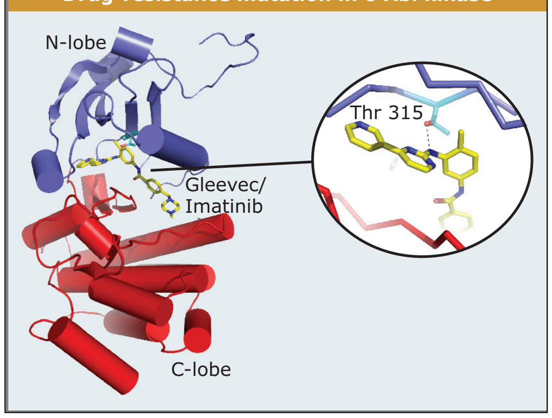

FIGURE 1.82 Drug-resistance mutation in c-Abl kinase. Image generated from Protein Data Bank file 1IEP.

1.16 What's next? Structural biology in the postgenomic era

The advent and success of large-scale genome sequencing has simultaneously shown how diverse biological systems are at all levels of organization and complexity. Although the structural database now contains information for around 40,000 proteins and mutational variants, we must still remember that the human genome alone may encode upward of 30,000 different proteins, of which we know the structures of only a relatively small fraction. If we remember that this basic set of proteins may be posttranslationally modified and that numerous variants may be produced by, for example, differential mRNA splicing, it is clear that much remains to be done merely to characterize individual molecules. To this end, a number of structural genomics consortia have been established around the world, with a view toward substantially increasing the available database of protein structures. These large-scale efforts have had a degree of success, although the question of whether they have fulfilled expectations is still open to debate. It remains to be seen whether the current rate of progress is maintained as more difficult problems, be they individual proteins or complexes, move to the top of the list of targets!

Still elusive is a real understanding of the physical processes that drive and regulate protein folding. The exponential increase in structural information, however, is beginning to influence the efforts of mathematical biologists to predict structure from sequence—not necessarily from first principles, but from a set of empirical rules or guidelines derived from the database of known structures. In addition, although we know the basic principles involved in the binding and specificity of proteins with other proteins and ligands, it is still difficult or even impossible to confidently predict the structural, kinetic, and thermodynamic bases of all but the very simplest of interacting systems. This is crucial because as we have seen, proteins, in general, do not function alone, and the most interesting—and therefore most difficult—challenge for structural biologists is the detailed understanding of the biologically relevant complexes that exist in biological systems.

Many examples of how structural biology has begun to address these outstanding issues have been described in the foregoing sections, and it is clear that technological developments are still being made, allowing more complex and larger structures to be determined at high resolution. In particular, the growth of NMR and cryo-EM into mature methods has considerably added to the arsenal of structural biologists, and the powerful combination of EM with crystal and NMR analyses has already had some impressive successes. Nonetheless, the detailed structural characterization of cellular structures, such as the nuclear pore complex, the centrosome, and other 'mega-complexes' that may be constructed from hundreds of proteins, still seems only a distant possibility.

Perhaps most exciting is the comprehensive integration of structural and mechanistic information into the framework of gene expression, biochemistry, cell biology, and physiology—a goal that is central to the emerging field of 'systems biology.' Clearly, there is much left to do!

1.17 Summary

The three-dimensional structure of protein molecules is intimately associated with their biological function. Over the last 50 years or so we have seen an explosion in the growth of structural databases. X-ray crystallography has, to date, been by far the most successful method for the high-resolution analysis of protein structures and complexes. Modern high-field heteronuclear NMR approaches, however, and the developments in single-particle cryo-EM are now making substantial contributions.

Proteins are made up from a basic and universal set of 20 α amino acids with L-configuration. The sequence of the amino acids in a protein chain encodes the final folded or tertiary structure that may contain elements of secondary structure, α-helices, and β strands. In spite of the diversity of proteins sequences that is observed, the number of distinct structures that exist is likely to be rather small, with perhaps only a few thousand folds covering all globular protein domains.

