6

The Development of the Prediction of Protein Structure

Gerald D. Fasman

Gerald D. Fasman • Graduate Department of Biochemistry, Brandeis University, Waltham, Massachusetts 02254.

G. D. Fasman (ed.), *Prediction of Protein Structure and the Principles of Protein Conformation*

I. INTRODUCTION

The tenet of structural biology that function follows form had its seeds in the monograph by C. B. Anfinsen, *The Molecular Basis of Evolution* (Anfinsen, 1959), wherein he stated "Protein chemists naturally feel that the most likely approach to the understanding of cellular behavior lies in the study of structure and function of protein molecules." The achievement of protein crystallography over the past 30 years has confirmed this view whereby the description of the structure and function of proteins is now frequently understood at the atomic level.

The classical experiments of Anfinsen and co-workers (Anfinsen *et al.*, 1961; Anfinsen, 1973) proved that ribonuclease could be denatured and refolded without loss of enzymatic activity. This implied that the amino acid sequence contains sufficient information to define the three-dimensional structure of a protein in a particular environment. The acceptance of this tenet has led to multifarious efforts to predict the conformation of proteins based only on the consideration of sequence alone, which has been termed the protein-folding problem. Both theoreticians (e.g., Levitt and Warshel, 1975; Nemethy and Scheraga, 1977; Karplus and Weaver, 1979; Schulz and Schirmer, 1979; Sternberg, 1983) and experimentalists (e.g., Shoemaker *et al.*, 1987; Creighton, 1978) have tackled the chain-folding problem with very limited success.

It is thought that the native structure of a protein will lie near the minimum of free energy; however, it will not fold by sampling every possible conformation (Levinthal, 1968; Wetlaufer, 1973), but there will be one or more pathways along which the protein folds (Creighton, 1979). Anfinsen (1973) had proposed that one or more regions of secondary structure, e.g., α helices or a two-stranded antiparallel β sheet, having marginal stability would act as nucleation sites and direct the refolding.

The advent of recombinant DNA techniques has led to an explosion of information concerning the sequences of receptors and enzymes that will be important for drug, herbicide, and pesticide design. Technological developments of industrial, clinical, and agricultural importance may be achieved in the coming years by imitation of the interactions between macromolecules and ligands that occur naturally in the living cell (Blundell *et al.*, 1987). Engineered hormones may have more desirable properties than their native counterparts in terms of stability or activity. All these promising prospects have fired the desire to predict the conformation of and design protein molecules with a high degree of sophistication.

The theoretical efforts could be categorized into three main areas: energetic, heuristic, and statistical. The school following the assumption that a protein folds so as to minimize the free energy of the system has had many contributors (e.g., Levitt and Warshel, 1975; Nemethy

and Scheraga, 1977; McCammon *et al.*, 1977; Weiner *et al.*, 1984). These researchers have developed potential functions to describe the energy surface of a polypeptide chain. Chain folding is simulated computationally, directed by surface gradients to find the energy minimum. Alternatively, conformation space is probed from a starting point by integrating the equations of motion over time. These theoretical predictions of structure have been influenced by both thermodynamic and kinetic arguments. The thermodynamic properties of the polypeptide chain were the first to be considered. Liquori and co-workers (DeSantis *et al.*, 1965) and Ramachandran *et al.* (1963) first demonstrated that the peptide unit can adopt only certain allowed conformations. They constructed ϕ, ψ plots, which were subsequently improved by semiempirical energy calculations (Lewis *et al.*, 1973b), to predict preferred regions for the various secondary structures of polypeptides. Recent energy calculations have been refined, either by choosing parameters to fit crystal structures (Hagler *et al.*, 1974) or by performing complex molecular orbital calculations (Pullman and Pullman, 1974). On the basis of the thermodynamic hypothesis, by calculating the total free energy of a protein and finding the global minimum, it should be possible to predict the native structure. However, minimization schemes have failed to predict chain folding accurately (Hagler and Honig, 1978; Cohen and Sternberg, 1980a,b). The lack of success presumably stems from the difficulties in modeling protein–solvent interactions, the use of analytical functions to approximate the chemical potential, and the compounding of these errors in the computed gradient; in addition, the energy surface has multiple minima, which make it nearly impossible to locate a global minimum. Attempts at solving these problems are underway. Molecular dynamics offers solutions to these problems but remains computationally limited as a technique for studying chain folding (McCammon *et al.*, 1977). Chain folding is thought to take place in the millisecond time scale (Baldwin, 1980). Elaborate computing resources must be applied to sample 100 nsec in the life of a small protein (Post *et al.*, 1986).

One of the first attempts to predict the conformation of a protein, ribonuclease, was that of Scheraga (1960). This model used the available chemical information, deuterium–hydrogen exchange data, and the knowledge of the α helix. It later was shown to have little similarity to the native molecule.

The prediction of the secondary structure of protein conformation was pioneered by Guzzo (1965). By analysis of the known sequences and structures of myoglobin and the α- and β-hemoglobins, he found that groups of amino acids were helix disrupters. He also emphasized the role of hydrophobic interactions. He noted that Blout (1962) had pointed out that the known poly(α-amino acids) fall into two categories, helix formers and breakers. However, earlier, Davies (1964) had detected a correlation between amino acid composition and protein structure. By using optical properties for 15 proteins and their known sequences, it was possible to compare estimates of helicity with amino acid composition. Other early researchers in this area included Prothero (1966, 1968), Cook (1967), Periti *et al.* (1967), and Low *et al.* (1968), who all examined the few x-ray structures available and attempted to ferret out patterns of recognition. Schiffer and Edmundson (1967) adopted an innovative way of locating hydrophobic arcs by the "helical wheel" method. Dunhill (1968) slightly improved this approach by constructing helical net diagrams to locate hydrophobic clusters. Ptitsyn (1969) made a statistical analysis of the distribution of different amino acid residues among helical and nonhelical regions of seven globular proteins: myoglobin, α- and β-hemoglobin, lysozyme, ribonuclease, α-chymotrypsin, and papain. It was found that the distribution of a number of amino acid residues differed essentially from the distribution averaged over all amino acids. Hydrophobic amino acids, Ala and Leu, showed a tendency to occupy internal turns of helical regions. Negatively and positively charged amino acids have a tendency to concentrate, correspondingly, at the amino and carboxyl ends of helical regions. Amino acids with heteroatoms (O or S atoms) near the main chain (Ser, Thr, Cys, Asn) have a tendency to concentrate in nonhelical

regions (including the regions with β structure). Proline can be located either on the amino ends of helical regions or in nonhelical regions. On examination it was found that Guzzo's (1965) empirical rules concerning the relationship between the amino acid composition of the given regions of the polypeptide chain and its structure was in disagreement with the results of the statistical analysis, whereas Prothero's (1966) rule agreed with them. Lewis *et al.* (1970, 1971), using the Zimm–Bragg (1959) helix–coil theory, suggested that better predictions may result when σ and *s* parameters for all 20 amino acids were experimentally determined.

Kotelchuck and Scheraga (1968, 1969) were among the first to attempt energy-minimization calculations to test the hypothesis that short-range interactions seem to play an important role in helix formation. Calculations were carried out to obtain the energy of interactions of individual amino acids along the backbone. A series of papers followed (see review by Scheraga, 1985), but these energy-minimization schemes have failed to predict chain folding accurately (Hagler and Honig, 1978; Cohen and Sternberg, 1980a,b). The reasons for this have been discussed above.

There have been between 15 and 20 proposals for the prediction of the secondary structure of protein conformation from the amino acid sequence. The number of methods counted depends on the degree of difference one defines between methods.

To predict the various secondary structures first requires a precise definition of the various secondary structures. Various criteria have been used by x-ray crystallographers, and thus secondary structures are delineated in different ways in various publications (see Blundell and Johnson, 1976). The crystallographers' assignments of secondary structure in the Brookhaven Protein Data Bank are often subjective and often incomplete. In an attempt to overcome this problem, Levitt and Greer (1977) described a computer program to analyze automatically and objectively the atomic coordinates of a large number of globular proteins (62) in order to identify the regions of α-helix, β-sheet, and reverse-turn secondary structure. The most successful criterion was based on the patterns of peptide hydrogen bonds, inter-C^{α} distances, and inter-C^{α} torsion angles. A more recent compilation of secondary structures has been published by Kabsch and Sander (1983b). This is a pattern-recognition process of hydrogen-bonded and geometrical features extracted from x-ray coordinates. Cooperative secondary structure is recognized as repeats of the elementary hydrogen-bonding patterns "turn" and "bridge." Repeating turns are "helices," repeating bridges are "ladders," and connecting ladders are "sheets." Geometric structure is defined in terms of the concepts of torsion and curvature of differential geometry. Local chain "chirality" is the torsional handedness of four consecutive C^{α} position and is positive for right-handed helices and negative for ideal twisted β sheets. Curved pieces are defined as bends. Solvent exposure is given as the number of water molecules in possible contact with residues. This dictionary is available in computer-readable form for protein structure prediction work.

II. PROTEIN TOPOLOGY

Proteins may be classified into three main categories: (1) globular proteins, (2) membrane proteins, and (3) fibrous proteins. Although fibrous proteins play very important roles in the anatomy and physiology of animals, they are not considered in this chapter.

The globular proteins pack tightly to form distinct tertiary conformations, usually producing a hydrophobic core. The packing involves the association of the various secondary structures, α helices and β strands with intervening β turns, into various domains. Often disulfide cross links between cysteine residues further stabilize the structure. These domains are formed from compact, contiguous chain structures (Wetlaufer, 1973; Liljas and Rossmann, 1974). These units have been proposed as intermediates in the folding process, as functional units

within the native structure (Richardson, 1981; Rossmann and Argos, 1981; Wetlaufer, 1981; Kim and Baldwin, 1982), and possibly even as fundamental genetic units (Gilbert, 1978; Blake, 1979). Schulz (1977) outlined a conceptual, partially hierarchic framework that emphasized the role of domains within the known protein structures. Levitt and Chothia (1976) had earlier observed that assemblies of a few adjacent secondary structural units often pack and serve as structural building blocks. Lewis *et al.* (1971) and Kuntz (1972), in defining turns, laid the foundation for the concept of domains. Turns constitute recognizable structural units in proteins and are situated at the solvent-accessible surface of the molecule.

The domains have been classified into five classes according to their secondary structural arrangement (Levitt and Clothia, 1976; Richardson, 1981). These classes are (1) all α, with α helices and no β strands; (2) all β, with only β strands and no α helices; (3) α/β, in which the chain alternates between α helices and β strands; (4) α and β, in which the α and β regions tend to segregate into separate regions; and (5) coil, in which there is little or no regular secondary structure. The excellent review article by Richardson (1981) is amply illustrated with schematic diagrams, which have made the visualization of protein structure more aesthetic and a much simpler task. Levitt and Greer (1977) developed a computer program to analyze, automatically and objectively, the atomic coordinates of a large number of globular proteins in order to identify regions of α helix, β sheets, and β turns (reverse turns). The pattern of hydrogen bonds, inter-C^{α} distances, and inter-C^{α} torsion angles was used to find the secondary structure. Rose (1979) devised an automatic procedure for the identification of domains in globular proteins. The known protein structures were shown to be iteratively subdivisible into a hierarchy of disjunct contiguous-chain regions ranging in size from whole-protein monomers down to individual helices and strands. This analysis was later expanded as a control study to a set of 1000 simulated chain folds, and their organization was found to be similar to that of authentic molecules (Yuschok and Rose, 1983). The hierarchic nature of proteins was also observed by Crippen (1978) using a different algorithm. Rashin (1981) has proposed an algorithm to calculate domains defined as globular fragments, and Wodak and Janin (1980, 1981) have also shown that the presence of domains is easily detected by an automatic procedure based on surface areas only. More recently, Kabsch and Sander (1983b) described a program similar to Levitt and Greer's (1977), a pattern-recognition process of hydrogen-bonded and geometrical features plus solvent exposure of each residue, to produce a *Dictionary of Protein Secondary Structure* for 62 different globular proteins.

