The Classification And Origins Of Protein Folding Patterns

Article *in* Annual Review of Biochemistry · February 1990

DOI: 10.1146/annurev.bi.59.070190.005043 · Source: PubMed

CITATIONS
336

READS
264

2 authors, including:

Alexei V Finkelstein
Russian Academy of Sciences
283 PUBLICATIONS **8,953** CITATIONS

SEE PROFILE

Annu. Rev. Biochem. 1990. 59:1007–39

THE CLASSIFICATION AND ORIGINS OF PROTEIN FOLDING PATTERNS

Cyrus Chothia

MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 2QH, England

Alexei V. Finkelstein

Institute for Protein Research, Academy of Sciences of the USSR, Pushchino, Moscow Region, USSR

KEY WORDS: secondary structure assemblies, close packing, chain topology, sequence determinants of folds.

CONTENTS

0066-4154/90/0701-1007$02.00

PERSPECTIVES AND SUMMARY

In nearly all proteins, the local folding of the polypeptide chain leads to the formation of α-helices or β-sheets, and these assemble to give the molecules their globular three-dimensional structures. The fold of a protein describes in outline three major aspects of its three-dimensional structure: the secondary structures of which it is composed, their relative arrangement, and the path taken through the structure by the polypeptide chain. In proteins with quite different functions and with no detectable evolutionary relationships, these features of their structures can be very similar. Such proteins share the same folding pattern. Though structures of proteins are complex and irregular at the atomic level, their folding patterns are surprisingly simple and elegant.

A small number of folding patterns describe in outline most of the known protein structures. This fact implies that folding patterns arise from the intrinsic general properties of the polypeptide chains and of the secondary structures that form globular proteins. Their precise description and the understanding of their origins will contribute to the solution of a central problem in molecular biology: the relation between the amino acid sequence and three-dimensional structure of proteins. Here we describe the known protein folding patterns and discuss recent work on the factors that determine their structure.

Folding patterns describe quite accurately the outline structures of nearly all small and medium sized proteins. Large proteins are composed of substructures or domains whose outline structures are also well described by folding patterns. The contacts between domains are usually small and irregular.

We begin with a review of the general principles that govern folding patterns. These concern both the general arrangement of secondary structures and the pathways taken by polypeptide chains through the folded structures.

We then describe the particular folding patterns that have been found in proteins composed of (*a*) α-helices, (*b*) β-sheets, and (*c*) both α-helices and β-sheets. We discuss the packing, overall architecture, and chain topology in proteins formed by these secondary structures.

The last part of the review is a discussion of the features of protein sequences that underlie the structure and stability of particular folds and the general implications of this work for folding patterns.

Work on protein folding patterns published prior to 1984 was reviewed in a previous article in the *Annual Review of Biochemistry* (1), and the present review is essentially concerned with work published since that time. To provide a coherent account of the recent work, however, brief sketches of some of the earlier results are given here. Two related reviews have been published recently (2, 3). One contains a detailed discussion of how the common folding patterns are produced by the requirements of (*a*) favorable topologies for the polypeptide chain, (*b*) a compact structure for the folded protein, and (*c*) the random distribution in sequences of polar and nonpolar residues (2). The other includes a description of the roles by the individual amino acids in protein structures (3). A comprehensive account of the general features of protein structure is given in (4); a detailed discussion of the chain topologies in proteins is given in (5); and clear simple descriptions of the folds in proteins of known structure are given in (6).

GENERAL PRINCIPLES FOR PROTEIN FOLDING PATTERNS

Protein folding patterns describe both the arrangements of the secondary structures and the paths made through the folded structures by polypeptide chains—the chain topologies. Protein folds must also be sufficiently compact to provide stability to the structure.

Secondary Structures and their Packings

In proteins the requirement that a structure be compact and that hydrogen bonds be formed by buried polar groups necessitates the formation of α-helices or β-sheets by a large proportion of the polypeptide chain. The only clear exceptions are found in a few small proteins, or in small domains in large proteins, where few residues are totally buried and main chain polar groups can form hydrogen bonds to the solvent. The α-helices or strands of β-sheet run across the molecule. The loops connecting secondary structures form few intramolecular hydrogen bonds and so they nearly always occur at the surface. The few that are buried have an environment structured to include water molecules to which they hydrogen bond (7).

The general shape of secondary structures governs the ways in which they form compact folds (2, 8). β-sheets form layer structures with helices or other β-sheets packed on their faces. The cylindrical shape of α-helices allows them either to stack around a central core or to form layer structures.

Protein interiors are close packed. This means that the exact geometry of

packed secondary structures will depend upon the residues that form the interface between them. Inspection of protein structures shows, however, that there are preferences and regularities in the geometries of their packings. In many cases these geometrical packing patterns can be explained in outline by simple models that embody the average features of the surfaces and conformations of secondary structures (9).

Chain Topology

Inspection of protein structures shows that the observed pathways are subject to definite limitations and preferences that can be summarized as follows:

1. Pieces of secondary structure that are adjacent in the sequence are often in contact in three dimensions and usually pack in an antiparallel, rather than parallel, manner (Figure 1*a*).
2. The connections in β-X-β units (where the βs are parallel strands in the same sheet, though not necessarily adjacent, and X is an α helix, a strand in a different sheet, or an extended piece of polypeptide) are right handed (Figure 1*b,c*).
3. The connections between secondary structures neither cross each other nor make knots in the chain.

The various analyses of protein structures that have led to the three rules have been reviewed previously (1–6). For certain protein folds, combinations of these basic rules give rise to further topology rules. These are discussed in later sections of this review.

The initial explanations for these empirical rules suggested that they might arise from structural features of the folded state or structural and kinetic features of intermediates in folding pathway (5, 6, 8, 10–14). Recently, it has been shown that a completely general explanation is provided by the inflexibility of the polypeptide chains (2).

The flexibility of a polymer chain is described by the persistence length model (15, 16). From this model an estimate of the free energy required to bend a chain is given by

$$\Delta G = \frac{RTa}{2L}(\Delta\theta)^2$$

where R is the gas constant; T the temperature; a the persistence length, about 17 Å for proteins; L the chain length, equal to $3.5m$ Å for m residues; and $\Delta\theta$ is the bending angle. At room temperature

$$\Delta G = \frac{1.5}{m}(\Delta\theta)^2$$

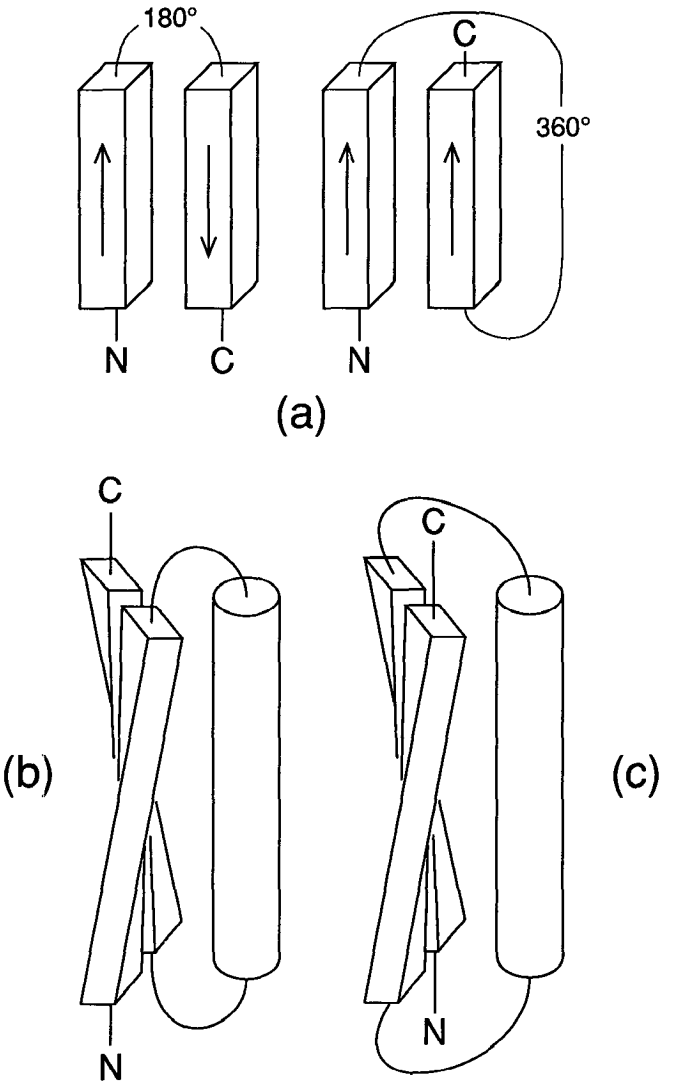

Figure 1 Features of the chain topology in protein structures.

(*a*) Pieces of secondary structure adjacent in the sequence tend to be adjacent in the structure and to pack in an antiparallel manner. Antiparallel packing involves the connecting loop bending through about 180°, while parallel packing involves bending about 360°.

(*b* and *c*) The connections in β-X-β units (where the βs are parallel strands in the same sheet, though not necessarily adjacent, and X is an α helix, a strand in a different sheet, or an extended piece of polypeptide) are right handed. Here the helix is shown as a cylinder and the strands as thick ribbons. Because of the right-handed twist found in β-sheet, right-handed connections between helices and strands (*b*) involve loops that bend some 25° less than left-handed connections (*c*).

This expression shows that the rigidity of polypeptide chains imposes restrictions on the favorable paths of loops (2). For a loop of 10 residues, the bending free energy is about 3.5 kcal/mole for a bending angle of 180° and about 6 kcal/mol for a bend of 360°. Thus antiparallel associations in which the bend angle is 180° are favored over parallel associations where the bend is about 360° (Figure 1*a*). This effect is more pronounced for loops attached to β-strands, which they leave in a direction along the strand axis, than for loops attached to α-helices, which they leave in a direction perpendicular to the helix axis.

The rigidity of polypeptide chains also favors right-handed over left-handed connections in β-X-β units (2). The right-handed twist normally found in β-sheets means that the bending angle for β-X-β units with right-handed connections is less than that for left-handed connections (Figure 1*b,c*). This effect is cooperative, because, if connections are not to cross each other (rule 3), all, or nearly all, connections must have the same hand.

The crossing of loops is unfavorable because one loop would be inside the structure, with the polar atoms in the buried loop losing contact with the solvent and having difficulty in making alternative hydrogen bonds.

If a protein is formed by two layers of secondary structures, the formation of knots in the polypeptide chain involves loop crossings. But in structures made of three or more layers, crossings are not obligatory for the formation of knots. The chance of their formation is still very low, however. Calculations show that for chains of 100–300 residues, the probability of knotting is 1–10% if the chains are extremely thin, and it is much less if the chains have the width of polypeptides or secondary structures (17–21).