A specific protein or enzyme may associate with itself or with others to form tight quaternary arrangements that may be crucial for biological activity, and provide structural stability or functional flexibility. Such higher-order arrangements are necessary for allosteric control of activity, one of the most commonly observed regulatory mechanisms. Additional functional and evolutionary versatility is provided by modular protein architectures where individually folded protein domains with different and complementary activities are encoded within a single protein sequence. Fine-tuning and precise control of protein function may also be achieved through posttranslational modifications that may directly affect activity, stability, localization, and so forth.

Proteins are, in the main, rather dynamic, and motions range from small and rapid atomic 'vibrations' through to large-scale conformational changes in different biological contexts. Such motions may arise and be driven in different ways. They may be linked to the hydrolysis of ATP, they may occur as a result of—or prerequisite for—binding of proteins to themselves and/or other ligands, or they may be a product of posttranslational modification. Unwanted conformational changes may also be brought about by mutation, resulting in many genetic diseases, including cancer.

All proteins function through interacting with other proteins or a variety of ligands such as DNA, RNA, lipids, carbohydrates, and small organic and inorganic molecules. The surfaces that they employ in these interactions are tailored by evolution to have affinity and specificity that are appropriate for specific biological functions. Interactions may be transient, with lifetimes of milliseconds or less, or may be longer lived, as seen in many multisubunit complexes. The nature of the interface in each of these types of interactions differs in structure and composition. It is often the case that the primary function of an enzyme, for example, may be carried out by only a few active-site and substrate-binding residues, while the many hundreds of remaining amino acids of the molecule are needed to form a structural scaffold that presents these specific residues to the substrate with exquisitely precise stereochemistry.

In many cases, proteins contain both folded domains along with unstructured regions, usually at the N- and C-terminal ends, that may play a variety of regulatory roles. Indeed, regions within modular proteins are usually connected through unstructured 'linker' regions that allow sufficient structural flexibility to permit, for example, intramolecular interdomain interactions. It now appears that up to ~30% of all protein sequences within eukaryotic genomes may encode unstructured regions. In some cases, an entire protein may contain no tertiary structure, and these molecules have been classified as 'natively unfolded.' The functional advantages remain largely unclear, but certainly include properties of flexibility, ease of posttranslational modification, and the ability to engage binding partners through more extensive interacting surfaces than are possible for folded domains.

Knowledge of the structure of proteins and their complexes informs not only the understanding of underlying biological processes and pathways, but has provided invaluable insights into the nature and cause of many human diseases. For this reason, structural analysis is now firmly established as a major tool in the identification of novel therapeutic compounds and will continue to underpin and drive the developments of new and more powerful computational approaches to the design of new drug molecules and the improvement of existing therapies.

References

1.2 X-ray crystallography and structural biology

References

Hendrickson, W. A., Horton, J. R., and LeMaster, D. M. (1990). Selenomethionyl proteins produced for analysis by multiwavelength anomalous diffraction (MAD): a vehicle for direct determination of three-dimensional structure. *EMBO J* 9, 1665–1672.

Kendrew, J. C., Bodo, G., Dintzis, H. M., Parrish, R. G., Wyckoff, H., and Phillips, D. C. (1958). A three-dimensional model of the myoglobin molecule obtained by x-ray analysis. *Nature* 181, 662–666.

Muirhead, H., and Perutz, M. F. (1963). Structure of haemoglobin: a three-dimensional Fourier synthesis of reduced human haemoglobin at 5-5Å resolution. *Nature* 199, 633–638.

Yang, W., Hendrickson, W. A., Crouch, R. J., and Satow, Y. (1990). Structure of ribonuclease H phased at 2 Å resolution by MAD analysis of the selenomethionyl protein. *Science* 249, 1398–1405.

1.3 Nuclear magnetic resonance

Reviews

Wuthrich, K. (1990). Protein structure determination in solution by NMR spectroscopy. *J Biol Chem* 265, 22059–22062.

References

Fiaux, J., Bertelsen, E. B., Horwich, A. L., and Wuthrich, K. (2002). NMR analysis of a 900K GroEL GroES complex. *Nature* 418, 207–211.

