5

Structure Determination

Most of the protein structures described and discussed in this book have been determined either by X-ray crystallography or by nuclear magnetic resonance (NMR) spectroscopy. Although these techniques both depend on data derived from physical techniques for probing structure, their interpretation is not unambiguous and entails assumptions and approximations often depending upon knowledge of the protein from other sources, including its biology. This chapter briefly describes how structures are determined by X-ray crystallography and NMR, how the data are interpreted, and what contributes to the accuracy of a structure determination.

5-1 The Interpretation of Structural Information

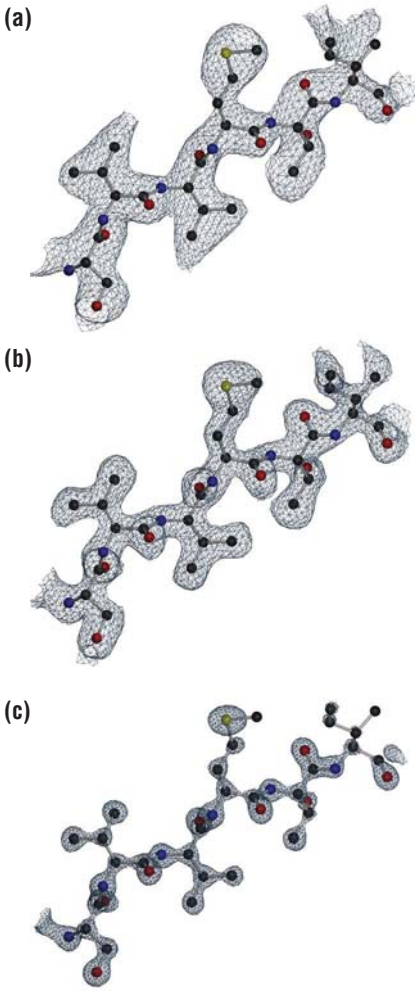

Figure 5-1 Portion of a protein electron density map at three different resolutions The peptide corresponding to the electron density map (serine, valine, valine, methionine, threonine, isoleucine) is superimposed. **(a)** At 3-Å resolution the fold of the polypeptide chain can be seen and the approximate positions of side chains can be determined. Interatomic distances can only be measured to ± about 0.5 Å. **(b)** At 2-Å resolution side chains are well delineated and the peptide carbonyls of the backbone are discernible, allowing the chain to be oriented with precision. Interatomic distances can be measured to a precision of about 0.2 Å. Approximately three times as many data are required for this resolution as were used at 3-Å resolution. **(c)** At 1-Å resolution atoms are visible and resolved. Interatomic distances can be measured to a precision of a few hundredths of an Ångstrom. Almost 30 times more data are required for this resolution as were used at 3-Å resolution. In favorable cases, the position of hydrogen atoms can be inferred at this resolution. Note that at high resolution (c) there is no electron density for the methyl group of the methionine side chain because it is disordered. At lower resolutions (a and b), it appears that such density is present, but this is actually the density of the heavier sulfur atom. Since at low resolution electron density is more diffuse, the entire thio-ether portion of the side chain can be fitted into the sulfur density. Kindly provided by Aaron Moulin.

Experimentally determined protein structures are the result of the interpretation of different types of data

Much of the content of this book depends on a detailed understanding of the structure of proteins at the atomic level. The atomic structures of biological macromolecules can be determined by several techniques: although a few have been determined by electron microscopy, the vast majority have been obtained by either single crystal *X-ray diffraction*, generally known as *X-ray crystallography*, or by *nuclear magnetic resonance spectroscopy*, generally known as NMR. How this is done, in both cases, is briefly explained in the next section. Here, we describe the kind of information these techniques produce, and how it can be interpreted.