These arguments suggest that the origin of the regularities and preferences found for the topology of polypeptide chains in protein structures is mainly due to the relative entropic cost of different folding intermediates. Such intermediates usually have only marginal stabilities, but we might expect rare occasions when the larger entropy losses are compensated for by additional interactions within intermediates. Thus the rules for the topology of polypeptide chains are statistical in nature. Indeed, a few exceptions, left-handed β-X-β units and chain threading (which resembles unfinished knotting), have been observed (1, 6).

Stability and Accessible Surface Area of Protein Folds

The free energies of protein folds are small, usually just a few kcals (22). These values represent small differences in the large terms favoring the unfolded or folded state, particularly conformational entropy of the polypeptide chain and hydrophobic free energy. Conformational entropy favors the unfolded state and depends upon the length of the polypeptide chain, while hydrophobicity depends upon contacts with the solvent and favors the folded state (23).

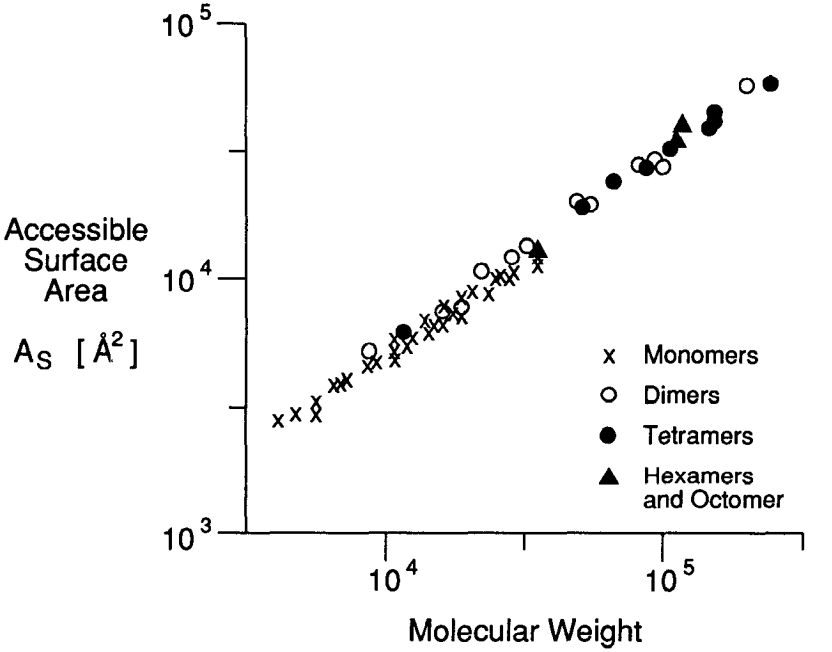

Figure 2 The correlation for monomeric and oligomeric proteins of the surface area that is accessible to solvent (90) and molecular weight. [Adapted from (91).]

The surface of unfolded proteins is proportional to molecular weight. The balance between the favorable and unfavorable energy terms implies that the buried and accessible surface area of a folded protein should be correlated with molecular weight. The correlation for monomeric proteins was established some time ago (1). Recently it has also been established for oligomeric proteins (24) (Figure 2). Although the size of total buried surface is essentially constant for oligomeric proteins of the same molecular weight, the separate contributions of the surfaces buried within and between the subunits vary greatly (24).

FOLDING PATTERNS IN PROTEINS FORMED BY HELICES

The Packing of α-Helices

The relative orientation of two α-helices packed face to face can be described by the angle between the helix axes when projected onto their plane of contact. The values for this angle in observed helix packings cover the whole possible range, though there is a major peak in the distribution at $-50°$ and a minor peak at $+20°$ (Figure 3). The preference for these orientations arises from the general characteristics of helix surfaces (1, 25, 26). Residues on the

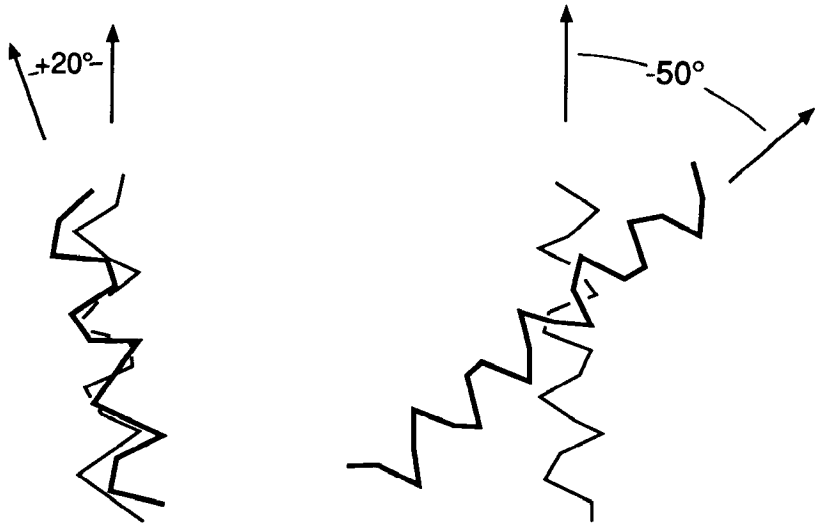

Figure 3 The packing of α-helices. The two preferred orientations of packed helices are those with the helix axes inclined at −50° and +20° (1, 25, 26).

surface of an α-helix tend to form ridges separated by grooves. The packing together of the ridges and grooves on the surface of two ideal helices will incline their axes at an angle of −50° or +20°. Irregularities in the surface ridges give packings that depart from these ideal orientations.

Assemblies of α-Helices

The common occurrence of one type of assembly of helices, that built from four helices packed with their axes antiparallel, was noticed some time ago (27), and the chain topologies and the packing geometries found for this fold have been described in some detail (28, 29). Larger assemblies of helices, however, are more complicated and usually consist of a compact aggregate of helices around a central core. Recently, the general principles that govern how α-helices form the three-dimensional structure of globular proteins have been embodied in a simple geometrical model. This model, the "quasi-spherical polyhedron model" (30, 31), describes accurately the outline structure of the large majority of the observed assemblies of α-helices (31).

The basis of the polyhedron model was the realization that the geometries of α-helices packed around a central core can be described by polyhedra; an example of such a description is illustrated in Figure 4*a*. The model for helix packings was developed by first determining the polyhedra that describe the ideal packing of 3, 4, 5, and 6 ideal helices (30). The helix assemblies actually observed in proteins were then shown to have geometries close to those described by the ideal models (31).

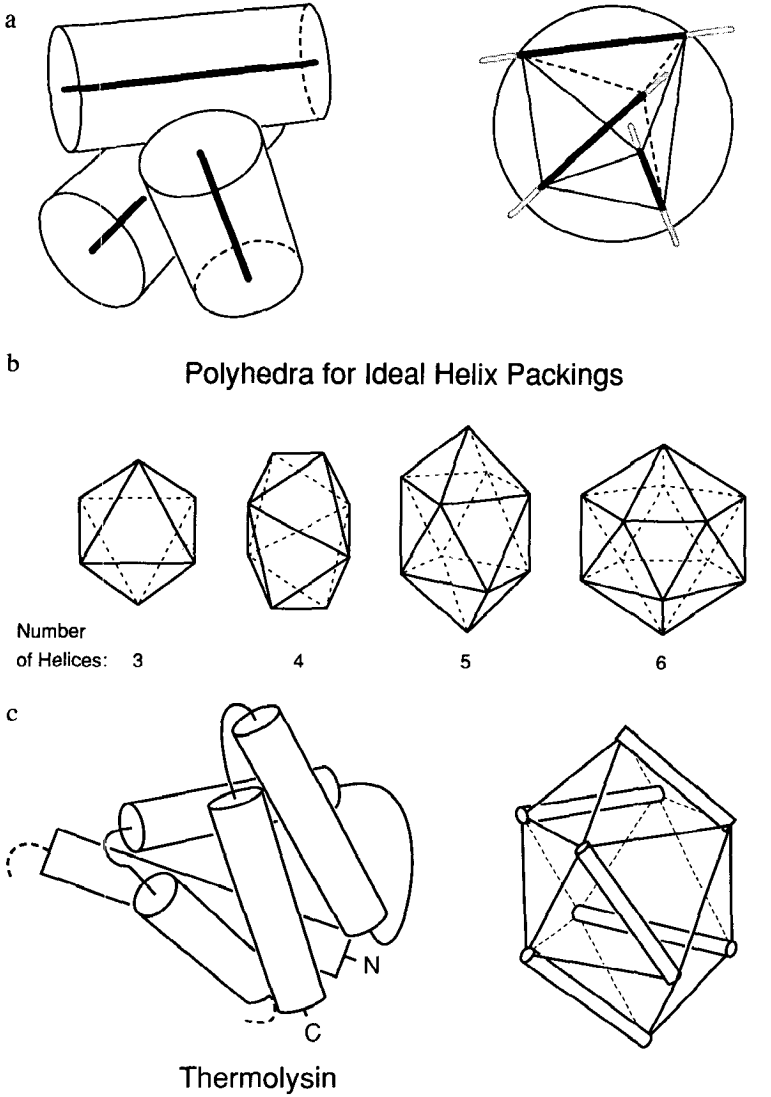

Figure 4 The polyhedron model for the architecture of assemblies of α-helices (30, 31). (*a*) This figure illustrates how the geometry of helix packings can be described by a polyhedron. Three packed helices are shown as cylinders. To construct this polyhedron, a sphere of diameter 11 Å is first drawn from the center of the packing. This sphere encloses the central section of each helix axis. These axes form one set of the ribs of the polyhedron. The polyhedron is completed by another set of ribs formed by the connections linking the ends of the nearest helices. For the three-helix packing, the polyhedron constructed by this procedure is the octahedron. Using this construction the geometry of any helix packing can be described by a polyhedron. (*b*) The polyhedra that describe the ideal packing of three, four, five, and six helices are the octahedron, the dodecahedron, the hexadecahedron, and the icosahedron, respectively. (*c*) The geometry of the five-helix packing in the second domain of thermolysin (92) corresponds closely to one of the ideal hexadecahedron models (31). [Reproduced from (93) with permission.]

The models for ideal helix assemblies embodied five of the features actually found in globular proteins:

1. helices have the general shape of cylinders;
2. they pack together with one face on the interior of the protein and the other exposed to the solvent;
3. the helices pack around a central hydrophobic core whose diameter is the length of two residues: ~11 Å;
4. the helices are close packed with a similar number of contacts, and
5. the assemblies of helices must be as spherical as possible.