Horst, R., Bertelsen, E. B., Fiaux, J., Wider, G., Horwich, A. L., and Wuthrich, K. (2005). Direct NMR observation of a substrate protein

bound to the chaperonin GroEL. *Proc Natl Acad Sci USA* 102, 12748–12753.
Ikura, M., Krinks, M., Torchia, D. A., and Bax, A. (1990). An efficient NMR approach for obtaining sequence-specific resonance assignments of larger proteins based on multiple isotopic labeling. *FEBS Lett* 266, 155–158.
Pervushin, K., Riek, R., Wider, G., and Wuthrich, K. (1997). Attenuated T2 relaxation by mutual cancellation of dipole–dipole coupling and chemical shift anisotropy indicates an avenue to NMR structures of very large biological macromolecules in solution. *Proc Natl Acad Sci USA* 94, 12366–12371.

1.4 Electron microscopy of biomolecules and their complexes

Reviews

Frank, J. (2002). Single-particle imaging of macromolecules by cryo-electron microscopy. *Annu Rev Biophys Biomol Struct* 31, 303–319.
Frank, J., Wagenknecht, T., McEwen, B. F., Marko, M., Hsieh, C. E., and Mannella, C. A. (2002). Three-dimensional imaging of biological complexity. *J Struct Biol* 138, 85–91.
Lucic, V., Forster, F., and Baumeister, W. (2005). Structural studies by electron tomography: from cells to molecules. *Annu Rev Biochem* 74, 833–865.
Rossmann, M. G. (2000). Fitting atomic models into electron-microscopy maps. *Acta Crystallogr D Biol Crystallogr* 56, 1341–1349.
Rossmann, M. G., Mesyanzhinov, V. V., Arisaka, F., and Leiman, P. G. (2004). The bacteriophage T4 DNA injection machine. *Curr Opin Struct Biol* 14, 171–180.

References

Schertler, G. F., Villa, C., and Henderson, R. (1993). Projection structure of rhodopsin. *Nature* 362, 770–772.

1.5 Protein structure representations—a primer

Reviews

Richardson, J. S. (1981). The anatomy and taxonomy of protein structure. *Adv Protein Chem* 34,167–339.

1.6 Proteins are linear chains of amino acids—primary structure

References

Pauling, L., and Corey, R. B. (1953). Stable configurations of polypeptide chains. *Proc R Soc Lond B Biol Sci* 141, 21–33.
Ramachandran, G. N., Ramakrishnan, C., and Sasisekharan, V. (1963). Stereochemistry of polypeptide chain configurations. *J Mol Biol* 7, 95–99.

1.7 Secondary structure—the fundamental unit of protein architecture

References

Blake C.C., Koenig D.F., Mair G.A., North A.C., Phillips D.C., and Sarma V.R. (1965). Structure of hen egg-white lysozyme: a three-dimensional Fourier synthesis at 2 Angstrom resolution. *Nature* 206, 757–761.
Crick, F. H. (1952). Is alpha-keratin a coiled coil? *Nature* 170, 882–883.
Hol, W. G., van Duijnen, P. T., and Berendsen, H. J. (1978). The alpha-helix dipole and the properties of proteins. *Nature* 273, 443–446.
Pauling, L., and Corey, R. B. (1951). Atomic coordinates and structure factors for two helical configurations of polypeptide chains. *Proc Natl Acad Sci USA* 37, 235–240.

1.8 Tertiary structure and the universe of protein folds

Reviews

Liu, Y., and Eisenberg, D. (2002). 3D domain swapping: as domains continue to swap. *Protein Sci* 11, 1285–1299.

References

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. *Science* 181, 223–230.
Holmgren, A., Kuehn, M. J., Branden, C. I., and Hultgren, S. J. (1992). Conserved imunoglobulin-like features in a family of periplasmic pilus chaperones in bacteria. *Embo J* 11, 1617–1622.
Lecomte, J. T., Vuletich, D. A., and Lesk, A. M. (2005). Structural divergence and distant relationships in proteins: evolution of the globins. *Curr Opin Struct Biol* 15, 290–301.
Levinthal, C. (1969) Are there pathways for protein folding? *J Chim Phys* 65, 44–45.
Muller, C. W., Rey, F. A., Sodeoka, M., Verdine, G. L., and Harrison, S. C. (1995). Structure of the NF-kappa B p50 homodimer bound to DNA. *Nature* 373, 311–317.
Nagano, N., Orengo, C. A., and Thornton, J. M. (2002). One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J Mol Biol* 321, 741–765.
Taylor, W. R. (2000). A deeply knotted protein structure and how it might fold. *Nature* 406, 916–919.
Taylor, W. R. (2002). A 'periodic table' for protein structures. *Nature* 416, 657–660.