Both X-ray crystallography and NMR produce information on the relative positions of the atoms of the molecule: these are termed **atomic coordinates**. Because X-rays are scattered from the electron cloud of the atoms, whereas NMR measures the interactions of atomic nuclei, X-ray crystallography provides the positions of all non-hydrogen atoms whereas NMR provides the positions of all atoms including hydrogen. However, the coordinate sets provided in this way represent the interpretation of the primary data obtained in each case. The precision and accuracy with which coordinates can be derived from these data depend on several factors that are different for the two methods. The objective end-product of a crystallographic structure determination is an **electron density map** (Figure 5-1), which is essentially a contour plot indicating those regions in the crystal where the electrons in the molecule are to be found. Human beings must interpret this electron density map in terms of an atomic model, aided by semi-automatic computational procedures. The objective end-product of an NMR structure determination is usually a set of distances between atomic nuclei that define both bonded and non-bonded close contacts in the molecule. These must be interpreted in terms of a molecular structure, a process that is aided by automated methods.

Because interpretation in each case requires assumptions and approximations, macromolecular structures can have errors. In most cases, these errors are small, and only affect a small part of the structure. However, such errors can be quite large: cases of completely incorrect structures, though rare, have been reported.

The choice of technique depends on several factors, including the molecular weight, solubility and ease of crystallization of the protein in question. For proteins and protein complexes with molecular weights above 50–100 kDa, X-ray crystallography is the method of choice. For smaller proteins or protein complexes, either method may be usable but X-ray diffraction will usually provide more precise structural information than NMR provided the species crystallizes easily and is well ordered in the crystal. For molecules that are hard to crystallize and can be dissolved at reasonably high concentrations without aggregation, NMR is the method of choice. In many cases however, both techniques can be used and provide complementary information, since crystallographic images are static but NMR can be used to study the flexibility of proteins and their dynamics over a wide range of time scales.

Both the accuracy and the precision of a structure can vary

It is essential to appreciate the distinction between the accuracy of an experimentally determined structure and its precision. The latter is much easier to assess than the former. We define the precision of a structure determination in terms of the reproducibility of its atomic coordinates. If the data allow the equilibrium position of each atom to be determined precisely, then the

Definitions

atomic coordinate: the position in three-dimensional space of an atom in a molecule relative to all other atoms in the molecule.

electron density map: a contour plot showing the distribution of electrons around the atoms of a molecule.

resolution: the level of detail that can be derived from a given process.

same structure determined independently elsewhere should yield atomic coordinates that agree very closely with the first set. Precise coordinates are usually reproducible to within a few tenths of an Ångstrom.

Accuracy refers to whether or not the structure is correct, but there are two ways to define correctness. A structure may be right but irrelevant: the protein may be in a conformation that is determined by the experimental conditions rather than its biological context. Such cases are rare, but they have occurred. It is also possible to interpret the experimental data incorrectly, yielding a structure that is not and could not possibly be correct, either in whole (rarely) or in part. Often it is possible to determine that a structure has been built incorrectly on the basis of the known rules of how proteins fold and their properties when they are folded. It is not possible to determine the correctness of minor details unless there is a biological experiment that can probe the structure at that position. For example, the involvement of a residue in catalysis predicted by the structure can be tested by mutating it to one that cannot perform the required catalytic function. Ultimately, the accuracy of a structure can best be assessed by the answer to a simple question: is the structure consistent with the body of biochemical and biological information about the protein?

Sometimes, part of a structure may be invisible to the experimental method used. This can arise because that part of the molecule is disordered and therefore does not contribute to the reflected X-rays in a crystallographic experiment or, in an NMR experiment, atoms may not be close enough to interact strongly.

The information content of a structure is determined by its resolution

Precision and accuracy of a structure are related: the more precise a structure determination is, the less likely it is to have gross errors. What links precision with accuracy is the concept of **resolution**.