For a given number of helices, there is only one polyhedron that can satisfy these five conditions (30). Ideal packings of three helices are described by an octahedron, four-helix packings by a dodecahedron, five-helix packings by a hexadecahedron, and six-helix packings by an icosahedron (see Figure 4*b*). For more than six helices, there are no such polyhedra.

Helices can be placed on the ribs of these ideal polyhedra in a number of different ways. Three helices fit on the octahedron in two different arrangements: that seen in Figure 4*a* and its enantiomorph. Four helices can be placed on the dodecahedron, and five helices on the hexadecahedron, in 10 different ways. Six helices can be placed on the icosahedron in eight different ways.

To determine whether these ideal models are relevant to real proteins, the geometries of the helix arrangements observed in real protein structures were compared with the geometries expected from the ideal models (31). In almost all cases, there is a close fit between the helix packings found in these proteins and one of those in the ideal models. The positions and orientations of helices in the observed structures differ from those found for helices in the corresponding ideal model by 2 Å and 20° on average (Figure 3). The larger deviations occur in proteins where the spherical nature of the packing is distorted by the helices clustering not about a point but about a somewhat more elongated or flattened region (31).

As discussed above, the ridge and groove character of helix surfaces means that pairs of close packed helices usually have their axes tilted at angles of approximately $-50°$ or $+20°$. The different polyhedron models put the helices in different relative orientations. The models that correspond to observed structures are, in most cases, those in which pairs of neighboring helices are inclined at angles of approximately $-50°$ or $+20°$ (31).

The absence of ideal polyhedra for assemblies of more than six α-helices implies that, in the few proteins that contain seven or more in one domain, the assembly will be of a different kind. The alternative is the formation of layer structures by the α-helices (8, 32, 33), as occurs in proteins containing β-sheets. This is seen in bacteriodopsin where seven helices form two layers (34), and in colicin A, where 10 helices pack in three layers (35).

Chain Topology in Helical Proteins

The polyhedron model provides an easy way to describe chain topology in assemblies of helices. In nearly all known structures, the connections between helices occur along the ribs of the polyhedra that connect adjacent helices (31). Only two exceptions are currently known. Thus the chain path is formed by the α-helices and connecting loops passing through all the vertices of the polyhedron without self-intersection.

FOLDING PATTERNS IN PROTEINS FORMED BY β-SHEETS

The Packing of β-Sheets

The intrinsic flexibility of β-sheets, which allows them to twist, coil, and bend (local coiling), also allows them to pack in quite different ways. Three distinct classes of β-sheet packings have been described in some detail. One of these is found in α/β barrel structures and is discussed in a later section. The other two are the aligned and the orthogonal β-sheet classes; the principles that govern these two types of association have been embodied in simple models (1, 36–38). Both involve the face-to-face packing of the β-sheets to form layer structures.

Aligned packing (1, 36, 37) involves the association of β-sheets that are essentially independent: strands in different sheets are linked by loops of polypeptide. The right-handed twist of β-sheets means that their surfaces are also twisted. Two ideal twisted β-sheets close pack if the rows of side chains that form the contacts at the interface are aligned. Although the side chains at the interface are aligned, the twist of the sheets results in the main chain direction in the two β-packed sheets being at an angle of about $-30°$ (Figure 5*a*).

Orthogonal packing (1, 38) involves β-sheets folding upon themselves to form layer structures in which the strand directions in the two layers are inclined at about 90° (Figure 5*b*). β-sheet strand(s) at one corner, or two diagonally opposite corners, pass from one layer to the next without interruption. The twist of the sheets makes the other two corners splay apart. Active sites or ligand-binding sites are often found at these splayed corners.

A different class of β-sheet packings is found in soybean trypsin inhibitor and interleukin-1β (39, 40). In this class three small β-sheets pack around a core. No analysis of how the β-sheets pack to form this folding pattern is available at present.

Assemblies of β-Sheets

The large majority of the known assemblies of β-sheets are two-layer structures in which the sheets associate in a manner very similar to that described

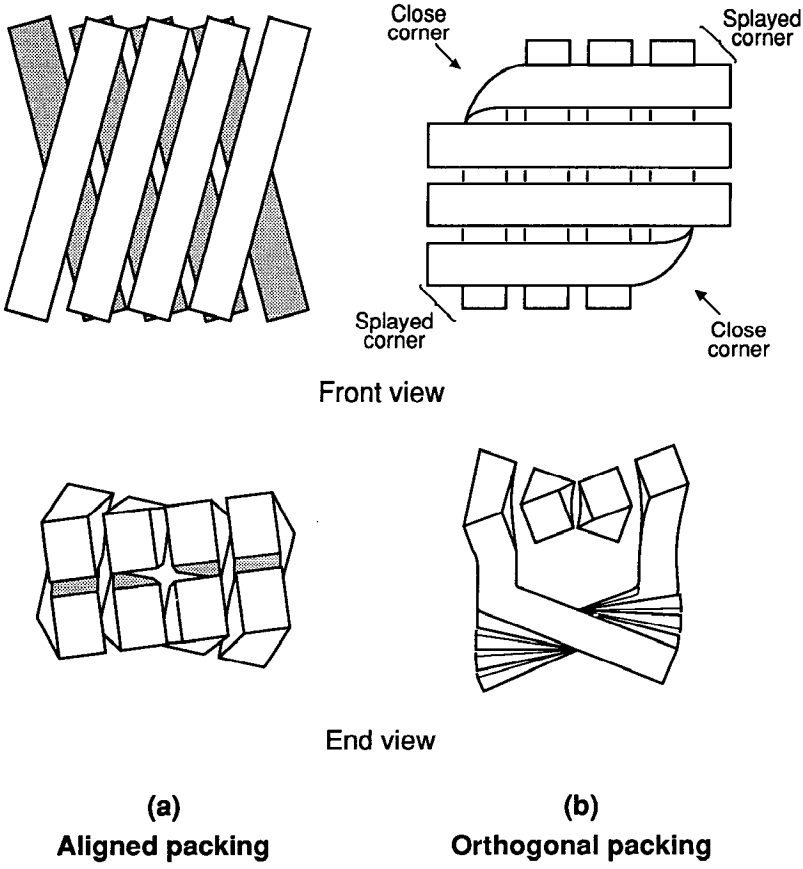

Figure 5 The packing of β-sheets (1, 36–38). In the drawings shown here, the strands of β-sheet viewed face on are shown as ribbons or, when viewed from one end, as rectangles. (*a*) The arrangement of the β-sheets in the aligned packings. (*b*) The arrangement of the β-sheets in the orthogonal packings.

by the models for aligned or orthogonal packing (36–38). They differ in the sizes of the β-sheets and the topologies of the chains. In one domain of the aspartate proteases, the β-sheet folds twice upon itself so two orthogonal packings form a three-layer structure (38).

A more complex type of assembly is found in the enzyme neuramidase (40). This protein has six β-sheets arranged in a propellerlike assembly. The β-sheets in this structure pack in the manner described by the aligned model. The propellerlike assembly is formed by the relative displacements of the packed sheets and their right-handed twist (Figure 6).

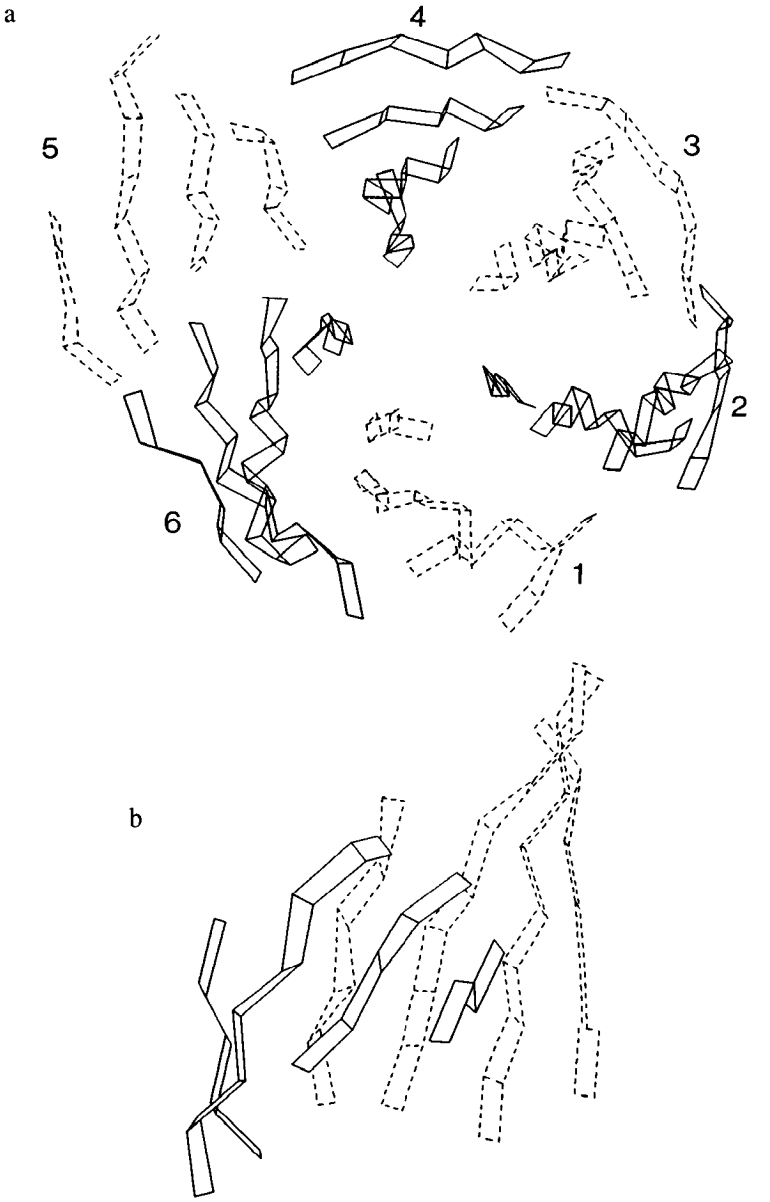

Figure 6 The architecture of the β-sheet assembly in neuramidase (41). (*a*) The strands that form the six twisted β-sheets in this protein are drawn as ribbons. The β-sheets are numbered 1 to 6 and drawn with alternating broken and continuous lines so they may be distinguished. (*b*) The relative positions of two adjacent sheets viewed face on. The packing of the sheets is described as the aligned model (Figure 5*a*). The propellerlike assembly is created by the combined effects of the relative displacements of the β-sheets and (*b*) their right-handed twist (*a* and *b*). (Figure drawn by A. M. Lesk from atomic coordinates supplied by P. Colman.)