1.9 Modular architectures and repeat motifs

Reviews

Andrade, M. A., Petosa, C., O'Donoghue, S. I., Muller, C. W., and Bork, P. (2001). Comparison of ARM and HEAT protein repeats. *J Mol Biol* 309, 1–18.

Groves, M. R., and Barford, D. (1999). Topological characteristics of helical repeat proteins. *Curr Opin Struct Biol* 9, 383–389.

Pawson, T. (1995). Protein modules and signalling networks. *Nature* 373, 573–580.

Sedgwick, S. G., and Smerdon, S. J. (1999). The ankyrin repeat: a diversity of interactions on a common structural framework. *Trends Biochem Sci* 24, 311–316.

Tskhovrebova, L., and Trinick, J. (2003). Titin: properties and family relationships. *Nat Rev Mol Cell Biol* 4, 679–689.

Tybulewicz, V. L. (2005). Vav-family proteins in T-cell signalling. *Curr Opin Immunol* 17, 267–274.

References

Couture, J. F., Collazo, E., and Trievel, R. C. (2006). Molecular recognition of histone H3 by the WD40 protein WDR5. *Nat Struct Mol Biol* 13, 698–703.

Emsley, P., Charles, I. G., Fairweather, N. F., and Isaacs, N. W. (1996). Structure of *Bordetella pertussis* virulence factor P.69 pertactin. *Nature* 381, 90–92.

Han, Z., Guo, L., Wang, H., Shen, Y., Deng, X. W., and Chai, J. (2006). Structural basis for the specific recognition of methylated histone H3 lysine 4 by the WD-40 protein WDR5. *Mol Cell* 22, 137–144.

Renault, L., Nassar, N., Vetter, I., Becker, J., Klebe, C., Roth, M., and Wittinghofer, A. (1998). The 1.7 Å crystal structure of the regulator of chromosome condensation (RCC1) reveals a seven-bladed propeller. *Nature* 392, 97–101.

Ruthenburg, A. J., Wang, W., Graybosch, D. M., Li, H., Allis, C. D., Patel, D. J., and Verdine, G. L. (2006). Histone H3 recognition and presentation by the WDR5 module of the MLL1 complex. *Nat Struct Mol Biol* 13, 704–712.

Schuetz, A., Allali-Hassani, A., Martin, F., Loppnau, P., Vedadi, M., Bochkarev, A., Plotnikov, A. N., Arrowsmith, C. H., and Min, J. (2006). Structural basis for molecular recognition and presentation of histone H3 by WDR5. *Embo J* 25, 4245–4252.

Sicheri, F., Moarefi, I., and Kuriyan, J. (1997). Crystal structure of the Src family tyrosine kinase Hck. *Nature* 385, 602–609.

Sondek, J., Bohm, A., Lambright, D. G., Hamm, H. E., and Sigler, P. B. (1996). Crystal structure of a G-protein beta gamma dimer at 2.1Å resolution. *Nature* 379, 369–374.

Xu, W., Doshi, A., Lei, M., Eck, M. J., and Harrison, S. C. (1999). Crystal structures of c-Src reveal features of its autoinhibitory mechanism. *Mol Cell* 3, 629–638.

1.10 Quaternary structure and higher-order assemblies

Reviews

Goodsell, D. S., and Olson, A. J. (2000). Structural symmetry and protein function. *Annu Rev Biophys Biomol Struct* 29, 105–153.

References

Caspar, D. L., and Klug, A. (1962). Physical principles in the construction of regular viruses. *Cold Spring Harb Symp Quant Biol* 27, 1–24.