In addition to the important hydrophobic interactions and the hydrogen-bonding networks, there are several other interactions that play a role in protein conformation. The role that electrostatic forces play in the structure of proteins has been frequently discussed (see Chapter 8 by Rogers). Paul (1982) presented an interesting proposition that it is the electrostatic interactions that have a major influence on the low-resolution features of protein tertiary structure. Barlow and Thornton (1983) examined the role of ion pairs in proteins. Their "working definition" for an ion pair has been derived on the basis of an analysis of the distance distributions for like- and oppositely charged groups in 38 proteins. Ion pairs defined according to this criterion ($\leq$4 Å between charged groups) have been analyzed in respect to (1) the frequencies of different pair types, (2) the residue separations and secondary structural locations of the residues involved, (3) the flexibility of the side chains involved, (4) their conformation, (5) their environmental accessibility to solvent and proximity to active site or ligand-binding regions, and (6) their conservation in related proteins. On the average, one third of the charged residues in a protein are involved in ion pairs, and 76% of these are concerned with stabilizing the tertiary structure. Only 17% of ion pairs are buried.

Rashin and Honig (1984) examined the environment of ionizable groups in 36 proteins and characterized them in terms of solvent accessibility, salt-bridge formation, and hydrogen bonding. An interesting finding was that there are on the average two completely buried

ionizable groups per protein, of which at least 20% do not form salt bridges. However, all buried ionizable groups form hydrogen bonds with neutral polar groups. In a recent paper Sundaralingam *et al.*(1987) surveyed 47 globular proteins to determine the probability of occurrence of ion pairs separated by various numbers of residues in α helices. Ion pairs of the type $i,i + 3$ and $i,i + 4$ were the most predominant. The normalized frequencies of occurrence of ion pairs were also found to increase generally with helix length. These results indicate that ion pairs may contribute to the stability of the solvent-exposed α helices.

Sawyer and James (1982) reported finding carboxyl–carboxylate interactions important in stabilizing the structures of protein crystals and multisubunit complexes at low pH. These were found in protease A and protease B found in *Streptomyses griseus* and in penicillopepsin. Leszczynski and Rose (1986) introduced the proposition that loops in globular proteins constitute an additional secondary structure. These so-called "omega" loops were usually previously classified as "random coil." The segment length of the omega loop must be between six and 16 residues, have their backbone groups packed closely together, have a distance between segment termini of less than 10 Å, and may not exceed two thirds the maximum distance between any two α carbons within the segment under consideration. The frequencies of amino acid residues in loops were calculated and normalized in 67 proteins analyzed. The overall conformation was distributed as follows: 26% in the α helix, 19% in the β, 26% in turns, and 21% in loops. Thus, a loop may be described as a continuous-chain segment that adopts a "loop-shaped" conformation in three-dimensional space with a small distance between its termini. Morgan and McAdon (1980) reported finding that sulfur–aromatic interactions in globular proteins are not random events and are promoted in the presence of positively charged side chains. It has been pointed out (Burley and Petsko, 1985; Singh and Thornton, 1985) that aromatic–aromatic interactions frequently occur in proteins and that about 60% of the aromatic side chains are involved in such pairs, 80% of which form networks of three or more interacting side chains. Phenylalanine–phenylalanine interactions occur frequently with dihedral angles of 90° to each other. They stabilize tertiary structure (80%) and quaternary structure (20%). Amino–aromatic interactions also occur frequently, with the positively charged or δ(+) amino groups of Lys, Arg, Asn, Gln, and His preferentially located within 6 Å of the ring centroids of Phe, Tyr, and Try, where they make van der Waals contact with the δ(−) π electrons and avoid the δ(+) ring edge (Burley and Petsko, 1986).

Blundell *et al.* (1986b) comment on the paper of Burley and Petsko (1986) concerning the distribution of dihedral angles between two aromatic residues in globular proteins. It can provide evidence of specific interactions between aromatic rings only if it differs significantly from a random distribution. The expected distribution can be considered in terms of the angles between the normals to the two ring planes. Singh and Thornton (1985) have pointed out that the distribution of this angle that would arise by chance varies as the sine of the angle and has a mean value of ~57°. The overall distribution in the paper of Burley and Petsko (1986) thus closely approximates a random distribution. However, it is also necessary to consider the spatial displacement of the two aromatic rings between the ring centers. When this is examined, it shows significant deviations from random arrangement. Thus, the importance of taking into consideration the available three-dimensional space and expected "random" distribution for such side-chain interactions is stressed. A striking preference for perpendicular packing of aromatic rings is observed for a small subgroup in a special spatial displacement.

III. TECHNIQUES OF PROTEIN PREDICTION

The approach termed "knowledge-based" prediction (Blundell *et al.*, 1987) depends on analogies between a protein that is to be modeled and other proteins of known three-dimen-

sional structure at all levels in the hierarchy of protein organization: secondary structure, motifs, domains, and quaternary or ligand interactions.

The first requirement is to have available the amino acid sequence. Protein sequences are collected and made available from the Protein Information Resource (PIR) databank at the National Biomedical Research Foundation, Maryland (Georgetown University Medical Center, 3900 Reservoir Road, N.W., Washington, DC 20007), where over 7000 entries are available (April, 1988, with about 1000 new entries/month), or the NEWAT data bank in the United States (for review of sequence data banks, see Kneale and Bishop, 1985). The second source is from the DNA data banks, which can be easily converted into the primary amino acid structure. The European Molecular Biology Laboratory (EMBL, Heidelberg), the National Biomedical Research Foundation, Maryland, and GenBank (Intelligenetics, 700 East El Camino Real, Mountain View, California 94040-2216) have vast collections of DNA sequences (see list at end of chapter).

A. Sequence Alignment

One of the earliest attempts to determine whether the relationships existing between protein sequences resulted from homology or chance was by the sequence alignment algorithm of Needleman and Wunsch (1970). The maximum match is a number dependent on the similarity of the sequences. Comparisons were made from the smallest unit of significance, a pair of amino acids, one from each protein. All possible pairs are represented by a two-dimensional array, and all possible comparisons were represented by pathways through the array. The use of dynamic programming methods to carry out global comparison of two sequences was soon amplified by many techniques (e.g., Sellers, 1974; Waterman *et al.*, 1976). Their proficiency depends on the degree of similarity and may give variable results depending on the gap-penalty parameters chosen, even for closely related sequences (Fitch and Smith, 1983; Barton and Sternberg, 1987a,b). For sequences with greater than 25% similarity, such automatic procedures will identify the homology above the background of randomized sequences. Because insertions and deletions often occur at the loop regions of proteins between secondary structures, improvements can be made by introducing penalties for insertions/deletions in α helices or β strands (Barton and Sternberg, 1987a; Lesk *et al.*, 1986).

For very distantly related sequences, homology may be restricted to a few key residues or sequence segments whose separation along the chain may vary considerably between proteins. There are several newer algorithms to identify such local homologies (Goad and Kanehisa, 1982; Sellers, 1979; Boswell McLachlan, 1984; Taylor, 1986a; Gribskov *et al.*, 1987). McLachlan (McLachlan, 1971; McLachlan and Stewart, 1976) was an early contributor to sequence alignment and published an improved method based on those of Fitch (1966a–c), Cantor and Jukes (1966), Needleman and Blair (1969), and Haber and Koshland (1970).

Crippen (1977a,b) examined the x-ray crystal structures of 19 selected proteins for correlations between amino acid sequences and long-range tertiary conformation. He found clear evidence for preferential association between certain types of amino acids, particularly among hydrophobic alphatic, aromatic, and cysteine residues. Because the likelihoods of forming these residue pair contacts are all less than 12%, packing and geometric requirements must take precedence over energetic considerations.

An algorithm for secondary structure determination based on sequence similarity (Levin *et al.*, 1986) claimed a prediction accuracy of 62.2% over three states for 61 proteins for a new set of seven proteins not in the original database. An example of the use of sequence analysis can be found in the work of de Groot *et al.* (1987). The amino acid sequences of the spike proteins from three distantly related coronaviruses were aligned initially by FASTP analysis (Lipman and Pearson, 1985). These alignments were further extended by reiterating FASTP

with nonaligned parts as query sequences and by DIAGON comparison (Staden, 1982). DIAGON plots revealed two repetitious regions in the C-terminal domains with a seven-residue periodicity. Further analysis showed the presence of so-called "heptad repeats" (Cohen and Parry, 1986). Because heptad repeats are indicative of a coiled-coiled structure, it was suggested that these spike proteins of coronaviruses had this supersecondary structure.

The following have published algorithms based on various sequence alignment techniques: Low *et al.* (1968), Remington and Matthews (1978), Lesk and Chothia (1980), Maizel and Lenk (1981), Staden (1982), Sippl (1982), Wilbur and Lipman (1983), Kabsch and Sander (1984), Murata *et al.*, (1985), Lipman and Pearson (1985), Sweet (1986), Lesk *et al.* (1986), Chothia and Lesk (1986), Taylor (1986a), Nishikawa and Ooi (1986), Bacon and Anderson (1986), Gribskov *et al.* (1987), Argos (1987a,b), Barton and Sternberg (1987b), Zvelebil *et al.* (1987), and Bashford *et al.* (1987).

B. Hydrophobicity

The hydrophobic effect, as first expounded by Kauzmann (1959), is primarily entropic, resulting from the unfavorable ordering of water molecules that associate with exposed nonpolar atoms. Since the publication of this report, the nature of the hydrophobic core in globular proteins has been a central focus in the studies of protein folding, self-assembly, and conformation. The first experimental values for this effect on the individual amino acids were published by Tanford (1962) (Nozaki and Tanford, 1971; also see Tanford, 1980). Since that date the controversy over the relative magnitude of the hydrophobicity of each of the amino acid residues in globular proteins had produced a vast literature without consensus. Cornette *et al.* (1987) discuss 38 published hydrophobicity scales, which are compared for their ability to identify the characteristic period of α helices, and also computed an optimum scale for this purpose using a new Eigenvector method. Scales have previously been compared and discussed (Rose *et al.*, 1985; Meirovitch *et al.*, 1980).

Several papers are discussed to demonstrate the differences in principle involved in developing these scales. Such scales can be classified as solution measurements, empirical calculations, or some combination of both. Solution scales are based on distribution coefficients between an aqueous phase and a suitably chosen organic phase, whereas empirical scales are based on partitioning between the solvent-accessible surface and the buried interior in proteins of known structure. Significant differences exist between scales. Residues that are strongly hydrophobic on one scale appear to be strongly hydrophilic on another scale [e.g., Try and Tyr were found to be hydrophobic by Nozaki and Tanford (1971) but found to be hydrophilic by Wolfenden *et al.* (Wolfenden *et al.*, 1981; Wolfenden, 1983)]. Five significant and different approaches are briefly discussed.

It has been pointed out that the hydrophobic indices of the amino acids have a poor correlation with the extent to which the residues are buried in the native folded-protein matrix (Chothia, 1976). Although the hydrophobic index is a measure of the preference of the nonpolar environment by a residue, it does not necessarily reflect to the same extent the environment in protein crystals. Ponnuswamy and co-workers (Manavalan and Ponnuswamy, 1977, 1978; Ponnuswamy *et al.*, 1980) have defined a parameter called "the surrounding hydrophobicity" for residues, which reflects realistically the preferred nonpolar environment of the residue in protein crystals. This environment is represented by a sphere of 8 Å radius around the residue. This preferred environment is obtained by each residue associating itself with a set of specific surrounding residues, and this requirement of each residue in the protein molecule drives the linear chain to fold into a specific compact globular shape. The hydrophobicity of a residue in a native protein is the same as the product of the surrounding residues

and their hydrophobic indices (using either values of Tanford, 1962, or Jones, 1975). This scale provides valuable information with regard to hydrophobic domains, nucleation sites, surface domains, loop sites, and the spatial positions of residues in protein molecules.