Chain Topology in β-sheet Structures

The operation of the general rules of chain topology on the layer structures formed by β-sheets results in the common occurrence of a limited number of topologies. The first detailed description of these topologies (13) was by the use of diagrams in which the sheet structures were drawn in one plane with strands that are adjacent in structure adjacent in the diagram. In such diagrams of the known structures, certain patterns or motifs are very common. These have been given the names "hairpin," "greek key," and "jellyroll," and concern two, four, and six strands, respectively (6, 13) (Figure 7).

For strands that are on the edge of a sheet, these diagrams do not differentiate between the neighbor in the same sheet and that in the sheet against which it packs. Thus the greek key motif, for example, applies to strand arrangements that differ in three dimensions (5). In Figure 7 we show the different structures that have been found for the motifs in β-sheet proteins. These are taken from a systematic analysis of the topologies found in the known β-sheet structures (42, see also 43, 43a). An examination of the structures in the Figure 7 shows how the motifs arise from the combination of the three basic rules for chain topology discussed above in the second section.

FOLDING PATTERNS IN α/β PROTEINS

Proteins whose structures are composed of α-helices and strands of β-sheets that roughly alternate along the chain can be divided into two groups. In the first, the β-sheet is just twisted and the helices pack upon its faces, usually both faces, though occasionally on just one. In the second, the sheet is twisted and coiled upon itself to form a barrel structure with helices packed on its outside surface. The second group, the α/β barrel proteins, is discussed in the next section of this review.

The Packing of α-Helices on β-Sheets

β-sheets in proteins have a right-handed twist when viewed in a direction parallel to the strands. In an α-helix, the two adjacent rows of residues, i, $i+4$, $i+8$, . . . and $i+1$, $i+5$, $i+9$, . . . also form a surface with a right-handed twist. Ideal α-helices pack onto ideal β-sheets with their axes parallel to the strands, because in this orientation the two rows of helix residues form a surface complementary to that of the β-sheet (Figure 8*a*). The large majority of observed packings have geometries close to that given by this model; very occasionally large departures from this geometry are produced by large or small side chains at the interface (1, 44, 45).

Assemblies of α-Helices and β-Sheets

The shape of the β-sheets means that most assemblies of α helices and β-sheets form simple layer structures. Some small α/β proteins contain just

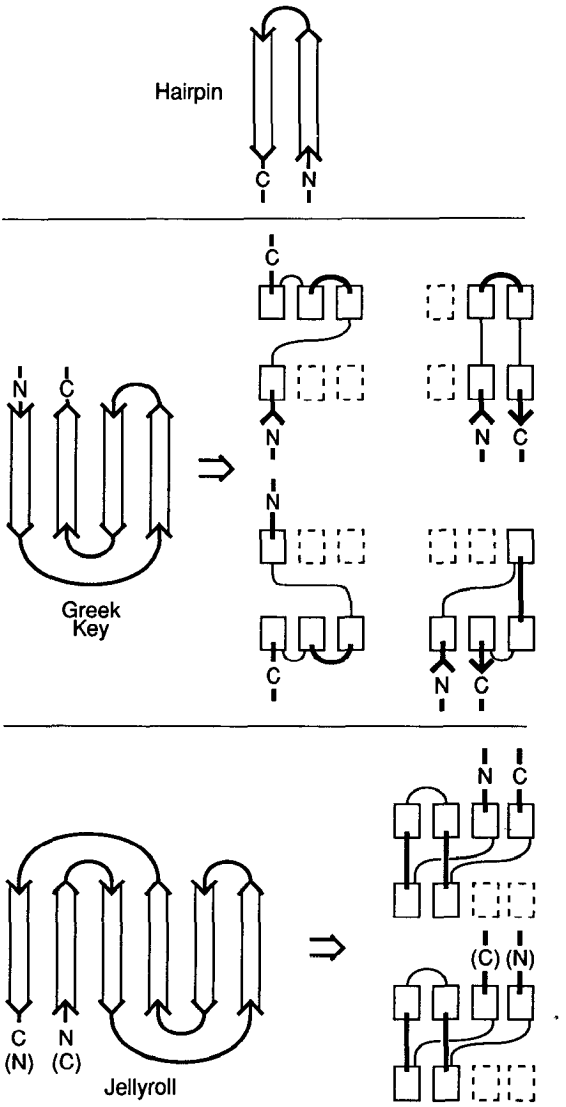

Figure 7 The chain topology in β-sheets proteins (6, 42, 43). On the left, the hairpin, greek key, and jellyroll motifs for the topology of β-sheets (6) are shown; see text. The layer structure of β-sheet packings (Figure 5) means that the motifs describe different spatial arrangements of β-sheet strands (42, 43). The spatial arrangements found in the currently known β-sheet proteins for strands that fit the greek key and jellyroll motifs (42) are shown on the left. The view is down the strands, and connections between strands close to the reader are shown by thick lines and those distant by thin lines. The sheets that form the layer structures are horizontal.

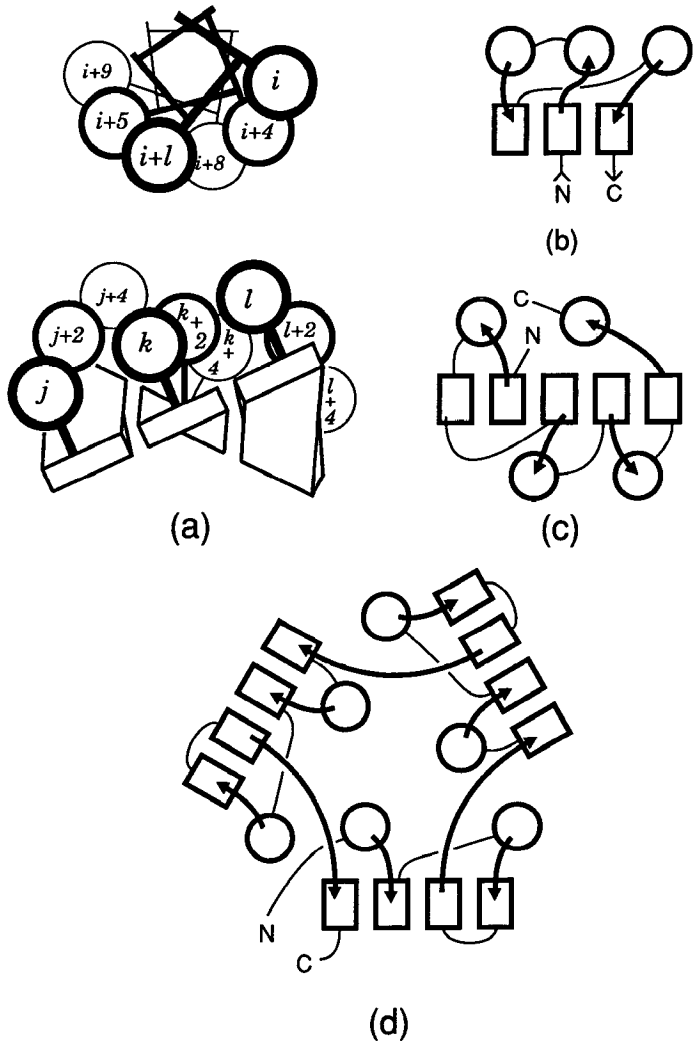

Figure 8 The packing and assemblies of secondary structures in α/β proteins (44–46, 94, 95) (*a*) Schematic pictures of an ideal three-stranded β-sheet and an α-helix are shown in which the side chains on one face of each are shown as large open circles. Because of the twist of the β-sheet, the side chains j, $j+4$, $j+6$, k, $k+2$, $k+4$, l, $l+2$, and $l+4$ form a surface that has a right-handed twist. On the α-helix the surface formed by the helix residues i, $i+1$, $i+4$, $i+5$, $i+8$, and $i+9$ also has a right-handed twist. Therefore, these ideal secondary structures close pack when the helix axis is parallel to the strand direction, because in this orientation the two surfaces are complementary. (*b,c*) Helices and β-sheets nearly always pack to form layer structures. These are ilustrated here in schematic diagrams in which the structures are viewed end on with helices represented as circles and the strands as rectangles (8). Occasionally α-helices and β-sheets form two-layer structures, as in (b) the L7/12 protein (94). More commonly, they form three-layer structures with α-helices packed on both faces of a β-sheet (*c*), as in Flavodoxin (95). (*d*) The type of assembly found in the enzyme EPSP synthase (46) is unique to that protein at present.

two layers, α-helices packed on one side of a β-sheet. The large majority are three-layer structures with α-helices packed on both sides of a β-sheet (Figure 8*b,c*) (6, 8, 44, 45).

Recently, a quite different architecture has been found in the enzyme EPSP synthase (46). This protein has two domains that have homologous structures. Each domain contains three copies of a simple α/β substructure formed by a four-stranded β-sheet and two α-helices. The three substructures pack symmetrically to form the domain. At the center, three helices pack approximately parallel and are surrounded by a triangular array of the three β-sheets (Figure 8*d*).

Chain Topology in α/β Proteins

The chain topology of α/β proteins is dominated by the preference of β-α-β units for right-handed topology (Figure 1). This preference is particularly striking in the way it determines the strand order in the β-sheets. To give two examples: if α-helices are packed on both sides of a parallel β-sheet, and if all the β-α-β units are right handed, it is necessary for the first strand to be near the middle of the β-sheet. Alternatively, the first strand of the β-sheet can start at one side, but then it must contain both parallel and antiparallel strands (Figure 8*b–d*) (5, 6, 47).

FOLDING PATTERNS IN α/β BARREL PROTEINS

The α/β barrel is the name given to the protein fold where a β-sheet coils round to form a closed cylinder and α-helices are packed on the surface of the cylinder. The first example of this fold was seen in 1975 in the enzyme triose phosphate isomerase (48). Recently, the same fold has been found in some 15 other proteins (49–51).

An example of α/β barrel structure is illustrated in Figure 9*a*. The known examples of this fold have remarkably similar features. The barrel structures are formed by eight parallel strands of β-sheet with the same geometry (see below). The strands are almost always linked by α-helices packed onto the β-sheet with their axes approximately parallel to the strand direction. The α-helix–β-sheet packing is the same as that previously described in α/β proteins. Discussions of the α/β barrel fold have concentrated on its unique feature: the barrel structures formed by the β-sheets.