Mattevi, A., Obmolova, G., Schulze, E., Kalk, K. H., Westphal, A. H., de Kok, A., and Hol, W. G. (1992). Atomic structure of the cubic core of the pyruvate dehydrogenase multienzyme complex. *Science* 255, 1544–1550.

Murakami, K. S., Masuda, S., Campbell, E. A., Muzzin, O., and Darst, S. A. (2002). Structural basis of transcription initiation: an RNA polymerase holoenzyme-DNA complex. *Science* 296, 1285–1290.

Winkler, F. K., Schutt, C. E., Harrison, S. C., and Bricogne, G. (1977). Tomato bushy stunt virus at 5.5-Å resolution. *Nature* 265, 509–513.

1.11 Posttranslational modifications and cofactors

Reviews

Fuentes-Prior, P., and Salvesen, G. S. (2004). The protein structures that shape caspase activity, specificity, activation and inhibition. *Biochem J* 384, 201–232.

Jenuwein, T., and Allis, C. D. (2001). Translating the histone code. *Science* 293, 1074–1080.

Seet, B. T., Dikic, I., Zhou, M. M., and Pawson, T. (2006). Reading protein modifications with interaction domains. *Nat Rev Mol Cell Biol* 7, 473–483.

Tsien, R. Y. (1998). The green fluorescent protein. *Annu Rev Biochem* 67, 509–544.

Yaffe, M. B., and Smerdon, S. J. (2004). The use of in vitro peptide-library screens in the analysis of phosphoserine/threonine-binding domain structure and function. *Annu Rev Biophys Biomol Struct* 33, 225–244.

1.12 Dynamics, flexibility, and conformational changes

Reviews

Endicott, J. A., Noble, M. E., and Tucker, J. A. (1999). Cyclin-dependent kinases: inhibition and substrate recognition. *Curr Opin Struct Biol* 9, 738–744.

References

Anderson, C. M., Zucker, F. H., and Steitz, T. A. (1979). Space-filling models of kinase clefts and conformation changes. *Science* 204, 375–380.

Elber, R., and Karplus, M. (1987). Multiple conformational states of proteins: a molecular dynamics analysis of myoglobin. *Science* 235, 318–321.

Koshland, D. E. (1958). Application of a theory of enzyme specificity to protein synthesis. *Proc Natl Acad Sci USA* 44, 98–104.

Koshland, D. E., Jr., Nemethy, G., and Filmer, D. (1966). Comparison of experimental binding data and theoretical models in proteins containing subunits. *Biochemistry* 5, 365–385.

Monod, J., Wyman, J., and Changeux, J. P. (1965). On the nature of allosteric transitions: a plausible model. *J Mol Biol* 12, 88–118.

1.13 Protein–protein and protein–nucleic acid interactions

Reviews

Auweter, S. D., Oberstrass, F. C., and Allain, F. H. (2006). Sequence-specific binding of single-stranded RNA: is there a code for recognition? *Nucleic Acids Res* 34, 4943–4959.

Draper, D. E. (1995). Protein-RNA recognition. *Annu Rev Biochem* 64, 593–620.

Fersht, A. R. (1987). Dissection of the structure and activity of the tyrosyl-tRNA synthetase by site-directed mutagenesis. *Biochemistry* 26, 8031–8037.

Janin, J. (1999). Wet and dry interfaces: the role of solvent in protein–protein and protein–DNA recognition. *Structure* 7, R277–279.

Janin, J., Henrick, K., Moult, J., Eyck, L. T., Sternberg, M. J., Vajda, S., Vakser, I., and Wodak, S. J. (2003). CAPRI: a critical assessment of predicted interactions. *Proteins* 52, 2–9.

Noller, H. F. (2005). RNA structure: reading the ribosome. *Science* 309, 1508–1514.

Nooren, I. M., and Thornton, J. M. (2003). Structural characterisation and functional significance of transient protein–protein interactions. *J Mol Biol* 325, 991–1018.

Steitz, T. A. (1990). Structural studies of protein–nucleic acid interaction: the sources of sequence-specific binding. *Q Rev Biophys* 23, 205–280.