Wolfenden (Wolfenden, 1983; Wolfenden *et al.*, 1979, 1981, 1983) determined the equilibria of distribution of amino acid side chains between their dilute aqueous solutions and the vapor phase at 25° by dynamic vapor pressure measurements. The resulting scale of "hydration potentials" or free energies of transfer from the vapor phase to neutral aqueous solutions spans a range of ~22 kcal/mole. These hydration potentials are closely correlated with the relative tendencies of various amino acids to appear at the surface of globular proteins. Guy (1985) calculated the energies required to transfer amino acid side chains from water to less polar environments, and these were compared with several statistical analyses of residue distributions in soluble proteins. He found that an analysis that divides proteins into layers parallel with their surface was more informative than those that simply classify each residue as exposed or buried. Most residues were found to be distributed as a function of the distances from the protein–water interface in a manner consistent with partition energies calculated from partitioning of amino acids between water and octanol phases rather than from solubilities of amino acids in water ethanol and methanol.

Kyte and Doolittle (1982) described a computer program that progressively evaluates the hydrophilicity and hydrophobicity of a protein along with its amino acid sequence. The hydrophobic properties of each of the 20 amino acid side chains are taken into consideration. The scale is based on an amalgam of experimental observations derived from the literature. They draw attention to the fact that the extent to which residues are buried depends not only on the strict hydrophobicity but also on steric effects that determine packing between the secondary structure in the crowded interior of the macromolecule. An excellent discussion of the literature bearing on the hydropathicity of proteins can be found in this paper.

Eisenberg *et al.* (1982a) analyzed the structure of proteins in terms of the "hydrophobic moments" (1) of the entire molecule and (2) of the segment of secondary structure that makes up the polypeptide chain. The zeroeth moment is defined as the sum of the hydrophobicities of the amino acid residues (an analogue of the net charge of a cluster of point charges), and the first moment, or hydrophobic dipole moment, is the analogue of the electric dipole moment of a cluster of charges. The hydrophobic dipole can be used to measure the amphilicity of the structure and can be applied to relating the function and secondary structure of a region to its amino acid sequence (Eisenberg *et al.*, 1982b). It can also be useful in the analysis of interactions of a segment or domain of a protein with neighboring regions in the protein (the "hydrophobic field") and to detect the periodicity in protein hydrophobicity (Eisenberg *et al.*, 1984a,b) (see Chapter 16 by Eisenberg *et al.*).

Direct measurements have been made, on long water-soluble double-chained alkylammonium acetate surfactants adsorbed onto sheets of muscovite mica, of the forces between electrically neutral planar hydrophobic surfaces in aqueous solution (Pashley *et al.*, 1985). Such forces reflect interactions caused by surface-induced water structure and are long-ranged, with an exponential decay length of about 1.4 nm. Over 0 to 8 nm, the forces are 10 to 100 times stronger that the van der Waals forces that would operate in the absence of any surface-induced order in water. It was concluded that there is available to biological systems a hierarchy of attractive forces that operate between hydrophobic moieties. These forces depend on the dimensions and geometry of the surfaces and are much stronger, longer-ranged, and more variable than classical colloid science previously indicated.

Rose *et al.* (1985) derived two new scales that are based on accessibility to solvent for residues within proteins of known structure. These two scales measure two quantities that can be distinguished: (1) the area lost when a residue is transferred from a defined standard state to

a folded protein—the area a residue buries on folding is proportional to the conformational free energy ΔG_{conf} (Richards, 1977)—and (2) the fractional accessibility of a residue, defined as its mean accessible area in protein molecules divided by the standard-state area. The fractional accessibility is an intrinsic measure of hydrophobicity. These results revealed a strong correlation between hydrophobicity and the surface area residues buried on folding (see Chapter 15 by Rose and Dworkin). Wolfenden *et al.* (Wolfenden *et al.*, 1979, 1981, 1983; Wolfenden, 1983) found that ΔG^0_h, the free energy of transfer of side-chain analogues between water and the dilute vapor phase, correlated with the empirical tendency of residues to be buried within proteins. Conversely, Chothia (1976, 1984) and Janin (1979), in an empirical analysis, and Wolfenden *et al.* (1981, 1983) noted a lack of correlation between the degree to which residues are buried and the Nozaki–Tanford ΔG_t^0.

C. Minimum Energy Calculations

At present, it is still impossible to predict the three-dimensional structure of a protein by the minimization of the free energy of an all-atom representation. There have been numerous quantum-mechanical studies of peptide systems. Some of these have been *ab initio* studies (e.g., Shipman and Christoffersen, 1973, and references herein), but the majority have used more approximate methods, mostly semiempirical (Scheraga, 1985). Many investigators have developed potential functions to describe the energy surface of a polypeptide chain. However, to date energy-minimization schemes have failed to predict chain folding accurately (Hagler and Honig, 1978; Cohen and Sternberg, 1980a,b). Scheraga and co-workers were among the first to tackle this exceedingly complex task (see review by Nemethy, 1974). The main problems involved are (1) the large number of variables and interaction energy terms, (2) uncertainties about the function form of the potential energy terms, (3) the existence of many conformations corresponding to a minimum in potential energy, (4) interaction of the protein with the solvent, and (5) vibrations and other free energy contributions. A basic question that can be asked is whether the native conformation of a protein molecule is its thermodynamically most stable state. Many investigators recognized that the kinetic question of the folding pathway was significantly more complex than the thermodynamic issue of predicting the optimal folded conformation. One major problem in this approach is that the contributions to the free energy cannot be modeled adequately. Many workers (e.g., Weiner *et al.*, 1984) have developed a new force field simulation for proteins. They have used the powerful Cartesian-coordinate energy refinement of Lifson and Warshel (1968) and developed empirical force fields within this context. The related force field for proteins was similar and contained minor modifications of the parameters used by Gelin and Karplus (1979). The most important changes concerned the explicit inclusion of H-bonding hydrogens parameters and the use of partial charges taken from Mulliken populations of *ab initio* calculations. They focused on ϕ, ψ maps of glycyl and alanyl dipeptides, hydrogen-bonding groups, and energy-refinement calculations on insulin. They emphasize that the single crudest aspect of this work in the application of the force field is the way solvation effects are modeled, and this is the area that requires the most refinement in the near future.

Another approach has been to predict the tertiary structure on the basis of sequence homology and regularize a predicted structure by energy calculations. Lewis and Scheraga (1971) carried this out for α-lactalbumin. Following the first energy calculations on known protein conformations (Levitt and Lifson, 1969), more attention has been given to the energy refinement of x-ray coordinates (e.g., Levitt, 1974; Burgess and Scheraga, 1975), but with little more success. Further work (Levitt, 1976; Levitt and Warshel, 1975) introduced a simplified representation of protein conformation for rapid simulation of protein folding using the concept of time-averaged forces, but this did not produce significant improvements.

Conformational energy calculations were used to analyze the interactions of structural substructures in subtilisin BPN (Honig *et al.*, 1976). These substructures are kept fixed or "rigid" so that the only variables in the calculations are the backbone segments that separate them. The flexible segments are assumed to be free turns. By using this representation of the protein, it is possible to predict both a likely order of events along the folding pathway and preferred modes of conformational changes of the native protein. These results suggest an approach to the folding problem based on the piecemeal formation of tertiary structure from smaller prefolded fragments. The major improvement of the Robson and Osguthorpe (1979) procedure over previous methods was to retain a more realistic and complete representation of the protein backbone and alternately to reduce the number of variables by coupling their behavior. They attempted to satisfy the criteria essential for folding as discussed by Nemethy and Scheraga (1977). This "preliminary investigation" left much work to be done before such computer simulation could become a realistic approach to protein folding. Finney *et al.* (1980), using detailed hydrogen bonding, surface exposure, internal environment, and solvent interaction calculations, in conjunction with data from quantum mechanical hydrogen-bonding studies, estimated various contributions to the free energy of folding. Their conclusion stated that "a picture emerges of globular proteins as extremely well-fitting jigsaw puzzles, in which no single driving force dominates the marginal stability of the native conformation. Rather, the folded structure is seen as the result of a complex global minimization of several strongly interacting driving forces." The necessity to maintain a very efficient internal hydrogen bonding and the role of solvent as a hydrogen-bond sink are stressed as strong constraints on the (incomplete) maximization of hydrophobic effects.

Robson and Platt (1986) reappraised the interatomic potential functions for protein structure calculations, using the all-atom approximation (except CH, CH_2, and CH_3, which were treated as "united atoms"). This produced a more efficient and robust folding algorithm. The potentials were calibrated for the rigid geometry approximation, since use of fixed standard bond lengths and valence angles (and fixed transplanar peptide groups) reduces the number of conformational variables and saves a great deal of computer time. These potentials do not generally give the nativelike form as the least energy form, in common with other similar methods. A statistical approach to the calculation of conformation of proteins was published by Crippen (1977a,b) and applied to the reduced pancreatic trypsin inhibitor. The theory applies to both thermodynamically and kinetically determined processes, so that not only can the equilibrium result be calculated but also the time course of folding. The results look promising.

Bash *et al.* (1987) have applied recent advances in statistical mechanical theory and molecular dynamics to the understanding of the role of solvation in determining molecular properties. The free energies of solvation of all the chemical classes of amino acids side chains were calculated using supercomputers. The effect of a site-specific mutation on the stability of trypsin was predicted, which yielded results in agreement with available experiments.

IV. APPROACHES TO PROTEIN CONFORMATION

The x-ray crystallographic study of proteins has opened up the opportunity to evaluate protein conformation from many points of view. Each of these opens a different window to view the architecture of protein structure. These various windows are briefly reviewed.

A. Solvent Accessibility

The counterparts of the buried hydrophobic areas are those that are accessible to solvent, usually water. Lee and Richards (1971) developed a program that permits the accessibility of

atoms, or groups of atoms, to solvent or solute molecules of specified size to be quantitatively assessed. This accessibility was found to be proportional to the surface area. This approach can also define internal cavities. It was found that about 40–50% of the surface area of each protein is occupied by nonpolar atoms. The numerical results are sensitive to the choice of the van der Waals radii of the various groups. For the atoms in ribonuclease S, lysozyme, and myoglobin, the average change in accessibility for the atoms in going from a hypothetical chain to the folded conformation of the native protein is about a factor of 3. Another alternative would be to consider the contact surface, those parts of the molecular van der Waals surface that can actually be in contact with a probe examining the surface (Richards, 1977). Richards and co-workers (Richards, 1974, 1977; Richmond and Richards, 1978; Richmond, 1984) have examined the total volume, group volume distributions, and packing densities of proteins. Richmond and Richards (1978) discussed the packing of α helices in sperm whale myoblobin and proposed an algorithm for picking potentially strong helix–helix interaction sites in peptides of known sequence. Richmond (1984) has expanded on this idea. A computer program, using the equations for area, has been tested and has had limited application to the docking of protein α helices. Chothia (1974, 1975, 1976) found that the accessible surface area is proportional to the hydrophobicities of residues in the protein. The loss of accessible surface area by monomeric proteins on folding is proportional to the hydrophobic energy and is simply proportional to the two-thirds power of their molecular weight. Wodak and Janin (1980) proposed an analytical substitute to the protein surface area that is accessible to solvent. A statistical approach leads to an expression of accessible surface areas as a function of distance between pairs of atoms or of residues in the protein structure, assuming only that these atoms or residues are randomly distributed in space and not penetrating each other. More recently, Zehfus *et al.* (1985) published equations that approximate the accessible area of a continuous protein segment using the surface area of an inertial ellipsoid that approximates the molecular volume from the number of nonhydrogen atoms in a segment.