The Formation of β-Barrels by β-Sheets

A general model for the formation of barrels by β-sheets was described some time ago (52), and recently it has been demonstrated that this model accurately describes the main geometrical features found in the known α/β-barrel

a

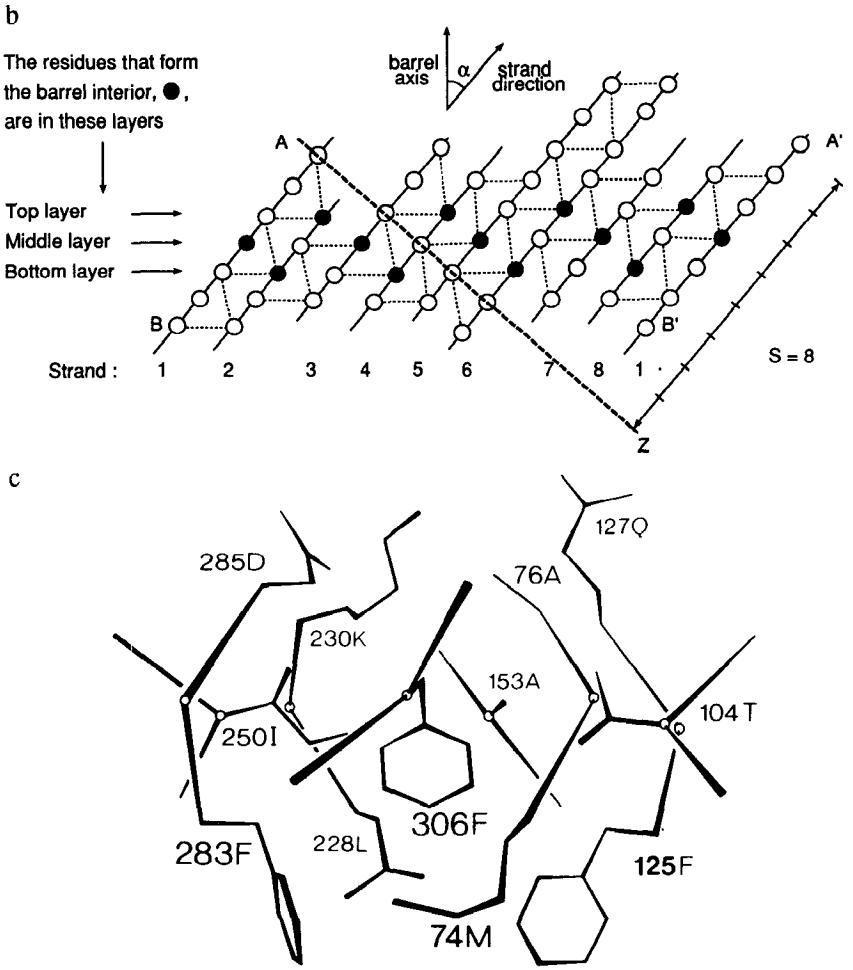

Figure 9 The geometry and packing of the β-sheets in α/β-barrel proteins (*a*) (See opposite page.) Two orthogonal views of the α/β barrel proteins Triose Phosphate Isomerase: views looking onto the sheet face *(left)* and down the barrel axis *(right)*. The helices are shown as cylinders and the sheet strands as ribbons. (*b*) The β-sheet of an α/β barrel. Observed structures contain eight strands. Here nine strands are shown, as the first strand is drawn on both the left and right edges of the sheets. To make the eight-stranded barrel structure, the leftmost strand must be superposed onto the rightmost by curling the paper back and glueing of residues A and B to residues A' and B'. Superposing residues A and B onto A' and B' creates a barrel with strands tilted with respect to its axis, by an angle α. The shear number S is a measure of this stagger of the β-sheet strands. If a line drawn perpendicular to the strand direction goes round the barrel and comes back to the strand from which it started, S is the number of residues by which it is displaced from its starting point. For the sheet illustrated here, it is the displacement of Z from A': 8. The 12 residues that pack at the center are shown as filled circles. (*c*) Residue packing at the center of the β-barrel in glycolate oxidase (50). The strands of β-sheet are pruned to three Cα per strand that form the central region. Twelve side chains fill the center of the barrel. They pack in three layers: 74 Met, 125 Phe, 228 Leu, and 283 Phe at the bottom; 104 Thr, 153 Ala, 250 Ile, and 306 Phe in the middle; and 76 Ala, 127 Gln, 230 Lys, and 285 Asp at the top. [Reprinted from (50) with permission.]

structures (50). A more detailed model shows that in some structures the barrel shape is distorted to form a hyperboloid (49).

The geometries of barrel structures formed by β-sheets can be defined by the following terms (52):

1. those intrinsic to β-sheets;
   - *a*, the Cα-Cα distance along the strands;
   - *b*, the distance between neighboring strands, and
   - τ, the twist of the β-sheet, defined by the number of turns by which the plane of the β-sheet twists on moving from one residue to the next along a strand;

and 2. features particular to a barrel structure;
   - *n*, the number of strands that form the barrel;
   - α, the angle between the direction of the strands and the axis of the barrel (see Figure 9*b*)
   - *R*, the mean radius of the barrel, and
   - *S*, the shear of the sheet. This is a measure of the extent to which the strands of the β-sheet are staggerred, and the definition of *S* values is shown in Figure 9*b*.

It can be shown (52) that the relationships among these quantities are given by the equations:

$$S = nb\tan\alpha/a$$
$$R = b/[2\sin(\pi/n)\cos\alpha]$$

and

$$\tau = a\sin\alpha\cos\alpha/2\pi R$$

Because the geometry of the sheet can vary only within fairly narrow limits, the geometries of β-sheet barrels are determined in practice by the number of the strands and the shear of the β-sheet. In all the known α/β barrel structures, the β-sheets have eight strands and the shear number (*S*) is eight. For such β-sheets, the equations predict that (*a*) the radius of the barrel should be 14 Å, (*b*) the tilt of the strands to the barrel axis should be $-36°$ (the negative sign of the tilt arises from the right handed twist of the β-sheets) and the twist angle between successive strands is 26°. The observed values are the same as those given by these predictions (50).

The one significant departure from this simple model that is found in the observed structures is that in some proteins the barrel cross section is ellipsoidal rather than circular (49).

The Packing of Residues Inside β-Barrels

There are no reasons intrinsic to the sheet geometry in β-barrels why α/β barrels should have only structures made with sheets with eight strands and a

shear number of eight. This preference arises from the packing of residues in the interior of the barrel (50).

Figure 9*b* shows the hydrogen-bonding net that arises by "unrolling" the barrel. In the observed structures, the strands vary in length, but at the central region of the sheet all have three residues at the same height relative to the barrel axis, which form a continuous hydrogen bonded net around the barrel. On each strand of sheet, alternate side chains point toward the region inside the sheet and out toward the helices. The packing inside the barrel is formed by the interactions of the 12 residues that have their side chains pointing inward (Figure 9*c*). One residue is contributed by each of the first, third, fifth, and seventh strands and two residues by each of the second, fourth, sixth, and eighth strands (Figure 9*b*).

The side chains occupy three tiers or layers. Each layer contains side chains from four alternate strands (Figure 9*c*). The layering is a consequence of the tilting of the strands with respect to the axis. The tilt angle of $-36°$ that produces the layering is a consequence of the sheet geometry; in particular, of the number of strands and the shear.

Closed barrel structures with eight strands with a shear number of eight involve normal sheet geometry and twist. They also place at the center of the barrels 12 side chains in a regular array and in a volume close to that of the average side chain (50). A 10-stranded barrel with a radius of about 8.9 Å would have an interior too large to be filled by a normal set of residues (49, 50). Similarly, a barrel of six strands, with a radius of about 5.5 Å, would have an interior that is too small. Small barrel structures that are not closed are very occasionally found. These are not subject to the same limitations as closed barrels, and so can have a variety of geometries (49).

Two Classes of α/β-Barrels

Underlying the similarity of folding of α/β-barrels, there are really two distinct classes (50). In the central girdle of three residues, alternate strands contribute one or two side chains, respectively, to the region within the sheet. Thus, if the strands are numbered 1 through 8 from the N-terminus of the sequence, class 1 structures are those for which the N-terminal strand (and the other odd-numbered strands) contribute one residue (Figure 9*b*); and class 2 structures are those for which the N-terminal strand (and the other odd-numbered strands) contribute two residues (50).

Chain Topology in α/β Barrel Proteins

The right-handed topology of β-α-β units and the absence of knots require a sequential arrangement of parallel β-strands with the last ending up adjacent to the first (6, 12, 47). Enolase is different from the other α/β-barrels. It has two strands followed by two α-helices and then a sixfold repeat of the

β-strand-α-helix units (53). This means that the first α-helix is antiparallel to the other α-helices and the second β-strand is antiparallel to the other β-strands.

SEQUENCE DETERMINANTS OF PARTICULAR PROTEIN FOLDS

Although the amino acid sequence determines the three-dimensional structure of a protein, not all sites are of equal importance. At some sites the substitution of very different residues has little effect on structure, while at other sites the same substitutions prevent the formation of a stable structure. However, it is not just the contribution of the individual residues to stability that is important for understanding the relation between sequence and folding patterns. We would expect, for example, the contribution of a buried nonpolar residue of a particular type to be similar in proteins with quite different folds. Also, very different sequences can produce very similar folds. To understand how protein folds are determined we need to understand how they are produced by the collective features of amino acids in sequences.

Homologous and Engineered Protein Sequences

Two related approaches have been used to investigate the relation between the sequences and folds of particular proteins. One is the analysis of naturally occurring families of proteins that have the same basic three-dimensional structure but for which many widely different sequences are known. The second is to use the new recombinant DNA techniques of saturation mutagenesis to produce all mutations at designated sites and to determine which sequence changes are consistent with the retention of structure and function.

Both approaches assume that if a sufficient range of sequences are available, the aspects that determine the fold of a protein should be apparent as conserved features. There are, however, limitations on the range of sequences that are available for a given fold. For naturally occurring proteins, the range is limited by the necessity to retain function as well as structure; by the random character of natural mutations, and by the small free energies of protein folds restricting sequence variations to those given by mutations that involve only small energy changes. Also, the analysis of homologous sequences is complicated by large changes in sequences resulting in peripheral loops and secondary structures having different folds (see below). However, if the atomic structures of the distantly related members of the family are known, the regions that change their fold can be determined.

Experiments with engineered mutations are not subject to the same limitations. But they are limited by it not being possible, in practice, to carry out simultaneous saturation mutageneses at more than a few sites at one time.

This is a serious limitation because sites in proteins are not autonomous but related, so that what is acceptable at one site will depend upon what occurs at other sites.

Analyses have been carried out on the sequences of the globins and of the variable domains of immunoglobulins (54, 55). At present, globin sequences have been determined for more than 400 species, and close to 500 complete sequences and twice as many partial ones are known for the immunoglobulin V domains (56). The globins have about 145 residues that form seven or eight regions of helix. The variable domain sequences have about 110 residues that form two β-sheets.

Saturation mutagenesis studies have been carried out on the arc repressor protein (57) and the N-terminal domain of the λ repressor (58). Both are dimers, with the arc monomer having 53 residues and the λ monomer 102 residues. The λ repressor domain is formed by six α-helices.