Crystallographers express the resolution of a structure in terms of a distance: if a structure has been determined at 2-Å resolution then any atoms separated by more than about this distance will appear as separate maxima in the electron density contour plot, as we explain in the next section, and their positions can be obtained directly to high precision. Atoms closer together than the resolution limit will appear as a fused electron density feature, and their exact positions must be inferred from the shape of the electron density and knowledge of the chemical structure of amino acids or nucleotides (Figure 5-1). Since the average C–C single bond distance is 1.5 Å, the precision with which the individual atoms of, say, an inhibitor bound to an enzyme can be located will be much greater at 1.5-Å resolution than in a structure determined at, say, 3-Å resolution.

NMR spectroscopists express resolution somewhat differently. NMR structures are determined as ensembles of similar models (Figure 5-2): we explain in the next section how these are derived. However, since more than one closely related model will fit the data, the effective resolution of an NMR structure is given by the extent of the differences between these models, expressed as a root-mean-square deviation (RMSD) of their atomic coordinates. The smaller the deviation, the more precise (and, presumably accurate) the NMR structure is believed to be, and therefore the higher its effective resolution.

In the case of both X-ray and NMR structures, what sets the resolution limit is the intrinsic order of the protein plus the amount of data that the experimenter is able to measure. The greater the amount of data, the higher the resolution, and the higher the resolution, the more precise and accurate the information that can be extracted from the structure.

In favorable cases, the resolution of a macromolecular structure determination by crystallography is such that the relative positions of all non-hydrogen atoms are known to a precision of a few tenths of an Ångstrom.

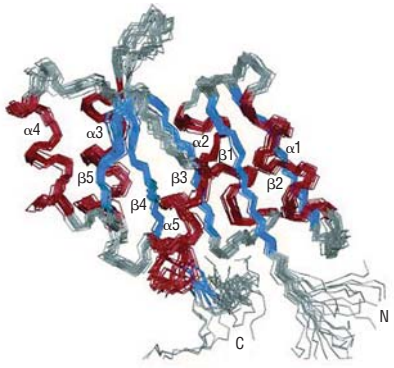

Figure 5-2 NMR structure ensemble The figure shows the superposition of the set of models derived from the internuclear distances measured for this protein in solution. Note that different portions of the structure are determined with different precision. Blue represents beta strands, red represents alpha helices, grey represents loops. This figure should not be taken to indicate the flexibility of segments of the protein: different regions may be poorly defined because there are insufficient data to constrain the structure, and not because the structures are mobile.

5-2 Structure Determination by X-Ray Crystallography and NMR

Protein crystallography involves summing the scattered X-ray waves from a macromolecular crystal

The steps in solving a protein crystal structure at high resolution are diagrammed in Figure 5-3. First, the protein must be crystallized. This is often the rate-limiting step in straightforward structure determinations, especially for membrane proteins. Then, the X-ray diffraction pattern from the crystal must be recorded. When X-rays strike a macromolecular crystal, the atoms in the molecules produce scattered X-ray waves which combine to give a complex diffraction pattern consisting of waves of different amplitudes. What is measured experimentally are the amplitudes and positions of the scattered X-ray waves from the crystal. The structure can be reconstructed by summing these waves, but each one must be in the correct registration with respect to every other wave, that is, the origin of each wave must be determined so that they sum to give some image instead of a sea of noise. This is called the **phase problem**. Phase values must be assigned to all of the recorded data; this can sometimes be done computationally, but is usually done experimentally by labeling the protein with one or more heavy atoms whose position in the crystal can be determined independently. The phased waves are then summed in three dimensions to generate an image of the electron density distribution of the molecule in the crystal. This can be done semi-automatically or by hand on a computer graphics system. A chemical model of part of the molecule is docked into the shape of each part of the electron part of the electron density (as shown in Figure 5-3). This fitting provides the first picture of the structure of the protein. The overall model is improved by an iterative process called refinement whereby the positions of the atoms in the model are tweaked until the calculated diffraction pattern from the model agrees as well as can be with the experimentally measured diffraction pattern from the actual protein. There is no practical limit to the size of the protein or protein complex whose structure can be determined by X-ray crystallography.