B. Packing of Residues

Solvent accessibility is the result of failure to achieve maximum close packing (Richards, 1974) of the atoms in the folding of the polypeptide chain. Although similar in principle to solvent accessibility, its approach is based on a different outlook. The result of maximum packing is to produce a center of residues, which has been classified as the hydrophobic core, that has maximum shielding from solvent (Richards, 1977; Chothia, 1975). This core consists mainly of nonpolar atoms or polar atoms that form either hydrogen bonds or salt bridges. Thus, the degree to which ideal packing occurs determines the degree of relative stability of the particular protein. Crippen and Kuntz (1978) surveyed the atom packing in high-resolution x-ray crystal structures of 21 proteins and concluded that the atom density around a given central atom is determined primarily by its covalently bonded neighbors and proximity to the surface of the protein. Long-range hydrophobic, hydrogen-bonding, and electrostatic interactions are strictly of secondary importance.

The interior of globular proteins has very significant density inhomogeneities on a scale of 100–1000 $Å^3$. The interior densities range from less than 0.5 g/cm^3 to over 3 g/cm^3. The low local densities are primarily associated with clusters of nonpolar side chains, and the high-local-density regions arise from the protein backbone secondary structures: helices and β sheets (Crippen and Kuntz, 1978). Lesk and Rose (1981) developed a method to identify all compact, contiguous-chain structural units in a globular protein from x-ray coordinates. These units were then used to describe a complete set of hierarchic folding pathways for the molecule. The analysis showed that the larger units are combinations of smaller units, giving rise to

structural hierarchy ranging from the whole protein monomer through supersecondary structures down to individual helices and strands. Thus, there is hierarchic condensation. In this model, neighboring hydrophobic chain sites interact to form folding clusters, with further stepwise cluster association giving rise to a population of folding intermediates.

C. Distance Geometry

The mathematics of distance geometry constitutes the basis of a group of algorithms for revealing the structural consequences of diverse forms of information about the macromolecule's conformation. An excellent review of this approach was published by Havel *et al.* (1983). The article presents the basic theorems of distance geometry in Euclidean space and gives formal proofs of the correctness and, where possible, of the complexity of the algorithms. Crippen (1977a,b) originally conceived the idea as a means of circumventing the longstanding local minimum problem of molecular minimization, but since then it has led to the development of a data structure for representing classes of related conformations consistent with experimental, energetic, and functional information that is compact and succinct.

Rackovsky and Goldstein (1987) extended a previous differential geometric analysis of the conformational properties of the various amino acids to study their influence on folding over a larger backbone interval. In addition, statistical effects associated with the variation in the number of individual amino acids in the data base were treated in greater detail, using a simulation method. It was found that the amino acids could be divided into three groups on the basis of their conformational influence over four-C^{α} units in the interval $i - 6 \leq j \leq i + 6$. Group Ia is composed of seven amino acids (His, Leu, Ala, Met, Lys, Glu, Ile) that encourage the formation of a right-handed α-helical structure. Group Ib (Glu, Phe, Trp, Val, Asp) is composed of amino acids with some helix-forming tendency but that also showed a positive extended strand formation tendency. They therefore act as a bridge between group Ia and group II (Cys, Gly, Asn, Pro, Arg, Ser, Thr, Tyr), which contains amino acids that encourage the formation of extended structure and bends. It was shown that, in general, such influences extend further in the N-terminal direction than in the C-terminal direction. Insofar as comparison is possible, these results correlate with those of Robson and Pain (Robson, 1974; Robson and Pain, 1974a–c) and Robson and Suzuki (1976) and are complementary to earlier work of Maxfield and Scheraga (1976, 1979). In general, they also agree with the propensity parameters of Chou and Fasman (1974a,b).

D. Amino Acid Physicochemical Properties

A large number of physical-chemical properties, manifest in the amino acid side chains, have been thoroughly examined by many investigators. Attempts have been made to correlate these properties with their relatedness between protein sequences. Sneath (1966) sought the relationship between chemical structure and biological function in peptide hormones using a new measure of difference in chemical structure. A large number of physicochemical properties (32) were investigated, and a principal component analysis yielded four vectors, which represent major composite chemical factors. Zimmerman *et al.* (1968) performed three different but related comprehensive statistical analyses of amino acid sequences in proteins. The goal was to find evidence of significant sequence structure related to a purely random arrangement of amino acid residues and to attempt to relate any significant structure uncovered to secondary and/or tertiary conformation of the protein. A continuous physical scale property for six scales was employed. These were bulkiness, polarity, R_F, p*I*, p*K*, and hydrophobicity. No striking conclusions could be made. Grantham (1974) developed a formula for the difference

between amino acids that combines the properties that correlated best with protein residue substitution frequencies. The properties that yielded the best correlations were composition, polarity, and molecular volume.

Jones (1975) replaced the amino acid sequence of a protein by a numerical sequence of values representing a physical or chemical property of the amino acids, and the resulting numerical sequence was amenable to autocorrelation analysis. Similarly, certain geometrical parameters, calculated from the three-dimensional structure of a protein to form a configurational series, can be analyzed by cross-correlation techniques. On the basis of ten proteins, it was found that the hydrophobicity of an amino acid residue in a protein influences the orientation angle of the amino acid side chain. Padlan (1977) used a technique to compare the dissimilarity of physicochemical properties of the amino acids in immunoglobulin sequences. Exterior residues showed greater structural variability than the interior residues. The dissimilarities were taken from the work of Sneath (1966) and Grantham (1974). Yockey (1977) utilized Grantham's (1974) method of representing amino acid residues to present a prescription that predicts all functionally equivalent residues expected at a given site in a protein sequence if at least two such residues are known. The prescription established a conceptual framework wherein the validity of Grantham's hypothesis can be tested objectively and, if necessary, generalized.

Levitt (1978) analyzed the secondary structure of 50 different globular proteins to give the frequency of occurrence of the 20 naturally occurring amino acids in the α helix, β sheet, and reverse turn. These preferences correlate well with the chemical structure and stereochemistry of the particular amino acid. The rules that emerged from this study can be summarized as follows: bulky amino acids, namely, those that are branched at the β carbon or have a large aromatic side chain, prefer β sheet. The shorter, polar side chains prefer reverse turns, as do Gly and Pro, the special side chains. All other side chains prefer α helix except Arg, which has no preference. The polar side chains with hydroxyl groups disrupt α helix; the other polar side chains disrupt reverse turns. This conclusion agrees well with previous analyses of a similar nature (Nagano, 1973; Chou and Fasman, 1974a,b; Robson and Suzuki, 1976; Chou and Fasman, 1977a–c; Chou *et al.*, 1975). Similar conclusions were also previously drawn from studies on poly-α-amino acids (for review see Blout, 1962; Fasman, 1987). Wertz and Scheraga (1978) examined the x-ray structures of 20 proteins to determine which amino acid residues reside in the inside or outside of the molecule and to assign a conformational state. The data confirm that polar groups are generally found on the outside of proteins and that nonpolar residues are generally found on the inside. Seven amino acids (Ala, Arg, Cys, His, Pro, Ser, Tyr) have inside/outside preferences that are not consistent with their usual assignment as either polar or nonpolar residues. It is suggested that differences in entropy play an important part in the inside/outside preferences of backbone structures.

There are generally significant changes in the conformational preferences of the residues in going from the inside to the outside of proteins; environmental (rather than local) solute–solvent interactions seem to be the predominant cause of these changes in conformational preferences. Kubota *et al.* (1981) examined internal homologies in an amino acid sequence and in amino acid sequences of two different proteins using correlation coefficients calculated from the sequences when residues are replaced by various quantitative properties such as hydrophobicity. These quantities inherent in amino acids were those previously discussed by Zimmerman *et al.* (1968). The sequences of α-tropomyosin, calmodulin, troponin C, and L2 light chain of myosin were evaluated in this manner.

Argos and Palau (1982) probed the compositional distribution of the 20 amino acids for particular positions within the secondary structures (α helices, β strands, and turns) in a 44-protein sample. Correlation coefficients between positional composition of the amino acids

and various of their physicochemical characteristics indicated considerable asymmetry in the properties of the residues comprising regions within and adjacent to secondary structures, modes of helix formation, physical parameters most sensitive to the buriedness of residues in β strands, and possible improvements in the accuracy of secondary structural prediction methodologies. The physical parameters used were bulkiness and polarity (Jones, 1975), hydrophobic index (Nozaki and Tanford, 1971), conformational propensity (Levitt, 1978), the "residue surrounding" hydrophobicity (Manavalan and Ponnuswamy, 1978), hydration potential (Wolfenden *et al.*, 1979), and free energy transfer (Janin, 1979). Taylor (1986b) classified the amino acid type based on a synthesis of physicochemical and mutation data. The major sets group the amino acids by size and hydrophobicity. These relationships were displayed as a Venn diagram, from which subsets are derived that include groups of amino acids likely to be conserved for similar structural reasons.

There have been several previous evaluations of amino acid mutations relative to evolution (e.g., Grantham, 1974; Sellers, 1974). The assessment of the structural effect of introducing a new amino acid into a known structure is often based on the likehood matrix of amino acid mutabilities derived by Dayhoff (1972, 1978). The prediction of secondary structure by evolutionary comparison was made by Crawford *et al.* (1987) for the α subunit of tryptophan synthase from ten different microorganisms. Both the Chou and Fasman (1974a,b) and Garnier *et al.* (1978) methods were applied, as well as profiles of hydropathy (Kyte and Doolittle, 1982) and chain flexibility values (Karplus and Schulz, 1985), to give a joint prediction. There was good agreement (1) among predicted β strands, maximal hydropathy, and minimal flexibility and (2) among predicted loops, great chain flexibility, and protein segments that accept insertions of various lengths of individual sequences.

Hellberg *et al.* (1986) predicted a set of bradykinin-potentiating potencies of pentapeptides with a variation in amino acid sequences. This was performed by varying three parameters per amino acid position. The variables were derived from a principal-component analysis of a property matrix for the 20 amino acids. The resulting structure descriptor describes the observed activity of the peptides to 97% by means of a multivariable partial least-squares (PLS) model. It was demonstrated that this quantitative structure–activity relationship (QSAR) can be used to predict the activity of new peptide analogues. The relationship is based on the same principle of quantitative analogy models that previously have been shown to apply to structure–activity relationships in organic chemistry, e.g., the Hammet (1970) relationship. Three scales are used to characterize each individual amino acid. These scales are derived by a principal-component analysis of a matrix of 20 properties of each of the amino acids. Vondervisz *et al.* (1986) studied the occurrence of all di- and tripeptide segments of proteins from a large data base containing 119,000 residues. It was found that the abundance of the amino acids does not determine the frequencies of the various di- and tripeptide segments. The pair-frequency distribution of amino acids is highly asymmetric, pairs formed from identical residues are generally preferred, and amino acids cannot be clustered on the basis of their first-neighbor preferences. These data indicate the existence of general short-range regularities in the primary structure of proteins. The consequences of these short-range regularities were studied by comparing Chou–Fasman parameters with analogous parameters determined from the results of conformational energy calculations (Zimmerman *et al.*, 1977) of single amino acids. This comparison showed that the Chou–Fasman parameters carry significant information about the environment of each amino acid. The success of Chou and Fasman's (1974a,b) prediction and the properties of the pair and triplet distributions of the amino acid residues suggests that every amino acid has a characteristic sequential residue environment in proteins.