In spite of the different constraints and limitations on the results that can be obtained from the analyses of homologous sequences and from studies of engineered mutations, the views that this work give of the relations between the sequence and the fold of a protein are remarkably similar.

Hydrophobic and Hydrophilic Surfaces

In relating sequence and structure, the basic classification of residues is usually in terms of their hydrophobic and hydrophilic character. This classification is based on the initial theoretical proposal that the origin of protein stability is the removal from contact with water of hydrophobic residues (23), and on the experimental observations that the free energies associated with the transfer from water to organic solvent of the polar, neutral, and nonpolar groups of amino acids are correlated with the extent to which they occur in the interiors and the surfaces of proteins (59–62). Several scales for the hydrophobic-hydrophilic character of residues have been published. Most of these scales differ only in details and so show high correlations (60, 63–64a).

The comparison of the first known protein sequences and structures produced the fundamental observation that the sites that form the protein interior are occupied mainly by nonpolar residues and occasionally by neutral residues (65–66a). The polar atoms in neutral residues usually form hydrogen bonds within their own secondary structures, so the surfaces between secondary structures are almost entirely hydrophobic in character (59). More recently, analyses of the much larger body of sequence data now available have shown that the converse is also true: the sites in a protein that are highly exposed to the solvent are nearly always occupied by polar or neutral residues (54, 55).

The proportion of sites that conserve their hydrophobic or hydrophilic character in the many globin and variable domain sequences is shown in

Figure 10. About one third of the sites in the globin and immunoglobulin sequences are occupied by just nonpolar or neutral residues. Another one third of the sites are occupied by just polar or neutral residues (54, 55).

In naturally occurring sequences, functional requirements place restrictions on active site residues and on the structural changes that can occur in response to mutations. The saturation mutagenesis experiments on the arc repressor sequence (57) involved selection of two kinds: first for mutants that retain both structure and function, and second for mutants that retain structure but not function. The sequences in the latter set do show, of course, less conservation than the sequences of the former, but the proportion of sites that conserve their hydrophilic or hydrophobic character is only a little smaller (Figure 10). For both sets of engineered mutants, these proportions are close to those found in the naturally occurring sequences of the globins and variable domains (Figure 10).

Though these different proteins have similar proportions of conserved hydrophilic and hydrophobic sites, the distribution of these sites in their sequences differ: the distribution of hydrophobic and hydrophilic sites common to the globin sequences has been shown to be unique to that fold (54). The formation of the native secondary structure brings together sites of the same character to form hydrophobic and hydrophilic surfaces. The association of hydrophobic surfaces forms the protein interiors and leaves the hydrophilic surfaces accessible to solvent (Figure 11).

Protein Interiors

The residues in the interiors are close packed (67, 68). Small cavities do occur (69). But in most proteins whose structures are known at high resolution, empty cavities the size of methyl groups are either not found or only occur once (70, 70a). In addition, the side chains buried in the interior have low energy conformations (71–74). These features mean the volumes and shapes of the interior residues not only determine the arrangement of the packed secondary structures, but make the major contribution to the stability of protein structures (74–79, but see also 79a).

All the mutations that are known to make T4 lysozyme temperature sensitive occur at sites of low solvent accessibility (79). Similarly, in mutants of λ repressor that retain function, the number of substitutions allowed at different sites correlates with the accessibility of the site (76).

Engineered mutations that increase the size of buried residues by the equivalent of two methyl groups can destabilize native proteins by 2–4 kcal (78, see also 78a). This loss of stability arises from both steric hindrance and the introduction of strained conformations (74). Reduction in the size of internal residues can lead to the loss of 1.0–1.6 kcal per methylene group (77). In at least one case, however, a small reduction in size, Cys to Ala, stabi-

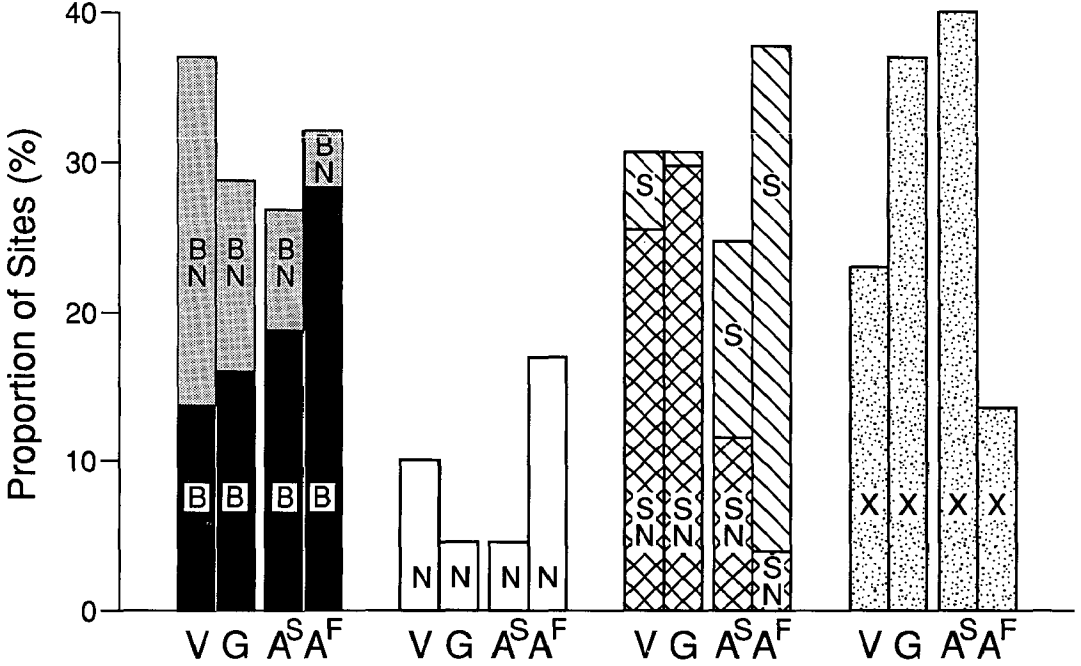

Figure 10 The proportion of sites in the sequences of the globins (G) and the immunoglobulin variable domains (V), and in the mutants of the arc repressor protein (A), that conserve their hydrophobic, neutral, or hydrophilic character. Two sets of results are given for the arc protein: one is from the sequences that conserve both structure and function (A^F), and the other is from the sequences that conserve structure but not function (A^S).

The proportion of sites where the residues are only polar (Arg, Lys, Glu, Asp, Gln, or Asn) is indicated by S; the proportion where the residues are neutral (Gly, Ala, Ser, Thr, Pro, His, or Tyr) by N; and the proportion where they are hydrophobic (Cys, Val, Leu, Ile, Met, Phe, or Trp) by B. The proportions of sites where the different sequences have both hydrophobic and neutral residues is indicated by BN and those where they have hydrophilic or neutral residues by SN. The proportion of sites not showing any strong residue conservation is indicated by X. [This figure is drawn from data given in (54–57).]

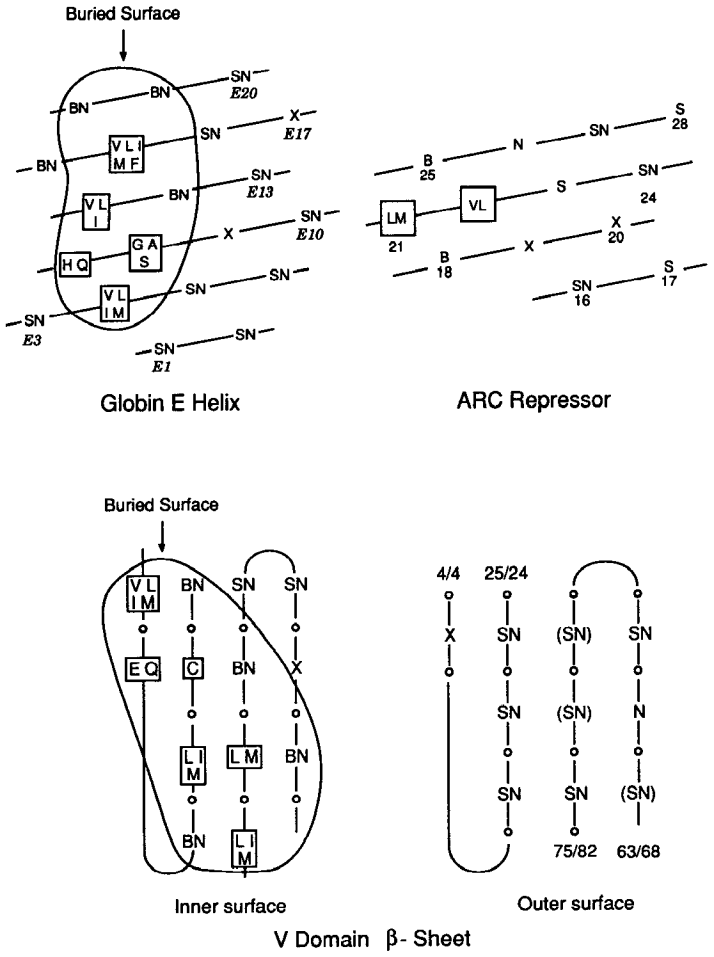

Figure 11 The patterns formed by the sites that show residue conservation in sequences of a helical region in the globins, a helical region in the arc protein, and a sheet region in the immunoglobulin variable domains. The data shown here come from the analysis of the many sequences that give the globin and immunoglobulin folds and of the mutants consistent with the fold of the arc repressor protein (54–57). The sites in the globins and arc repressor protein are drawn on flat projections of the helices. In the β-sheet drawing the sites pointing toward the reader are shown on the left and those that point away are shown on the right. The regions of the globin helix and variable domain β-sheet that are buried in the known structures are indicated.

For the globin helix and immunoglobulin sheet we indicate the conservation found in at least 97% of the many known sequences, except for bracketed sites ( ) where the conservation is at least 90%. If closely related residues are conserved, they are indicated by their one-letter code, e.g. V, I, and L at site E11. Sites where the residues are only polar (Arg, Lys, Glu, Asp, Gln, or Asn) are indicated by S; those where the residues are neutral (Gly, Ala, Ser, Thr, Phe, His, or Tyr) by N, and those where they are hydrophobic (Cys, Val, Leu, Ile, Met, Phe, or Trp) by B. Sites where the different sequences have both hydrophobic and neutral residues are indicated by BN and those where they have hydrophilic or neutral residues by SN. Sites not showing any strong residue conservation are indicated by X. [This figure is drawn from data given in (54–57).]

lizes the structure (78) and shows that the internal packing may contain imperfections.