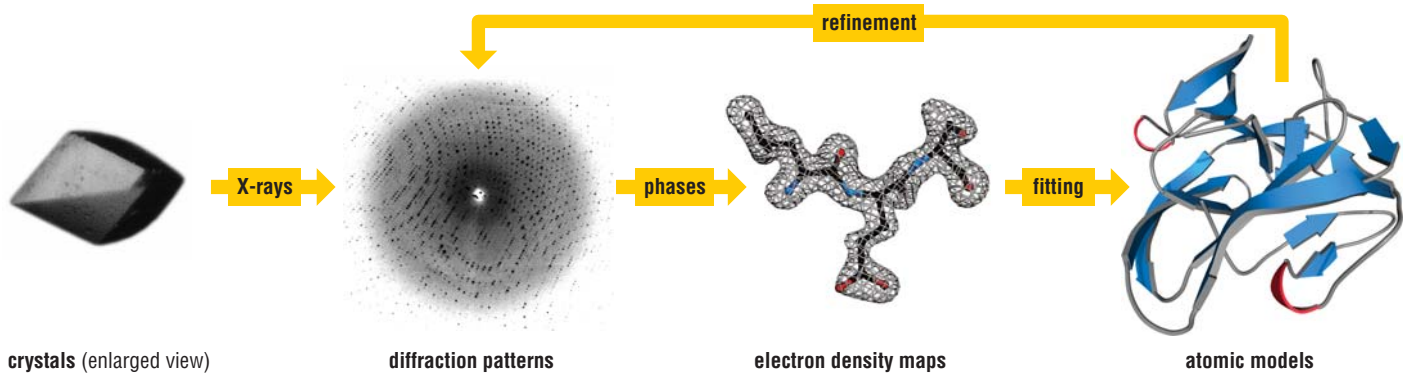

Figure 5-3 Structure determination by X-ray crystallography The first step in structure determination by X-ray crystallography is the crystallization of the protein. The source of the X-rays is often a synchrotron and in this case the typical size for a crystal for data collection may be $0.3 \times 0.3 \times 0.1$ mm. The crystals are bombarded with X-rays which are scattered from the planes of the crystal lattice and are captured as a diffraction pattern on a detector such as film or an electronic device. From this pattern, and with the use of reference—or phase—information from labeled atoms in the crystal, electron density maps (shown here with the corresponding peptide superimposed) are computed for different parts of the crystal. A model of the protein is constructed from the electron density maps and the diffraction pattern for the modeled protein is calculated and compared with the actual diffraction pattern. The model is then adjusted—or refined—to reduce the difference between its calculated diffraction pattern and the pattern obtained from the crystal, until the correspondence between model and reality is as good as possible. The quality of the structure determination is measured as the percentage difference between the calculated and the actual pattern.

Definitions

phase problem: in the measurement of data from an X-ray crystallographic experiment only the amplitude of the wave is determined. To compute a structure, the phase must also be known. Since it cannot be determined directly, it must be determined indirectly or by some other experiment.

NMR spectroscopy involves determining internuclear distances by measuring perturbations between assigned resonances from atoms in the protein in solution

Unlike crystallography, structure determination by NMR (Figure 5-4) is carried out on proteins in solution, but the protein must be soluble without aggregating at concentrations close to those of a protein in a crystal lattice. NMR structure determination requires two types of data. The first is measurement of nuclear magnetic resonances from protons and isotopically labeled carbons and nitrogens in the molecule. Different nuclei in a protein absorb electromagnetic energy (resonance) at different frequencies because their local electromagnetic environments differ due to the three-dimensional structure of the protein. These resonances must be assigned to atoms in specific amino acids in the protein sequence, a process that requires several specific types of experiments. The second set of data consists of internuclear distances that are inferred from perturbing the resonance of different atoms and observing which resonances respond; only atoms within 5 Å of each other show this effect, and its magnitude varies with the distance between them. The set of approximate internuclear distances is then used to compute a structure model consistent with the data. Since the distances are imprecise, many closely related models may be consistent with the observations, so NMR structures are usually reported as ensembles of atomic coordinates. In practice, if a structure is determined by both NMR and X-ray diffraction, it is usually found that the average of the NMR ensemble closely resembles the crystal structure. As a general rule, there are practical limitations to the determination of the structure of a complete protein by NMR: a molecular weight of about 50 kDa is considered a very large protein for NMR. In special cases, domains or portions of much larger proteins or complexes can be studied.