Kelly and Holladay (1987) compared the scales of amino acid side-chain properties by conservation during evolution of four proteins. As the amino acid sequence of a given protein

changes along the phylogenetic tree, enough of the overall folding pattern must be conserved to ensure that the protein fulfills its biological function. Eighteen published scales that tabulate various side-chain properties were compared by computing the variance of each scale when applied to each of several protein families. The conservation of each scale of side-chain properties was examined for the 20,627 residues in 60 mammalian myoglobins, 31 mammalian ribonucleases, insulin A and B chains (29 sequences each), and 29 vertebrate and 28 plant cytochrome *c*s. Those scales that are the most highly conserved may well be the best prediction of protein folding patterns. The mean-area-buried scale (Rose *et al.*, 1985) and the optimized matching hydrophobicities scale (Sweet and Eisenberg, 1983) are more conserved than other scales. An additional result is the relatively poor conservation of the Chou–Fasman (1974a,b) secondary structure predictions.

V. PREDICTION OF THE SECONDARY STRUCTURE OF PROTEINS: α HELIX, β STRANDS, AND β TURN

The origins of the prediction of the secondary structure of proteins have been previously outlined (Section I). Over 20 different methods have been proposed in addition to variations of several of these. These predictive schemes mostly assume that the local sequence (short-range interactions) determines local structure. The following authors have proposed the main schemes representing significantly different approaches: Kotelchuck and Scheraga (1968a,b), Lewis *et al.* (1970), Ptitsyn and Finkelstein (1970a,b), Toitskii and Zav'yalov (1972), Kabat and Wu (1973a,b), Chou and Fasman (1974a,b, 1978a,b), Burgess *et al.* (1974), Lim (1974a,b), Robson and Suzuki (1976), Maxfield and Scheraga (1976), Nagano (1977), McLachlan (1977), Barkovsky and Bandarin (1979), and Barkovsky (1982).

These methods can be categorized into two broad classes. The empirical statistical methods use parameters obtained from the analysis of known sequences and structures (e.g., Chou and Fasman, 1974a,b). The second method is based on stereochemical criteria (e.g., Lim, 1974a,b). The most frequently used methods to date have been the empirical approaches of Chou and Fasman (1974a,b) and of Robson and co-workers (Garnier *et al.*, 1978) and the stereochemical method of Lim (1974a,b).

The Chou and Fasman (1974a,b) method has been widely used because of the simplicity of its application and the ease of understanding its premise. The statistical derivation of the propensity of residues to be in α helices, β strands, or turns (P_α, P_β, P_t) is straightforward. The relative magnitudes of these values assist in classifying residues as α formers, α indifferent, and α breakers. The β-strand residues are likewise categorized. The β-turn parameters (frequencies) are also used in a direct manner. There are rules for utilizing these parameters for nucleation, propagation, and termination of α helices and β sheets. A problem of interpretation of regions of equal potential for α helices and β strands has dismayed some users (Sternberg, 1983), and the lack of clear-cut rules has prevented programmers from writing computer algorithms that yield the same results. One factor, unfortunately, has been overlooked that gives this method an advantage over others. This ambiguity that arises in delineating sequences of equal potential for α and β structures is not a detrimental factor, as is so frequently pointed out, but rather gives insight into regions of sequence that can undergo conformational change. This has been illustrated for glucagon (Chou and Fasman, 1975) and for the preproparathyroid hormone (Rosenblatt *et al.*, 1980). The percentage accuracy of the method was stated to be 75–80%. The Chou and Fasman method is discussed in detail in Chapter 9.

The method of Robson and co-workers (Robson and Suzuki, 1976; Garnier *et al.*, 1978) is based on information theory and can be programmed easily and unambiguously. It considers the

effects of residues from positions $i - 8$ to $i + 8$ on the conformation of position i. The Garnier *et al.* (1978) method has recently been reevaluated (Gibrat *et al.*, 1987), and the validity of the approximations drawn from the theory examined. It was concluded that the existing data base does not allow evaluation of parameters required for an exact treatment of the problem. It was shown that the first-level approximation, involving single-residue parameters, is only marginally improved by an increase in the data base. The second-level approximation, involving pairs of residues, provides a better model. However, in this case, the data base is not big enough, and this method leads to parameters with deficiencies. This new version of the Garnier–Osguthorpe–Robson (GOR) method increases the accuracy of prediction by 7%, giving a 63% correctly predicted residues count for three states in 68 proteins, each protein to be predicted being removed from the data base and the parameters derived from the other proteins. If the protein to be predicted is kept in the data base, the accuracy goes up to 69.7%. The GOR method is discussed in detail in Chapter 10.

The method of Lim (1974a,b) is based on a stereochemical theory of globular protein secondary structure. This approach considers the "architectural" principles of packing of polypeptide chains and the interactions of proteins with water molecules. Thus, this procedure takes into account long-range interactions, as was first applied by Schiffer and Edmundson (1967). The most salient structural features utilized are compactness of form, the presence of a tightly packed hydrophobic core (cores), and a polar shell. These structural features impose the following requirements for possible conformations of any region of the polypeptide chain:

> (1a) All the helical, β-structural and irregular regions must be attached by non-covalent interactions to the main part of the globule. (1b) The linear dimensions of regular and irregular regions cannot exceed the linear dimensions of the globule. The following requirements result from the second structural principle: (2a) The overwhelming majority of massive hydrophobic side groups must be fully or partially immersed in hydrophobic cores, while hydrophilic groups must not penetrate into these hydrophobic cores; NH and CO groups of the backbone also must not penetrate into these cores without formation of a hydrogen bond. (2b) The conformation of the backbone of the region having hydrophobic side groups must be compatible with the tight packing of these hydrophobic groups in the hydrophobic core or cores. (2c) The hydrophilic shell must at least shield each hydrophobic core from water molecules; hydrophilic groups must, if possible, form salt and hydrogen bonds which together with hydrophobic interactions will attach the separate parts of the globule together (Lim, 1974a).

Lim (1974b) published an algorithm that was stated to predict accurately ~80% and 85% of the α-helical and β-structural regions, respectively, for 25 proteins.

Pongor and Szaley (1985) have described a quantitative procedure for the comparison of predicted secondary structures of homologous proteins. The predictive procedures of Chou and Fasman (1978a) and Garnier *et al.* (1978) and the hydrophilicity values of Hopp and Woods (1981) show that correlation coefficients of structural profiles can be used to describe similarity in secondary structures. The method is potentially useful to describe evolutionary changes in protein secondary structure as well as in the design of peptide analogues.

A. β Turns

A proposed mechanism for folding of polypeptide chains in proteins was put forth by Lewis *et al.* (1971). It was suggested that secondary structures (e.g., α helix or other ordered regions) could be stabilized only by long-range interactions arising from the proximity of two such ordered regions. These regions would be brought near each other by the directing influence of certain other amino acid sequences that have a high probability of forming β bends or variants thereof and also on the basis of short-range interactions. An analysis was made, on

three proteins of known structure, of the tendency of various amino acids to occur in β bends, and it was found that it was possible to predict the regions of the chain in which a β bend will occur with a high degree of reliability (~80%). The β bends were defined in the manner previously described by Venkatachalan (1968). Dickerson *et al.* (1971) determined the structure of horse heart ferricytochrome *c* to a resolution of 2.8 Å. They carefully analyzed the folding of the polypeptide chain and carefully delineated the β turns (calling them 3_{10} bends). The 3_{10} bend occurred six times in cytochrome *c*. Two variants of the 3_{10} bend, which were previously classified by Venkatachalan (1968) as type I and type II, were found: three type I and three type II bends, with Gly in the third α-carbon position in the latter.

Kuntz (1972) developed simple computer routines for locating regions in which the peptide backbone of a globular protein folds back on itself. These programs were used to locate "turns" in carboxypeptidase and α-chymotrypsin. The following generalizations were made: (1) turns occur on the surface of the protein; (2) turn segments are less hydrophobic than the protein as a whole; (3) uninterrupted sequences of three to eight hydrophilic residues are frequently associated with folding of the peptide chain. It was found that Ser, Thr, Gly, Pro, Gln, and Asn frequencies were about 20% enhanced in turns. Bunting *et al.* (1972) predicted the β bends in the immunoglobulin light and heavy chains using the method of Lewis *et al.* (1971). They concluded "that there was an overall common design in light chains, and in heavy chains, which, in the absence of significant quantities of other recognizable conformational features, may be at least partially influenced by the position of β-bends."

It is now known that immunoglobulins contain a large amount of β-sheet structure, and the β strands are frequently connected by β turns. Thus, the above conclusions have been verified by x-ray analysis. Esipova and Tumanyan (1972) suggested a model for the tertiary structure of proteins whereby relatively straight segments are interrupted by a folding point between them. Turns in proteins are regarded either as flexible hinge regions that permit the chain to bend in this region because of forces acting outside the regions of folding or as rigid segments that alter the course of the polypeptide chain through local interactions in the bending region. Folding regions were determined for seven proteins. These globular proteins were analyzed by means of stereo projections and special curvature functions. The suggested model encompasses not only regions with the secondary structure of a protein globule but also those regions that do not enter into secondary structure and previously have been negatively described as "unordered." It is primarily these latter segments that in the different portions of the protein determine the most important properties of the tertiary structure. It was shown that certain amino acids—Gly, Asn, Ser, etc.—have a specific significance in these regions of transition, forming hydrogen bonds between their side groups and the backbone of the chain. These transitions agree with those found by Lewis *et al.* (1971). Lewis *et al.* (1973a) analyzed eight proteins for the presence of chain reversals and found 135 bends. Of these, 129 belong to a set of ten types.

Conformational energy calculations were carried out on the pentapeptide N-acetyl-N′-methyl-Ala-Ala-Ala-Ala-amide in order to determine the backbone conformations of minimum energy for bend and nonbend structures. The conformations of lowest energy did not correspond to the conformations obtained from a minimization starting from the x-ray structural data, although in all three cases, one type of bend or another was found to be of lowest energy compared to other conformations studied. Crawford *et al.* (1973) defined a reverse turn as a tertiary conformation in globular proteins and defined it in terms of the dihedral angles, the C^{α}_1–C^{α}_4 distance (less than 5.7 Å), and the O_1....H-N_4 hydrogen-bond distance (less than 3.2 Å). In seven proteins, 125 examples of turns were found, comprising 33% of the amino acids in these proteins as compared with 34% of the residues forming helices and only 17% forming β sheets. The amino acid compositions of turns, helices, and β sheets were analyzed in detail.

They found Asn and Gly mainly in turns, Pro in turns (and at the beginning of helices), and Glu in helices. In these turns, 19% of Asp was found in the first position, 33% of Pro in the second position, 24% of Asn in the third position, and 26% of Trp in the fourth position.

Chou and Fasman's (1974b) algorithm for the prediction of the secondary structure of proteins predicted α helices, β sheets on the basis of the conformational parameters P_α and P_β (Chou and Fasman, 1974a). In order to compare the β-turn potentials for all the 20 amino acids, with their α and β potentials, the conformational parameter for the β turns, $P_t = f_t/\langle f \rangle$ ($f_t = n_t/n$, where n_t and n are, respectively, the total occurrence of each residue in the β turns and in all the regions of the 12 proteins; $\langle f_t \rangle = n_t/n$ is the frequency of all residues in the β-turn regions), was obtained in the same manner as P_α and P_β (Chou and Fasman, 1974a) by the process of normalization (see Chapter 9). These turns were chosen in part on stereo diagrams inasmuch as atomic coordinates were not available. There is almost an inverse relationship between the β-turn potential and the α potential for the 20 amino acids. The relative probability that a tetrapeptide will form a β turn (Lewis *et al.*, 1971), $p_t = f_i \times f_{i+1} \times f_{i+2} \times f_{i+3}$, where f_i, f_{i+1}, f_{i+2}, and f_{i+3} are, respectively, the frequency of occurrence for a certain residue at the first, second, third, and fourth positions of a β turn. The cutoff value of $p_t = 0.5 \times 10^{-4}$ was found to be a reasonable value in predicting the β turns of the 12 proteins studied herein.