References

Lee, B., and Richards, F. M. (1971). The interpretation of protein structures: estimation of static accessibility. *J Mol Biol* 55, 379–400.

Somers, W. S., and Phillips, S. E. (1992). Crystal structure of the met repressor–operator complex at 2.8 Å resolution reveals DNA recognition by beta-strands. *Nature* 359, 387–393.

Stucki, M., Clapperton, J. A., Mohammad, D., Yaffe, M. B., Smerdon, S. J., and Jackson, S. P. (2005). MDC1 directly binds phosphorylated histone H2AX to regulate cellular responses to DNA double-strand breaks. *Cell* 123, 1213–1226.

1.14 Function without structure?

Reviews

Fink, A. L. (2005). Natively unfolded proteins. *Curr Opin Struct Biol* 15, 35–41.

Spolar, R. S., and Record, M. T., Jr. (1994). Coupling of local folding to site-specific binding of proteins to DNA. *Science* 263, 777–784.

Wright, P. E., and Dyson, H. J. (1999). Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *J Mol Biol* 293, 321–331.

References

Russo, A. A., Jeffrey, P. D., Patten, A. K., Massague, J., and Pavletich, N. P. (1996). Crystal structure of the p27Kip1 cyclin-dependent-kinase inhibitor bound to the cyclin A-Cdk2 complex. *Nature* 382, 325–331.

Zor, T., Mayr, B. M., Dyson, H. J., Montminy, M. R., and Wright, P. E. (2002). Roles of phosphorylation and helix propensity in the binding of the KIX domain of CREB-binding protein by constitutive (c-Myb) and inducible (CREB) activators. *J Biol Chem* 277, 42241–42248.

1.15 Structure and medicine

Reviews

Bennett, M. J., Sawaya, M. R., and Eisenberg, D. (2006). Deposition diseases and 3D domain swapping. *Structure* 14, 811–824.

Nagar, B., Bornmann, W. G., Pellicena, P., Schindler, T., Veach, D. R., Miller, W. T., Clarkson, B., and Kuriyan, J. (2002). Crystal structures of the kinase domain of c-Abl in complex with the small molecule inhibitors PD173955 and imatinib (STI-571). *Cancer Res* 62, 4236–4243.

Ross, C. A., and Poirier, M. A. (2004). Protein aggregation and neurodegenerative disease. *Nat Med* 10 Suppl, S10–17.

Wilke, M. S., Lovering, A. L., and Strynadka, N. C. (2005). Beta-lactam antibiotic resistance: a current structural perspective. *Curr Opin Microbiol* 8, 525–533.

Wittinghofer, A., and Nassar, N. (1996). How Ras-related proteins talk to their effectors. *Trends Biochem Sci* 21, 488–491.

References

Cho, Y., Gorina, S., Jeffrey, P. D., and Pavletich, N. P. (1994). Crystal structure of a p53 tumor suppressor–DNA complex: understanding tumorigenic mutations. *Science* 265, 346–355.

Fernandez-Recio, J., Walas, F., Federici, L., Venkatesh Pratap, J., Bavro, V. N., Miguel, R. N., Mizuguchi, K., and Luisi, B. (2004). A model of a transmembrane drug-efflux pump from Gram-negative bacteria. *FEBS Lett* 578, 5–9.

Ho, W. C., Fitzgerald, M. X., and Marmorstein, R. (2006). Structure of the p53 core domain dimer bound to DNA. *J Biol Chem* 281, 20494–20502.

Perutz, M. F., and Mitchison, J. M. (1950). State of hemoglobin in sickle-cell anaemia. *Nature* 166, 677–679.

Scheffzek, K., Ahmadian, M. R., Kabsch, W., Wiesmuller, L., Lautwein, A., Schmitz, F., and Wittinghofer, A. (1997). The Ras-RasGAP complex: structural basis for GTPase activation and its loss in oncogenic Ras mutants. *Science* 277, 333–338.

1.16 What's next? Structural biology in the postgenomic era

References

Chandonia, J. M., and Brenner, S. E. (2006). The impact of structural genomics: expectations and outcomes. *Science* 311, 347–351.