Unlike X-ray diffraction, which presents a static picture (an average in time and space) of the structure of a protein, NMR has the capability of measuring certain dynamic properties of proteins over a wide range of time scales.

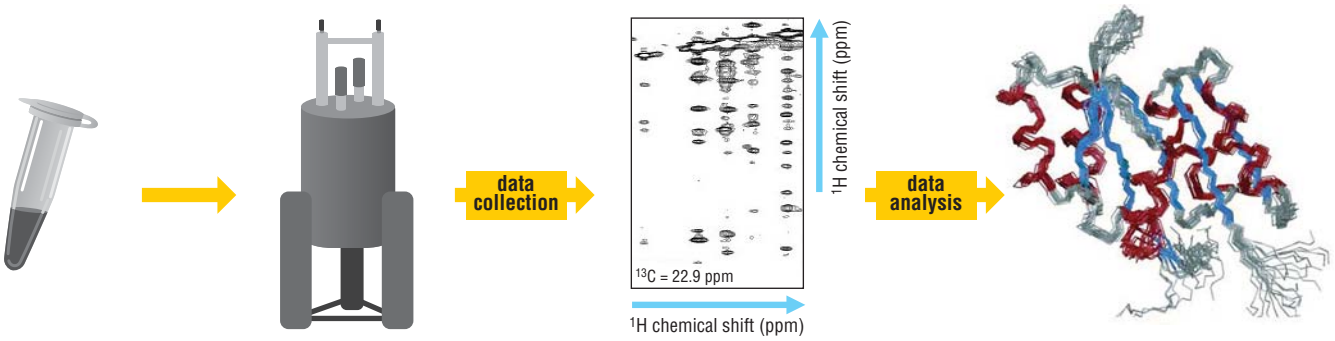

Figure 5-4 Structure determination by NMR For protein structure determination by NMR, a labeled protein is dissolved at very high concentration and placed in a magnetic field, which causes the spin of the hydrogen atoms to align along the field. Radio frequency pulses are then applied to the sample, perturbing the nuclei of the atoms which when they relax back to their original state emit radio frequency radiation whose properties are determined by the environment of the atom in the protein. This emitted radiation is recorded in the NMR spectrometer for pulses of differing types and durations (for simplicity, only one such record is shown here), and compared with a reference signal to give a measure known as the chemical shift. The relative positions of the atoms in the molecule are calculated from these data to give a series of models of the protein which can account for these data. The quality of the structure determination is measured as the difference between the different models.

References

Drenth, J.: *Principles of Protein X-Ray Crystallography* 2nd ed. (Springer-Verlag, New York, 1999).

Evans, J.N.S.: *Biomolecular NMR Spectroscopy* (Oxford University Press, Oxford, 1995).

Markley, J.L. *et al.*: **Macromolecular structure determination by NMR spectroscopy.** *Methods Biochem. Anal.* 2003, **44**:89–113.

Rhodes, G.: *Crystallography Made Crystal Clear: A Guide to Users of Macromolecular Models* 2nd ed. (Academic Press, New York and London, 1999).

Schmidt, A. and Lamzin, V.S.: ***Veni, vidi, vici*—atomic resolution unravelling the mysteries of protein function.** *Curr. Opin. Struct. Biol.* 2002, **12**:698–703.