Burgess *et al.* (1974) described methodology for describing a discrete number of conformational states of amino acids in proteins and used these to examine the relative importance of short-range, medium-range, and long-range interactions in light proteins of known structure. A prediction algorithm that assigns four states to each residue of a protein chain (α helix, extended structures, bend, and coil) was developed from consideration of both short- and medium-range interactions and was applied to 13 proteins of known structure. First the frequencies of occurrence of the various conformations of each of the 20 amino acids in several proteins are examined and represented on ϕ,ψ conformational maps. Second, these empirical frequencies are compared to those deduced (Lewis *et al.*, 1970) from the statistical weights of various conformations of these residues (obtained from conformational energy calculations on the N-acetyl-N′-methylamides of each residue; Lewis *et al.*, 1973b). Third, the empirical frequency data are used to develop an algorithm to predict simultaneously the location of α-helical, extended, bend, and coil regions of a protein, taking into account the conformational states of four residues on each side of the given one (the prediction of β turns was not as good as those for the α helix). The authors state that their method is more reliable than other predictive schemes, but in relative trials (Schulz and Schirmer, 1974) this was not borne out.

Nishikawa *et al.* (1974) examined the low-energy conformations of two dipeptides using "empirical" energy calculations and the virtual-bond method. The complete conformational space of the dipeptide N-acetyl-N′-methyl-L-alanineamide was searched systematically, and all conformations of minimum energy were found. The low-energy conformation of the alanine dipeptide, which closedly approximates a type II bend, does not correspond to a combination of single-residue minima. Conformations similar to type I bends are not found to be of minimum energy for either the glycine or alanine dipeptide. Chou and Fasman (1977b) expanded their sample and utilized the x-ray atomic coordinates from 29 proteins of known sequence and structure to elucidate 459 β turns in regions of chain reversal. Tetrapeptides whose C^α_i–C^α_{i+3} distances were below 7 Å and not in a helical region were characterized as β turns. In addition, β turns were considered to have hydrogen bonding if their computed $O_{(i)}$–$N_{(i+3)}$ distances were ≤3.5 Å. The torsion angles of 26 proteins containing 421 β turns were classified in 11 bend types based on (ϕ,ψ) dihedral angles of the $i + 1$ and $i + 2$ bend residues. The average frequency of β turns was found to be 32% as compared to 38% helices and 20% β sheets.

The relative frequencies for all the amino acids in the four positions were established (see Tables XI, XII, and XIII, Chapter 9). Residues with the highest β-turn potential in all four positions are Pro, Gly, Asn, Asp, and Ser, with the most hydrophobic residues (i.e., Val, Ile, and Leu) showing the lowest bend potential. An environmental analysis of β-turn neighbors shows that reverse chain folding is stabilized by antiparallel β sheets as well as helix–helix and α–β interactions. The β-turn potential at the 12 positions adjacent to and including the bend were plotted for the 20 amino acids and showed dramatic positional preference, which may be classified according to the nature of the side chains. Eight type VI bends were found with a *cis*-Pro at the third position. Chou and Fasman (1979a) then presented an automated computer prediction of the chain-reversal regions in globular proteins, using the bend frequencies and β-turn conformational parameters (P_t) determined from 408 β turns in 29 proteins calculated from x-ray atomic coordinates. The probability of bend occurrence at residue i is $p_t = f_i \times f_{i+1} \times f_{i+2} \times f_{i+3}$, with the average bend probability $\langle p_t \rangle = 0.55 \times 10^{-4}$. Therefore, tetrapeptides with $p_t > 0.75 \times 10^{-4}$ ($\simeq 1.5 \times \langle p_t \rangle$) as well as at $\langle P_t \rangle > 1.00$ and $\langle P_\alpha \rangle < \langle P_t \rangle > \langle P_\beta \rangle$ were selected by the computer as probable bends. When adjacent probable bends or overlapping bends occur, the tetrapeptide with the higher p_t value is predicted as the β turn. The percentage of bend and nonbend residues predicted correctly for 29 proteins by this computer algorithm is 70%, whereas 78% of β turns were localized correctly within ±2 residues. The average β-turn content in the 20 proteins is 32%, with helical proteins having fewer bends (17%) than the β-sheet proteins (41%). The accuracy of β-turn prediction using the probability of bend occurrence at residue i obtained from the frequencies of 2, 4, 8, and 12 residues were investigated, and very little change occurred beyond four residues.

Tanaka and Scheraga (1976d) developed a statistical mechanical treatment of protein conformation wherein chain-reversal regions were predicted in 23 proteins after predictions of helical and β-sheet regions. Their percentage of bends localized (±2 residues) was 59%. Lenstra (1977) has compared the accuracy of bend predictions in 25 proteins according to the computer methods of Nagano (1977) and Argos *et al.* (1976) and obtained correlating coefficients of $C_t = 0.40$ and 0.44, respectively. Although the automatic bend prediction (Chou and Fasman, 1979a) using four residues has a smaller correlation coefficient, $C_t = 0.38$, its prediction of turns is 70.4%

Rose and Seltzer (1977) presented an algorithm to identify peptide chain turns from x-ray-elucidated coordinate data. This algorithm uses only the C^α coordinate for every residue in the protein. No other information is required, and notions about hydrogen bonding at these loci are irrelevant to the geometric nature of the argument. The radius of curvature for each amino acid along the chain is calculated, and this is the basis of detecting turns. A turn corresponds to a locus where the chain direction vector is changing rapidly and the value of the radius of curvature is at a local minimum. The results are compared to data of Kuntz (1972), Crawford *et al.* (1973), and Lewis *et al.* (1971). Both Kuntz (1972) and Rose and Seltzer (1977) use visual parsing to detect β turns, and their results are quite similar. The authors stated that their algorithm yields a more comprehensive set of turns than the other three. Because the sample is relatively small, it is difficult to evaluate accurately these different sets of results.

Zimmerman and Scheraga (1977) calculated probabilities of bend formation in 47 amino acid sequences of the N-acetyl-N′-methylamide dipeptides with a statistical mechanical analysis using empirical conformational energies and compared these results with the fraction of bends formed in the same 47 dipeptide sequences in the x-ray structures of 20 globular proteins. Agreement was found for 26 dipeptides, suggesting that for those particular dipeptide sequences, local interactions dominate over long-range interactions in determining conformational preferences. Rose and Wetlaufer (1977), using the algorithm of Rose and Seltzer (1977) for determining turns, showed that the number of turns in a protein is a linear function of the

number of amino acid residues in the protein. In this sequence-dependent model, turns are a linear function of the molecular weight, whereas in a shape-dependent model, they are a function of the two-thirds power of the molecular weight. The number of turns (T) is a linear function of the number of amino acids (R) as given by $T = 0.125R + 2.28$.

Rose (1978) hypothesized that turns occur at those sites in the polypeptide chain where the hydrophobicity is at a local minimum. The measure of hydrophobicity of the amino acid side chains is taken to be the Nozaki and Tanford (1971) free energy of transfer from water to an organic solvent. One problem arises because one or more local minima in hydrophobicity are always found to be associated with helical secondary structure. An agreement of 78% between measure and predicted turns was claimed. Conformational energy calculations were carried out on the two terminally blocked tetrapeptides N-acetyl-Thr-Asp-Gly-Lys-N′-methylamide and N-acetyl-Ala-Asp-Gly-N′-methylamide (Simon *et al.*, 1978). The latter occurs as a bend at residues 94–97 in staphylococcal nuclease. Several groups of low-energy conformations were found. They were compactly folded structures, but they differed from the "standard" chain reversals. One group, containing Thr peptides, was stabilized by a network of hydrogen bonds involving polar atoms of both the backbone and side chains of Thr, Asp, and Lys.

Anderson *et al.* (1979) optimized energies of seven β bends, repeating C5 and C7 and right- and left-handed α-helical conformations for each of eight tetrapeptides have been computed using empirical methods. Eight tetramers were selected: four helix-forming sequences with hydrophobic residues such as Val, Leu, Ile, and Trp and four helix-breaking sequences with hydrophilic residues such as Asp, Asn, and Ser, as determined by their frequency of occurrence in β turns in proteins. Analysis of the optimized conformations with energies ≤2.1 kcal/mole from absolute minimum energy conformer for each tetramer revealed a correlation between low-energy conformations and those predicted from observed protein structure. These results indicated that energy calculations, performed with the Empirical Conformational Energy Program for Peptides (ECEPP) developed by Scheraga and co-workers (Momany *et al.*, 1975), on small peptide fragments may be useful in predicting protein structure. From the bend frequencies based on 29 proteins (Chou and Fasman, 1979a), the β-turn probability profiles were calculated for three sets of proteins: ten mammalian proinsulins, seven proteinase inhibitors, and 12 species of pancreatic ribonucleases. Despite relatively low sequence conservation in these three sets of proteins, β turns were predicted to be highly conserved: 33% sequence versus 78% bend for proinsulins, 20% sequence versus 85% bend for proteinase inhibitors, and 65% sequence versus 92% bend for ribonucleases.

It was suggested that chain-reversal regions play an essential role in keeping the active structural domains in hormones and enzymes intact for their specific biological function. The concept of β bends [nonhelical dipeptide sequences in which the distance $R_3(i,i+3)$ between C^α of residues i and $i+3$ is ≤7.0 Å] has been extended to define double bends as tripeptide sequences not in an α helix in which two successive distances $R_3(i,i+3)$ and $R_3(i+1,i+4)$ are both ≤7.0 Å, with analogous definitions for higher-order multiple bends, as examined by Isogai *et al.* (1980). A sample of 23 proteins consisting of 4050 residues contains 235 single, 58 double, and 11 higher-order multiple bends. Multiple bends may occur as combinations of the "standard" type I, II, and III chain reversals (as well as their mirror images), but usually they require distortions from these well-defined conformations. The frequencies of occurrence of amino acids often differ significantly between single and multiple bends. The probability of distribution of R_3 distances does not differ in single and multiple bends. However, R_4 (the distance between the C^α atoms of residues i and $i+4$) in multiple bends is generally shorter than in tripeptide sequences containing single bends. The value of R_4 in many multiple bends is near those for α helices. Double bends in which the signs of two successive virtual-bond dihedral angles differ have conformations that are very different from an α helix. They act as

chain reversals occurring over three residues. Multiple bends may play an important role in protein folding because they occur fairly frequently in proteins and cause major changes in the direction of the polypeptide chain. Such multiple β bends have been frequently predicted by the Chou and Fasman (1977b) method (G. D. Fasman, unpublished data).

A simple algorithm was developed by Kolasker *et al.* (1980) to detect β bends and "loop" chain reversals containing five amino acid residues, using only coordinates of C^{α} atoms from crystal structure data of globular proteins. Analysis of bends showed that the total number of bends in each protein (T_B) is linearly related to the total number of nonhydrophobic residues in that protein, which in turn is related linearly to the total number of amino acid residues. This result is similar to that reported by Rose and Wetlaufer (1977). They also reported that a large number of consecutive bends occur in each protein, which give rise to, on an average, only three independent residues per turn. Positional preferences of amino acid residues in chain reversals were stressed.

Nemethy and Scheraga (1980) characterized β bends in proteins by a range of dihedral angles. These were classified into eight groups according to the orientation of the three peptide groups comprising the bend. The possibility of formation of intrabend hydrogen bonds involving N-H and C-O groups depends on the relative orientation of the peptide groups and hence differs for various types of bends. On occasion, β turns were buried in the hydrophobic interior of the molecule, although most turns are always situated at the surface of the protein in contact with solvent water (Rose *et al.*, 1983). In every instance of a buried turn, one or more solvent molecules were also found in a hydrogen-bonded complex with main-chain atoms of the turn residues. These bound water molecules appear to function as an integral part of the protein structure.