Although the residues that form the protein interior are usually nonpolar or neutral, there are rare cases of buried polar residues. The sets of hydrogen bonds made by such residues are often particular to their side chains. This makes them difficult to replace during evolution. For example, the side chain of residue 6 Glu/Gln in the variable domains is buried and hydrogen bonded between the two β-sheets (Figure 11). Of the 1580 different sequences known for the initial regions of the variable domains, 99% have Gln or Glu at position 6 (56).

Sequence Divergence and Protein Folds

In proteins that have the same fold, the identity of the residues on both the surface and interior may differ. In families of related proteins, however, the deeply buried residues usually differ by no more than two methyl groups. Buried residues adjacent to the surface may show larger variations (54, 55) (Figure 10).

The analysis of the structures of homologous proteins and engineered mutants has shown how the adaptation to mutations occurs through changes in structure that are consistent with the maintenance of function (74, 79, 80). Individual acceptable mutations produce only very small changes in structure (74, 79), but the cumulative effects of many sequence changes produce large differences (80, 81) (see below).

Although structural change can be used to adapt to mutations, series of acceptable mutations must be cooperative if at each step they maintain the close packing and low energy conformations necessary for stability (80, 81). The cooperative nature of acceptable mutations has been demonstrated by the engineered mutations made to the buried core of the λ repressor protein (58). The buried interior of this protein is formed by seven nonpolar residues. In turn, one, two, and three of these sites were subjected to mutagenesis. Analysis of the sequence changes compatible with a functional protein shows that the changes allowed at a particular site are strongly influenced by the changes that occur at other sites. Thus, sets of buried residues that have same total volume but different shapes make quite different contributions to stability (58). Also the total volume variations that can be accommodated by a functional protein increases with the number of sites subjected to mutation. The residues at the buried seven sites in the native protein have a total volume of 549 Å^3. For the engineered sequences it was found that:

Number of sites subjected to mutation	Total volume of the seven buried residues consistent with a functional protein (Å^3)
1	531–567
2	526–585
3	500–585

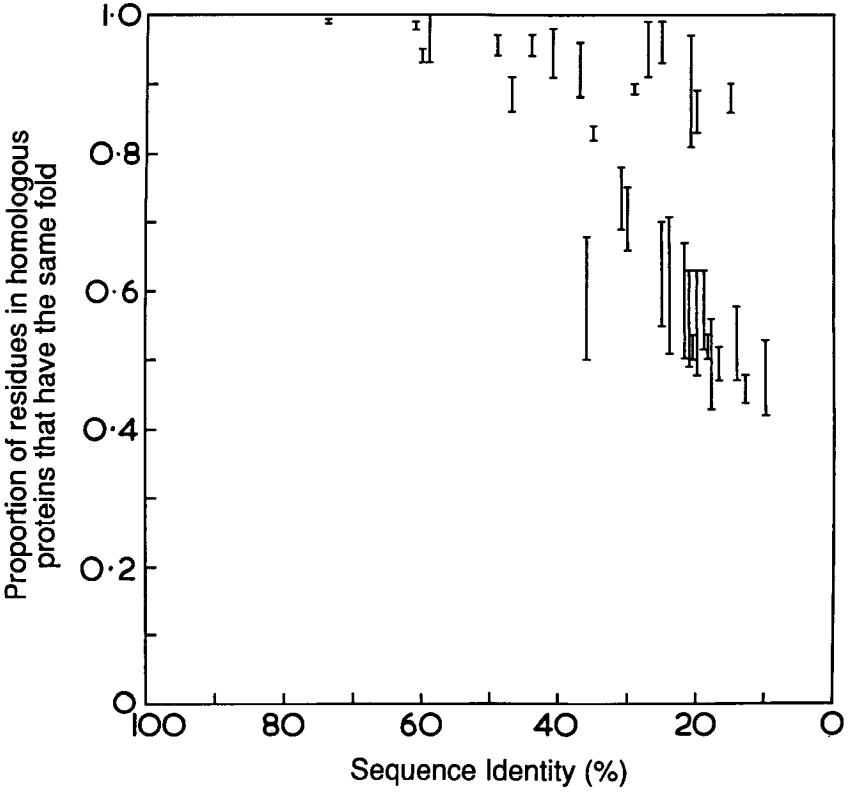

Figure 12 The relation between the divergence of protein sequences and folds.

If two related proteins, with m and n residues, have c residues with the same fold, the proportions of each structure with the same fold is c/m and c/n. For 32 pairs of related proteins from eight protein families the values of c/m and c/n, linked by a line, are plotted against the percentage of the c residues that are identical. For proteins whose sequence identity is 40% or greater, the regions with the same fold compose 90% or more of each structure. When the residue identities fall to 20%, it is usual for the regions with the same fold to form about half of each structure.

[Reprinted from (69) with permission.]

The extent of the structural changes increases exponentially with the differences in sequence (82, 83). Proteins whose residue identities are more than 40% have very similar structures, but as the identity decreases below this value, the structural differences become increasingly large (82). The major secondary structures shift in relation to each other by several Å. The peripheral elements of secondary structure and the peptides that link the major elements may change their fold or be deleted. For pairs of proteins with residue identities of 20%, it is common for only half of their structures to have the same fold (Figure 12).

Close similarities in the folding patterns of two proteins are sometimes taken to imply evolutionary relationships, even if there are no significant similarities in sequence. Such a relation may exist, but the preferences for particular arrangements of secondary structures and chain topologies mean that proteins can share the same folding pattern but be unrelated. A clear illustration of this point is given by a calculation carried out on two helical proteins (31). The two proteins have α-helices in the same relative arrangement but with very different connectivities and so are not related. In spite of this lack of relationship, the similarly arranged α-helices in the two proteins can be superposed with a root mean square difference in position of their Cαs of 2.5 Å.

AMINO ACID SEQUENCES AND PROTEIN FOLDING PATTERNS

In the initial part of this review we discussed structural and topological features of different protein folding patterns. These features are either independent of sequence or only involve some average properties of side chains. In the latter sections we were concerned with the relations between the folds of certain proteins and the different sequences that are consistent with those folds. In this last section we bring these and other results together to discuss the relations between amino acid sequences and folding patterns.

Experimental and structural studies have revealed the major determinants of the secondary structures formed by a region of a particular sequence. Different methods for the prediction of secondary structures on the basis of residue composition alone give correct results for 55–65% of residues (84, 84a). This accuracy is appreciably higher than the expected value for a random prediction (85), and clearly indicates that the particular residue composition of regions creates a predisposition for formation of helices, strands, or turns. Mutation studies have shown how interactions between side chains both stabilize and limit the extent of the nativelike secondary structures formed by protein fragments; see (86) in this volume for a review of this work.

In the observed structures a crucial role of the native secondary structure is the creation of separate hydrophobic and hydrophilic surfaces. The hydrophobic surfaces are formed by nonpolar and neutral residues; the hydrophilic surfaces by polar and neutral residues (see above). In the large majority of soluble proteins the ratios of the polar, nonpolar, and neutral residues are very similar, close to 3:3:4 (87), and the distribution of these residues through the sequence is random in character (88, 88a). It can be shown that, with the usual composition and distribution of residues, the average length of regions that can be expected to form amphipatic helices and

β-sheet strands are 11 and 6 residues, respectively, values close to the observed averages (2). Similarly, the formation of a β-sheet hydrophobic on both sides is expected to be quite probable, whereas a completely hydrophobic helix (twice as many residues as in a strand) is expected to be a rare event, as is observed (2). It is these features that result in most proteins being formed by two or three layers of secondary structures and that limit the size of domains.

The set of secondary structures and their hydrophobic surfaces formed by sequences are the primary determinants of the folding patterns. The manner in which secondary structures are assembled is mainly determined by (*a*) the preference in the folding pathway for intermediates of low energy favoring certain chain topologies, (*b*) the close packing of the buried surfaces that favors one of a small number of packing geometries for most sets of residues, and (*c*) sufficient surface being buried to provide stability.

For related proteins, the deeply buried homologous residues are similar in size (54, 55, 58) (see above). In unrelated proteins that share the same folding patterns this is not the case. Certain different sets of side chains can have low-energy conformations, be close packed, and give identical packing geometry (73). For proteins that share the same folding patterns, the packing geometries are only roughly similar and the differences in geometry will allow many quite different sets of residues to form the interior (81). Thus in unrelated proteins with the same folding pattern, the residues at equivalent buried sites are hydrophobic but usually have quite different identities.

CONCLUSION

Protein folding patterns were first described more than 10 years ago when less than a quarter of the presently known structures were available (8–14). Since then some new patterns have been discovered and the definitions of old ones have been refined, and this will continue. It is, however, remarkable that the initial set quite accurately describes the folding patterns found in the large majority of the new structures. This clearly implies that the intrinsic properties of proteins result in most sequences forming structures in which the secondary structures pack in one of a small number of basic geometries and the chain has one of a small number of topologies. The work reviewed here suggests that the properties of proteins that determine folding patterns are well understood in outline. We believe we know why certain geometries and topologies are preferred.

We do not yet, however, have an understanding that is sufficiently exact to predict, correctly and in general, a particular folding pattern for a particular sequence. The recent advances made by the study of engineered and homologue sequences, by particular predictions (2, 85, 89) and by the analysis of

folding pathways (86), suggests that progress will now be made in this direction.

ACKNOWLEDGMENTS

We thank Annette Lenton and John Creswell for the figure drawings, Drs. S. Abdel-Meguid and P. Colman for information prior to publication, the Royal Society for support, and the European Molecular Biology Organization for the award of a short-term fellowship to A. F.