Gorenstein, N.: *Nuclear Magnetic Resonance (NMR)* (Biophysics Textbooks Online): http://www.biophysics.org/btol/NMR.html

5-3 Quality and Representation of Crystal and NMR Structures

The quality of a finished structure depends largely on the amount of data collected

Both X-ray and NMR structure determination have statistical criteria for the quality of the atomic model produced. Crystallographers usually speak of R-factors, which represent the percentage disagreement between the observed diffraction pattern and that calculated from the final model. R-factors of around 20% or less are considered indicative of well determined structures that are expected to contain relatively few errors. NMR spectroscopists usually report overall root mean square deviation (RMSD) between the atoms in secondary structure elements in all coordinate sets in the ensemble of structures consistent with the experimental data. In practice, RMSDs of 0.7 Å are considered good, indicating a structure determination of high precision. RMSDs of around 1 Å are considered acceptable.

There is no substitute for high resolution. It makes the structure determination easier and more reliable. The closer one gets to true atomic resolution in X-ray crystallography (better than 1.5 Å), the less ambiguity one has in positioning every atom. Atomic resolution allows one to detect mistakes in the biochemically determined or genomically derived amino-acid sequence, to correct preliminary incorrect chain connectivity, and to identify unexpected chemical features in the molecule. Incorrect crystal structures are almost never reported from high-resolution data. Most of the mistakes in protein crystallography have been made because a medium-resolution structure has been misinterpreted or overinterpreted. In NMR, the general rule is the more internuclear distances measured the better. Most of the mistakes that have been made in NMR structure determination have resulted from either incorrect assignment of a set of resonances to a particular part of a protein or the failure to measure enough internuclear distances.

Different conventions for representing the structures of proteins are useful for different purposes

Atomic coordinate sets make for boring reading, so protein structures are presented visually. There are a number of different ways to render a protein structure, depending on the information that one wishes to convey. The fold of the polypeptide chain can be depicted as a wire model that follows the path of the backbone (Figure 5-5a), which is useful for example in comparisons of two conformations of the same molecule (see Figure 1-80); or it can be depicted as a ribbon diagram in which alpha helices and beta sheets are graphically stylized (Figure 5-5b): this not only makes the overall fold easily recognizable, but makes particular secondary structure elements or loops that may have particular functional significance easily recognizable. Detail of the structure at the atomic level can be rendered by means of a ball-and-stick model (Figure 5-5c) in which the balls are colored or sized by type of atom and covalent bonds are represented by perspective sticks; such drawings are to scale so relative bonded and non-bonded distances can be assessed, which is important for evaluating interactions. Atoms as volumes can be represented by space-filling drawings in which each atom is given a sphere scaled to its van der Waals radius (Figure 5-5d): this representation is particularly useful for assessing the fit of a ligand to a binding site (see Figure 2-8b). Finally, to emphasize the protein surface that is created by the space-filling nature of atoms, a surface topography image can be produced (Figure 5-5e). Such an image can be colored according to different local properties such as the electrostatic potential at different points in the molecule. In this book, all of these different methods of visualizing structures are used.

References

Holyoak, T. *et al.*: **The 2.4 Å crystal structure of the processing protease Kex2 in complex with an Ala–Lys–Arg boronic acid inhibitor.** *Biochemistry*, in the press.

Martz, E.: **Protein Explorer: easy yet powerful macromolecular visualization.** *Trends Biochem. Sci.* 2002, **27**:107–109.