Cohen *et al.* (1983) described an algorithm for assigning secondary structure of α/β proteins. Turns were identified very accurately (98%) by simultaneously considering hydrophilicity and the ideal spacing of turns throughout the sequence. The segments bounded by these turns are labeled by a pattern recognition scheme based on the physical properties of α helices and β strands in this class of proteins. Long-range as well as local information is incorporated to enhance the quality of the assignments. The algorithm successfully divides proteins into two classes: α/β and non-α/β. This method is discussed in detail in Section VI.

Tetrapeptide sequences of the type *Z*-Pro-*Y*-*X* were obtained from the crystal structure data on 34 globular proteins and used in an analysis of the positional preferences of the individual amino acids in the β-turn conformation (Ananthanarayanan *et al.*, 1984). This work is an extension of the work of Chou and Fasman (1977b). The effect of fixing proline as the second-position residue in the tetrapeptide sequence was studied. Differences were found in positional preferences for the two sequences *Z*-*R*-*Y*-*X* and *Z*-Pro-*Y*-*Y*. Murakami (1985, 1987) made an attempt to predict the conformations, particularly β turns, around the mutation regions of the p21 protein and its cancer-associated variants by using the prediction method of Chou and Fasman (1978a). Mutations affecting the 12th and 61st amino acid of p21 resulted in a decreased probability of β-turn occurrence. Point mutations at residues 12, 13, or 61 are involved in malignant activation of *ras* protooncogenes. Probabilities of β-turn occurrence at residues 10–13 or 58–61 of the p21 proteins are high. Thus, these critical amino acids lie within β turns.

β Hairpins are widespread in globular proteins and have often been suggested as possible sites for nucleation (Ptitsyn, 1981). Sibanda and Thornton (1985) examined the loop regions of β hairpins in proteins of known structure and found that the "tight" β hairpins, classified by the length and conformation of their loop regions, form distinct families, and the loop regions of the family members have sequences that are characteristic of that family. The two-residue hairpin loops include entirely I′ or II′ β turns, in contrast to the general preferences for type 1

and type 2 turns (Chou and Fasman, 1977b; Lewis *et al.*, 1973a). Milner-White and Poet (1986, 1987) demonstrated that β hairpins can be divided into four classes, each with a number of members. Hairpins from a single class are readily interconverted by loss or gain of hydrogen bonds, but interconversion between classes requires unzippering and reformation of the entire β hairpin.

Milner-White (1988) determined that there is a recurring loop motif in proteins that occurs in both right-handed and left-handed forms, found frequently at the C-terminal end of α helices with a characteristic hydrogen bond pattern, which is called a paperclip. It was also noted that several loops with the same structural features occur independently of α helices; e.g., two are situated at the loop end of β hairpins. These paperclips exist in two classes depending on the number of residues at the loop end. Two such loops belong to the common class except that the main-chain conformation is the mirror image of that normally found. The majority of paperclips were shown to have tightly clustered sets of main-chain dihedral angles.

These are somewhat similar to, but distinct from, a subgroup of another common family of loops that have been called β bulge loops, whose dihedral angles are also tightly clustered. The high degree of clustering in both cases is likely to be a result of steric constraints associated with hydrogen-bond patterns at the ends of loops. Sibanda and Thornton (1985) have classified β hairpins in terms of the number of "loop residues" as either two-residue, three-residue, four-residue, etc., loops. This leads to ambiguity, but if the class (one of the four mentioned above) is also specified, the description of β hairpins becomes straightforward.

Cohen *et al.* (1986a,b) extended the use of amino acid sequence patterns (Cohen *et al.*, 1983) to the identification of turns in globular proteins. The approach uses a conservative strategy combined with a hierarchic search and length-dependent masking to achieve high accuracy (95%) on a test of proteins of known structure. Applying the same procedure to homologous families gives a 90% success rate. The loops in globular proteins were considered in detail on 67 proteins of known structure by Leszczynski and Rose (1986) were categorized as a novel secondary structure. The protein loop, a novel category of nonregular secondary structure, is a segment of contiguous polypeptide chain that traces a "loop-shaped" path in three-dimensional space; the main chain of an idealized loop resembles a Greek omega (Ω). A systematic study revealed 270 omega loops. Although such loops are typically regarded as "random coil," they are, in fact, highly compact substructures and may also be independent folding units. Loops are almost invariably situated at the protein surface, where they can assume important roles in molecular function and biological recognition.

To identify a loop from x-ray coordinates, the following rules are applied. (1) The segment length must be between six and 16 residues. Loops of this length allow their side-chain atoms to pack within the loop's own core. (2) The distance between segment termini, that is, end-to-end distance, is measured as the distance from the first α carbon to the last α carbon in the segment, must be less than 10 Å, and may not exceed two-thirds the maximum distance between any two α carbons within the segment under consideration. The residue composition of loops was assessed by calculating the normalized frequency of occurrence, f, for each residue type, X, such that $f = X_L X_T / N_L N_T$, where X_L is the number of the residues of type X in loops, X_T is the total number of residues of type X, N_L is the total number of residues in loops, and N_T is the total number of residues in the data base. These frequencies, when compared to the p_t values of Chou and Fasman (1977c), reveal that the residues most often found in reverse turns are also found most often in loops (Gly, Pro, Asp, Asn, and Ser). Parrilla *et al.* (1986) described a simple PASCAL microcomputer program based on the Chou and Fasman (1974a,b) algorithm for the prediction of protein secondary structure. The program also performs an analysis of the hydrophobic character of the residues. The authors

compare their results with the original results of Chou and Fasman (1978a) for the prediction of β turns and report very good agreement.

Edwards *et al.* (1987) analyzed 129 loops of 70 βαβ units from 17 α/β proteins for patterns. There were many different conformations of the loop regions, but 18 of the loops could be classified into one of four loop families with distinctive conformations and sequence patterns: (I) adjacent α/β loops with one residue between the α helix and β strand—the residue is a Gly with conformationally restricted ϕ,ψ angles; (II) adjacent αβ loops of three residues with a conformationally restricted Gly as the first of the loops followed by an Ala or His and a third residue with helical ϕ,ψ angles; (III) adjacent βα loops of three or four residues previously reported to bind nucleotides and that have three Gly residues in the loop region; (IV) nonadjacent βα loops of zero residues with a Ser or Thr as the last residue of the β strand. The analysis provides information for model building of loops and prediction of secondary structure from amino acid sequences.

In a series of papers, Efimov (1985, 1986a,b) performed a stereochemical analysis of regions previously termed irregular in known protein structures and found new regions of "unknown standard conformations." Using Ramachandran maps (Ramachandran *et al.*, 1963), he assigned six regions of "condensation," designated by the symbols α, α_L, β, γ, δ, and ϵ. The conformation of an irregular region can be written by enumerating the conformations of its constituent residues, starting from the N terminus. The α, α_L, and β have the previous ψ,ϕ values of Ramachandran *et al.* (1963), whereas γ, δ, and ϵ are newly found regions of high occupancy in protein structures. For example, the two β strands in a β sheet can be linked by the structures $\beta\beta\alpha_L\beta$, $\beta\alpha\gamma\alpha_L\beta$, $\beta\alpha\alpha\alpha_L\beta$, etc., which are referred to as turns. The structures of α-helix–β-strand type examined were those in which the α helix and β strand were packed approximately antiparallel with formation of α–β hairpins. It was shown that for α–β hairpins with connector length no more than five or six peptide units, a limited number of standard conformations was found in proteins (Efimov, 1986a). Each of these standard α–β hairpins must have its own strictly determined alternation of hydrophobic, hydrophilic, and glycine residues. In general, two conditions are required for formation of any structure in globular proteins: (1) the prohibition of dehydration of polar groups and (2) the prohibition of strong steric tension. The simplest transition from an α into a β structure via an α–β hairpin is observed with an $\alpha_m\gamma\alpha_L\beta$–$\beta_n$ conformation where *m* is the number of residues in the α helix and *n* is the number of residues in the β structure. Efimov categorizes all possible connectors (or gaps) up to six residues. The β–α hairpin is also analyzed in the same manner (Efimov, 1986b). The standard β–α hairpins are two-layered structures in which hydrophobic clusters of α helices and β strands are oriented approximately toward one another and are located at different layers. They also differ primarily in the length and the conformation of the connectors and also by the degree of twisting of the β strands. A series of proteins with various numbers of amino acids in the connectors were examined, and a defined set of parameters was obtained for these regions previously termed "irregular" (e.g., $\beta_m\beta\alpha\beta\alpha_n$). Thus, it is not obligatory that each β–α hairpin in a protein be unequivocally governed by primary structure, but a certain pattern of residues can cause formation of β turns or other hairpins.

Wilmot and Thornton (1988) examined 59 nonidentical proteins whose x-ray structure had been determined with a resolution of ≤2 Å, and extracted 735 β turns. Using ϕ,ψ angles, these β turns were classified into seven conventional types (I, I′, II, II′, IV, VIa, VIb) and a new class turn, designated VIII, in which the central residues (i + 1, i + 2) adopt an $\alpha_R\beta$ conformation. Type I′ and II′ turns were found in 83% and 53%, respectively, of β hairpins. These two turn types were shown to be strikingly different in their sequence preference. Type I turns favor Asp, Asn, Ser, or Cys at i; Asp, Ser, Thr, or Pro at i + 1; Asp, Ser, Asn, or Arg at i + 2; and Gly, Trp, or Met at i + 3. Type II turns prefer Pro at i + 1; Gly or Asn at i + 2; and

Gln or Arg at $i + 3$. The positional trends for types I and II were incorporated into a simple empirical predictive algorithm originally developed by Lewis *et al.* (1971). With these new parameters, 72% of β turns were predicted within ±2 residues, compared to 41% using the original Chou and Fasman parameters.

Aubert *et al.* (1976) calculated the structure of the peptide segment around the carbohydrate–peptide linkage in glycopeptides by the Chou and Fasman (1974a,b) algorithm. Sequences around Thr and/or Ser for O-glycosidic and Asn for N-glycosidic linkages were studied. Nine O-glycosidically linked glycans and 28 N-glycosidically linked glycans were predicted. All O-glycosidic links were within β turns. The Asn residue was frequently found (19 out of 28) either in or near a β turn. A conformational study of α_1-acid glycoprotein (Aubert and Loucheaux-Lefebrve, 1976) and a prediction of its conformation showed that four of five glycan chains are linked to Asn residues that are situated either in a reverse β turn or in regions where charged residues are numerous. Beeley (1976) located the four carbohydrate groups of ovomucoid, a family of glycoproteins with antitryptic activity that have been isolated from the egg white of several avian species, on Asn groups. These Asn groups were very close to groups of amino acids that occur with high frequency in β turns (Chou and Fasman, 1974b). The sequence Asn(CHO)*X*-Ser/Thr has previously been proposed as a necessary requirement for glycosylation of Asn (Neuberger and Marshall, 1968). Beeley (1977) applied the Chou and Fasman (1974b) predictive scheme to predict the peptide chain conformation of the amino acid sequences adjacent to carbohydrate attachment sites of glycoproteins containing N-glucosylamine-type protein–carbohydrate linkages. Of 31 glycosylated residues examined, 30 occur in sequences favoring β-turn structures.