Literature Cited

1. Chothia, C. 1984. *Annu. Rev. Biochem.* 53:537–72
2. Finkelstein, A. V., Ptitsyn, O. 1987. *Prog. Biophys. Mol. Biol.* 50:171–90
3. Richardson, J. S., Richardson, D. C. 1989. In *Prediction of Protein Structure and the Principles of Protein Conformation,* ed. G. D. Fasman, New York: Plenum
4. Schulz, G. E., Schirmer, R. H. 1979. *Principles of Protein Structure*. New York: Springer-Verlag. 314 pp.
5. Ptitsyn, O. B., Finkelstein, A. V. 1980. *Q. Rev. Biophys.* 13:339–86
6. Richardson, J. S. 1981. *Adv. Protein Chem.* 34:167–339
7. Rose, G. D., Young, W. B., Gierasch, L. M. 1983. *Nature* 304:655–57
8. Levitt, M., Chothia, C. 1976. *Nature* 261:552–57
9. Chothia, C., Levitt, M., Richardson, D. 1977. *Proc. Natl. Acad. Sci. USA* 74:4130–34
10. Sternberg, M. J. E., Thornton, J. M. 1976. *J. Mol. Biol.* 105:367–82
11. Richardson, J. S. 1976. *Proc. Natl. Acad. Sci. USA* 73:2619–23
12. Sternberg, M. J. E., Thornton, J. M. 1977. *J. Mol. Biol.* 110:269–83
13. Richardson, J. S. 1977. *Nature* 268:495–500
14. Nagano, K. 1977. *J. Mol. Biol.* 109: 235–50
15. Flory, P. J. 1968. *Statistical Mechanics of Chain Molecules*. New York: Interscience Publ.
16. Landau, L. D., Lifshitz, E. M. 1959. *Statistical Physics,* Part 1. London: Pergamon
17. Crippen, G. M. 1974. *J. Theor. Biol.* 45:327–38
18. Vologodskii, A. V., Lukashin, A. V., Frank-Kamenetskii, M. D., Anshelevich, V. V. 1974. *Zh. Eksp. Teor. Fiz.* 66:2153–63
19. Frank-Kamenetskii, M. D., Lukashin, A. V., Vologodskii, A. V. 1975. *Nature* 258:398–402
20. Frank-Kamenetskii, M. D., Vologodskii, A. V. 1981. *Usp. Fiz. Nauk.* 134:641–74
21. Connolly, M. L., Kuntz, I. D., Crippen, G. M. 1980. *Biopolymers* 19:1167–82
22. Privalov, P. L. 1979. *Adv. Protein Chem.* 33:167–241
23. Kauzmann, W. 1959. *Adv. Protein Chem.* 14:1–63
24. Miller, S., Lesk, A. M., Janin, J., Chothia, C. 1987. *Nature* 328:834–36
25. Crick, F. H. C. 1953. *Acta Crystallogr.* 6:689–97
26. Chothia, C., Levitt, M., Richardson, D. 1981. *J. Mol. Biol.* 145:215–50
27. Argos, P., Rossmann, M. G., Johnson, J. E. 1977. *Biochem. Biophys. Res. Commun.* 75:83–86
28. Weber, P. C., Salemme, F. R. 1980. *Nature* 287:82–84
29. Presnell, S. R., Cohen, F. E. 1989. *Proc. Natl. Acad. Sci. USA* 86:6592–96
30. Murzin, A. G., Finkelstein, A. V. 1983. *Biofizika* 28:905–11
31. Murzin, A. G., Finkelstein, A. V. 1988. *J. Mol. Biol.* 204:749–70
32. Efimov, A. V. 1977. *Dokl. Akad. Nauk SSSR* 235:699–702
33. Efimov, A. V. 1979. *J. Mol. Biol.* 134:23–40
34. Henderson, R., Unwin, P. N. T. 1975. *Nature* 257:28–32
35. Parker, M. W., Pattus, F., Tucker, A. D., Tsernoglou, D. 1989. *Nature* 337: 93–96
36. Chothia, C., Janin, J. 1981. *Proc. Natl. Acad. Sci. USA* 78:4146–50
37. Cohen, F. E., Sternberg, M. J. E., Taylor, W. R. 1981. *J. Mol. Biol.* 148:253–72
38. Chothia, C., Janin, J. 1982. *Biochemistry* 21:3955–65
39. McLachlan, A. D. 1979. *J. Mol. Biol.* 133:557–63

40. Priestle, J. P., Schar, H.-P., Grutter, M. G. 1988. *EMBO J.* 7:339–343
41. Varghese, J. N., Laver, W. G., Colman, P. 1983. *Nature* 303:35–40
42. Chirgadze, Y. N. 1987. *Acta Crystallogr. A* 43:405–17
43. Efimov, A. V. 1982. *Mol. Biol. (USSR)* 16:799–806
43a. Gibson, T. J., Argos, P. 1990. *J. Mol. Biol.* 212:7–9
44. Janin, J., Chothia, C. 1980. *J. Mol. Biol.* 143:95–128
45. Cohen, F. E., Sternberg, M. J. E., Taylor, W. R. 1982. *J. Mol. Biol.* 156:821–62
46. Stallings, W., Abdel-Meguid, S. S. 1990. Submitted for publication
47. Sternberg, M. J. E., Cohen, F. E., Taylor, W. R., Feldmann, R. J. 1981. *Philos. Trans. R. Soc. London, Ser. B* 293:177–89
48. Banner, D. W., Bloomer, A. C., Petsko, G. A., Phillips, D. C., Pogson, C. I., et al. 1975. *Nature* 255:609–14
49. Lasters, I., Wodak, S. J., Alard, P., van Cutsem, E. 1988. *Proc. Natl. Acad. Sci. USA* 85:3338–42
50. Lesk, A. M., Branden, C.-I., Chothia, C. 1989. *Proteins* 5:139–48
51. Hofmann, B. E., Bender, H., Schulz, G. E. 1989. *J. Mol. Biol.* 209:793–800
52. McLachlan, A. D. 1979. *J. Mol. Biol.* 128:49–79
53. Lebioda, L., Stec, B., Brewer, J. M. 1989. *J. Biol. Chem.* 264:3685–92
54. Bashford, D., Chothia, C., Lesk, A. M. 1987. *J. Mol. Biol.* 196:199–216
55. Chothia, C., Bashford, D., Lesk, A. M. 1990. In preparation
56. Kabat, E. A., Wu, T. T., Reid-Miller, M., Perry, H. M., Gottesman, K. S. 1987. *Sequences of Proteins of Immunological Interest*. Washington, DC: Public Health Service, NIH. 4th ed.
57. Bowie, J. U., Sauer, R. T. 1989. *Proc. Natl. Acad. Sci. USA* 86:2152–56
58. Lim, W. A., Sauer, R. T. 1989. *Nature* 339:31–36
59. Chothia, C. 1976. *J. Mol. Biol.* 105:1–14
60. Kyte, J., Doolittle, R. F. 1982. *J. Mol. Biol.* 157:105–32
61. Rose, G. D., Geselowitz, A. R., Lesser, G. L., Lee, R. H., Zehfus, M. H. 1985. *Science* 229:834–38
62. Miller, S., Janin, J., Lesk, A. M., Chothia, C. 1987. *J. Mol. Biol.* 196: 641–56
63. Eisenberg, D., Schwarz, E., Komaromy, M., Wall, R. 1984. *J. Mol. Biol.* 179:125–42
64. Rose, G. D., Gierasch, L. M., Smith, J. A. 1985. *Adv. Protein Chem.* 37:1–109
64a. Cornette, J. L., Cease, K. B., Margalit, H., Spouge, J. L., Berzofsky, J. A., Delisi, C. 1987. *J. Mol. Biol.* 195:659–85
65. Perutz, M. F., Kendrew, J. C., Watson, H. C. 1965. *J. Mol. Biol.* 13:669–78
65a. Schiffer, M., Edmundson, A. B. 1967. *Biophys. J.* 7:121–35
66. Lim, V. I., Ptitsyn, O. B. 1970. *Mol. Biol.* 4:372–82
66a. Eisenberg, D., Weiss, R. M., Terwilliger, T. C. 1982. *Nature* 299:371–74
67. Richards, F. M. 1974. *J. Mol. Biol.* 82:1–14
68. Chothia, C. 1975. *Nature* 254:304–8
69. Richards, F. M. 1979. *Carlsberg Res. Commun.* 44:47–63
70. Rashin, A. A., Iofin, M., Honig, B. 1986. *Biochemistry* 25:3619–25
70a. Connolly, M. 1986. *Int. J. Peptide Protein Res.* 28:360–63
71. Janin, J., Wodak, S., Levitt, M., Maigret, B. 1978. *J. Mol. Biol.* 125:357–86
72. Gelin, B. R., Karplus, M. 1979. *Biochemistry* 18:1256–68
73. Ponder, J. W., Richards, F. M. 1987. *J. Mol. Biol.* 193:775–91
74. Karpusas, M., Baase, W. A., Matsumura, M., Matthews, B. W. 1989. *Proc. Natl. Acad. Sci. USA* 86:8237–41
75. Perutz, M. F., Lehmann, H. 1968. *Nature* 219:902–9
76. Reidhaar-Olson, J. F., Sauer, J. F. 1988. *Science* 241:53–57
77. Kellis, J. T., Nyberg, K., Fersht, A. R. 1989. *Biochemistry* 28:4914–22
78. Hughson, F. M., Baldwin, R. L. 1989. *Biochemistry* 28:4415–22
78a. Radding, J. A. 1987. *Biochemistry* 26:3530–36
79. Alber, T., Sun, D.-P., Nye, J. A., Muchmore, D. C., Matthews, B. W. 1987. *Biochemistry* 26:3754–58
79a. Sali, D., Byroft, M., Fersht, A. R. 1988. *Nature* 335:740–43
80. Chothia, C., Lesk, A. M. 1987. *Cold Spring Harbor Symp. Quant. Biol.* 52:399–405
81. Lesk, A. M., Chothia, C. 1980. *J. Mol. Biol.* 136:225–70
82. Chothia, C., Lesk, A. M. 1986. *EMBO J.* 5:823–26
83. Hubbard, T. J. P., Blundell, T. L. 1987. *Protein Eng.* 1:159–71
84. Kabsch, W., Sander, C. 1983. *FEBS Lett.* 155:179–82
84a. Ptitsyn, O. B., Finkelstein, A. V. 1989. *Protein Eng.* 2:443–47
85. Schulz, G. E. 1988. *Annu. Rev. Biophys. Biophys. Chem.* 17:1–21

86. Kim, P. S., Baldwin, R. L. 1990. *Annu. Rev. Biochem.* 59:631–60
87. Dayhoff, M. D. 1976. *Atlas of Protein Sequences and Structure,* Vol. 5, Suppl. 2. Washington, DC: Natl. Biomed. Res. Found.
88. Poroykov, V. V., Esipova, N. G., Tumanyan, V. G. 1984. *Mol. Biol. (USSR)* 18:541–47
88a. Ptitsyn, O. B., Volkenstein, M. V. 1986. *J. Biomol. Struct. Dyn.* 4:137–56
89. Crawford, I. P., Niermann, T., Kirschner, K. 1987. *Proteins Struct. Funct. Genet.* 2:118–29
90. Lee, B., Richards, F. M. 1971. *J. Mol. Biol.* 55:379–400
91. Janin, J., Miller, S., Chothia, C. 1988. *J. Mol. Biol.* 204:155–64
92. Holmes, M. A., Matthews, B. W. 1982. *J. Mol. Biol.* 160:623–39
93. Chothia, C. 1989. *Nature* 337:204–5
94. Leijonmarck, M., Eriksson, S., Liljas, A. 1980. *Nature* 286:824–26
95. Burnett, R. M., Darling, G. D., Kendall, D. S., Lesquesne, M. E., Mayhew, S. G., et al. 1974. *J. Biol. Chem.* 249:4383–92

Annual Review of Biochemistry
Volume 59, 1990

CONTENTS

(*continued*)