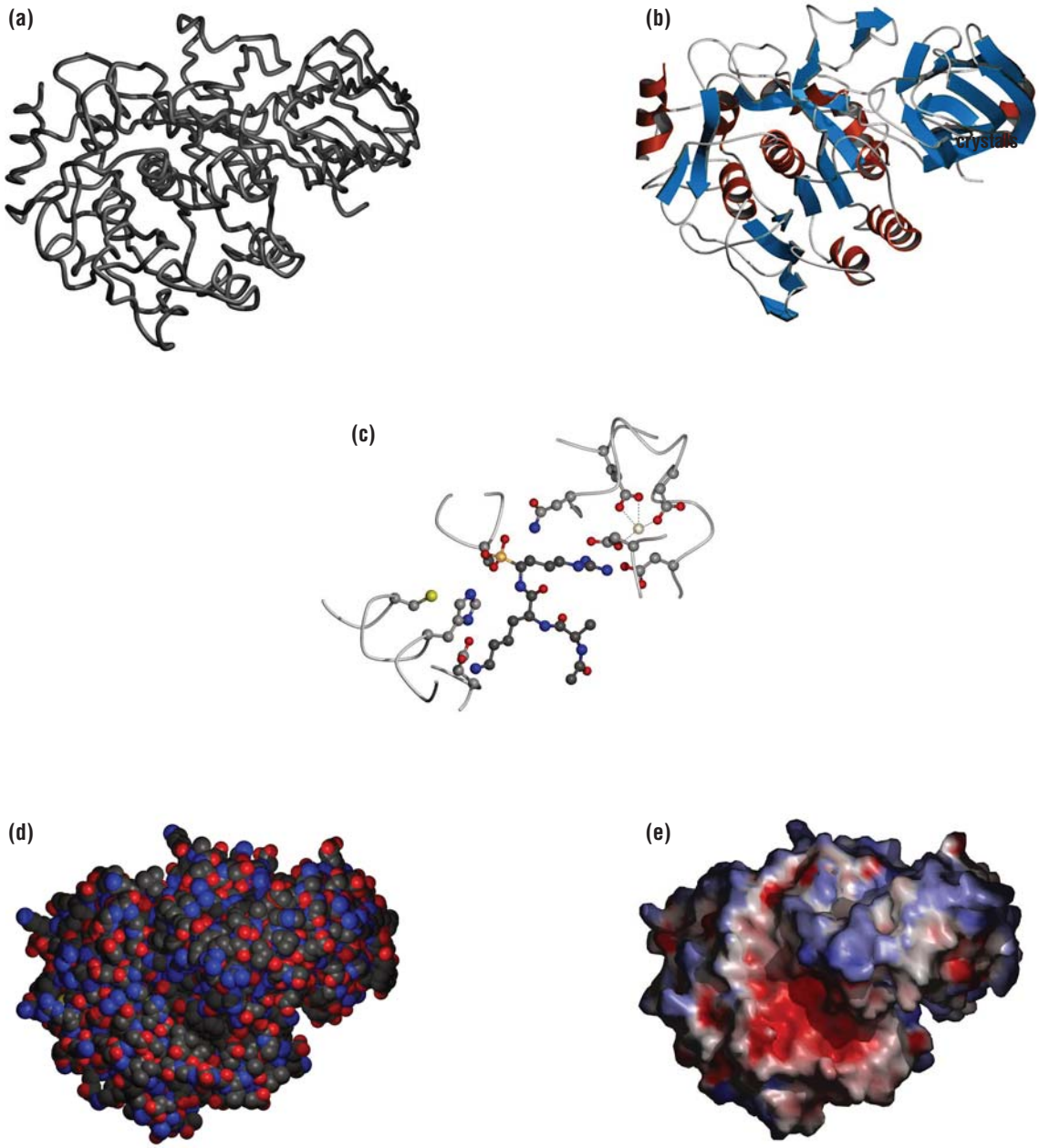

Figure 5-5 Different ways of presenting a protein structure The structure of the catalytic and P domains of the protease Kex2 is shown in different styles. **(a)** "Wire" diagram showing the path of the polypeptide backbone. **(b)** Ribbon diagram highlighting the secondary structure elements. Beta strands are depicted as arrows with the arrow head being at the carboxyl terminus. Alpha helices are drawn as coiled ribbons. **(c)** Ball-and-stick model of a small part of the protein structure, showing details of amino-acid interactions. **(d)** Space-filling representation in which every non-hydrogen atom is shown as a sphere of its van der Waals radius. **(e)** Surface representation (sometimes called a GRASP image after the program that computes it) in which the topography of the protein surface is shown and the electrostatic characteristics of the surface are highlighted in color (red for negative, blue for positive). Kindly provided by Todd Holyoak.