Small *et al.* (1977) examined 14 different proteins that were highly phosphorylated and found that 24 out of 30 phosphorylated residues (80%) existed within regions predicted to be β turns. Phosphorylated sites not predicted within turns were found to be adjacent to predicted turns (±2 residues) in four other cases. It was suggested that these β turns play a more active role in biological function in addition to their directional effect on the folding of globular proteins. Previously Mercier and Chobert (1976) had pointed out that in the caseinomacropeptides, phosphorylated hydroxy amino acid residues are located in a tripeptide sequence Ser-*X*-Glu. A more detailed predictive study of the K-caseins by Loucheux-Lefebrve *et al.* (1978) demonstrated that both glycosylation and phosphorylation sites were attached to β turns (Chou and Fasman, 1974a). Loucheux-Lefebrve (1978) also predicted β turns in different regions of peptide and glycopeptide antifreezes. Ricard *et al.* (1983) examined the structural requirement of the Asn-*X*-Thr(Ser) sequence for the N-glycosylation of proteins as a local conformation acting as a signal for the enzymatic process. Energy calculations were performed on the substrate Ac-Asn-Ala-Thr-NH_2, and the lowest-energy conformers have been characterized as β-bend structures. The structural comparison of protein sequences around potential N-glycosylation sites was made by Mononen and Karjalainen (1984). One hundred and five proteins with 139 glycosylated and 57 nonglycosylated Asn-*X*-Ser/Thr sites were analyzed. The primary sequence data indicated a lack of glycosylation in structures containing either Pro or Glu at position *X*. The study of predicted structures (Chou and Fasman, 1974b) showed that in both glycosylated and nonglycosylated acceptor sites, 70% of the Asn occurred in β turns, approximately 20% in β sheets, and 10% in helical conformations. The Asn had no preference for a particular position in the β turn. Statistical tests did not show significant differences in the predicted secondary structure, contradicting claims of the importance of the β turns. Parrilla *et al.* (1986) described a simple PASCAL microcomputer program for predicting secondary structure according to the Chou and Fasman (1978a,b) algorithm. They predicted the β turns in four proteins—myothemerythrin, superoxide dismutase, thioredoxin, and trypsin inhibitor—and compared the results to the x-ray-determined structures. The results were very satisfactory.

B. Evaluation of Predictive Methodologies

The three most frequently used predictive methods of the secondary structure of proteins—those of Chou and Fasman (1974a,b), Garnier *et al.* (1978), and Lim (1974a,b)—have come under considerable scrutiny and evaluation. These three methods are compared in this section.

The first comparison of various predictive schemes for predicting α helices, performed before the three above named methods were published, was that reported by Dickerson *et al.* (1971), who had determined the structure of both horse and bonito ferricytochrome *c* at 2.8-Å resolution. The methods of Prothero (1966), Schiffer and Edmundson (1967), Low *et al.* (1968), Kotelchuck and Scheraga (1969), and Lewis *et al.* (1970) were compared. All the methods except that of Kotelchuck and Scheraga (1969) identified the α-helical 91 to 101 region correctly. Most of the methods predicted helix in regions of residues up to 21 and 80–88. The first region, up to 21, has a helical entity in the crystal structure, but the predicted region from 80 to 88 is completely wrong. The helical segment in the sequence in the region of 60–70 is approximately correct for most schemes. Thus, for cytochrome *c*, the helix-predicting methods appeared to that date (1971) to have considerable validity, their chief defect being that they were not sufficiently discriminating, with a tendency to predict more helix than was actually present. The partition function of Lewis *et al.* (1973b) was particularly poor in this respect. Thus, the methods available in 1971 did not give satisfactory predictions.

The second comparison of various predictive schemes was carried out by Schulz in 1974 (Schulz *et al.*, 1974a) for adenylate kinase, whose 3-Å x-ray crystallographic structure he and co-workers (Schulz *et al.*, 1974a) had determined. He invited the authors of 11 predictive methods to use the known sequence and contribute to this joint effort. Figure 1 shows the results of this comparison of the predicted and the experimentally determined α helices, strands of β sheets, and β bends in adenylate kinase. The highest percentage of α helices correctly predicted was 82% (Lim, 1974a,b), with 22% β sheets and 38% turns; Chou and Fasman (1974a,b) predicted correctly 28% of bends, 20% of β sheets, and 70% of α helix. Matthews (1975) compared the predicted and observed secondary structure of T4 phage lysozyme. Within the amino-terminal half of the molecule, the location of helices, predicted by a number of methods, agrees moderately well with the observed structure; however, within the carboxyl half of the molecule, the overall agreement was poor. For 11 different helix predictions, the coefficients giving the correlation between prediction and observation ranged from 0.14 to 0.42 (Lim, 1974a,b, 60% α helix, 42% β sheet; Chou and Fasman, 1974a,b, 53% α helix, 42% β sheet, 36% β turn). This result is less successful than that reported by Schulz *et al.* (1974a) above. Burgess and Scheraga (1975) emphasized that these predictive algorithms gave little information about the three-dimensional structure of a protein. They stated, "However, starting from the results of a *perfect* prediction algorithm, it appears that conformational energy minimization (with long interactions included) can lead to a structure having the general features of the native protein."

Argos *et al.* (1976) computerized five secondary predictive methods (Nagano, 1973, 1974; Barry and Friedman, as quoted in Schulz *et al.* 1974a; Chou and Fasman, 1974a,b; Kabat and Wu, 1973a,b) based on protein amino acid sequence and combined them to calculate a joint prediction. Forty known structures were predicted. It was found, as previously, that accuracy of prediction of α helices is better than that for β sheets or turns and that the amino-terminal-half prediction is superior to that for the carboxyl-terminal half. However, there was a little improvement in this joint predictive scheme.

Lenstra *et al.* (1977) compared the primary structures of 24 homologous ribonucleases that differ in up to 34% of their amino acids. The results using the Chou and Fasman (1974a,b)

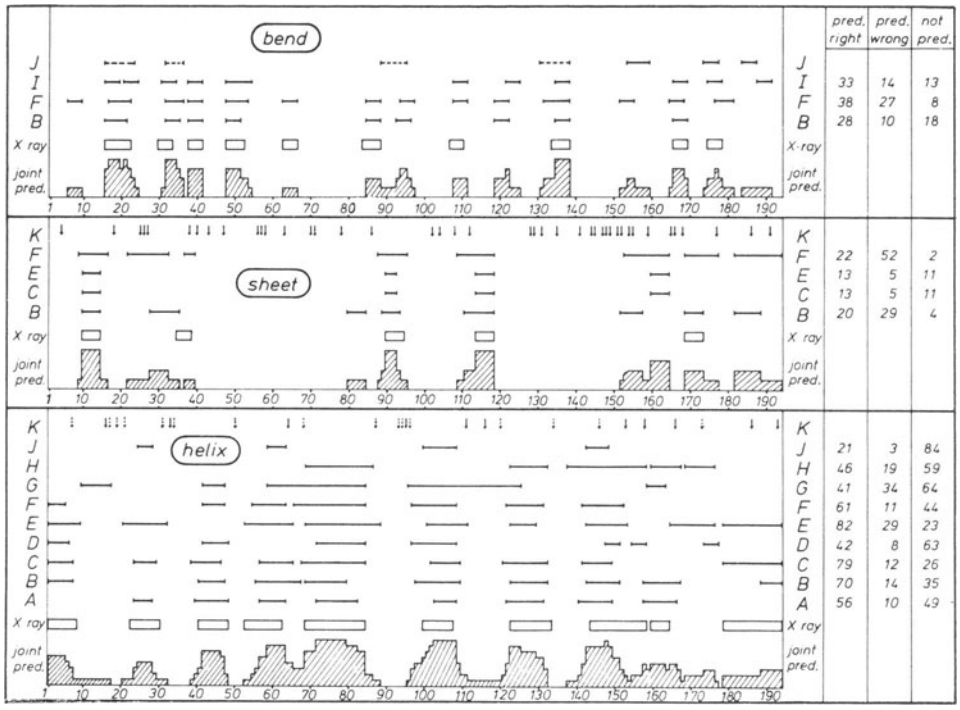

Figure 1. Comparison of predicted and experimentally determined α helices, strands of β pleated sheets, and bends in adenylate kinase. The experimental data (x-ray) have been derived from a 3-Å electron density map. At this resolution the exact geometry of bends cannot be evaluated with certainty. Therefore, the experimentally determined bends are defined as changes of more than 120° in the overall direction of the polypeptide chain and without reference to any hydrogen-bonding scheme. None of the predictions is biased by previous structural information about the enzyme. The predictions *A* to *K* have been supplied by *A*, Barry and Friedman; *B*, Chou and Fasman, 1974a,b; *C*, Ptitsyn and Finkelstein, 1970a,b; *D*, Levitt and Robson; *E*, Lim, 1974a,b; *F*, Nagano, 1973; *G*, Kotelchuck and Scheraga, 1969; *H*, Lewis *et al.*, 1971; *I*, Burgess and Scheraga, 1975; *J*, Burgess *et al.*, 1974; and *K*, Kabat and Wu, 1973a. They are based on prediction schemes that have been described in the corresponding references. (The method applied by Barry and Friedman has not been published yet. It uses an analysis of the distribution of amino acids occurring at both ends of helices to define potential starting and termination points of helices in a sequence. In this method the residues three before and three after, as well as the residue defined as the beginning (or end) of an helix, are considered to be structurally significant. The predictions are based on data for 90 helices from 23 distinct three-dimensional protein structures.) Although predictions are usually made as probability profiles, all groups converted their profiles beforehand to yes-or-no decisions for each residue by comparing the probabilities with a given threshold value. Thus, every prediction depends on the threshold value applied[32]. This procedure simplified the comparison appreciably. But concomitantly, it reduced the amount of information contained in any of the predictions. Kabat and Wu predicted helix and sheet-breaking residues, which are indicated by vertical arrows. Dashed arrows point to residues that are helix breaking but with a lower probability. Dashed lines in the bend prediction *J* indicate "multiple bend regions." A bend is predicted to be anywhere within such a region. Three joint prediction histograms have been produced by adding predictions *A* to *J* for helix, predictions *B*, *C*, *E*, and *F* for sheet, and predictions, *B*, *F*, *I*, and *J* for bends, respectively. Scores of correctly and incorrectly predicted residues as well as residues not predicted are listed on the right side of the figure. From Schulz *et al.* (1974a).

method were in better agreement with the x-ray structure of bovine RNAse than the method of Burgess *et al.* (1974). The method of Lim (1974a,b) gave the most satisfactory results. Lenstra (1977) also evaluated the predictive accuracy of the histogram method of Argos *et al.* (1976), the statistical method of Nagano (1973), and the stereochemical method of Lim (1974a,b). The method of Nagano yielded the best prediction of β structure, whereas the β-structure predictions of Lim and Argos *et al.* were not significantly different. The results of the α-helix and β-structure predictions according to the statistical mechanical method of Tanaka and Scheraga (1976a–d) were inferior to those obtained by the other three methods. For the prediction of turns, there was no significant difference between the methods of Nagano (1974) and Argos *et al.* (1976). The method of Chou and Fasman (1974a,b) was not utilized, as it was stated that their rules were ambiguous and difficult to program.

Kabsch and Sander (1984, 1985) pointed out the lack of structural significance of short sequence homologies. In 62 proteins with 10,000 residues, they found the longest isolated homologies between correlated proteins to be five residues long. In six out of the 25 cases, they found structural adaptability; the same five residues are part of an α helix in one protein and part of a β strand in another. These examples show quantitatively that pentapeptide structure within a protein is strongly dependent on sequence context. However, they state that this fact is essentially ignored in most protein structure prediction methods. This statement is misleading, because it does not apply to several predictive methods (e.g., Chou-Fasman, 1978a,b), where this is taken into account by the methodologies used. Unfortunately, this paper has been quoted to demonstrate the unreliability of predictive schemes (e.g., Harrision, 1985). Wilson *et al.* (1985) have used this observation, that the same pentapeptides can have different conformations in different proteins, to test the theories of immune recognition.

Two sequences that have been found to have different conformations (by x-ray diffraction studies) have been examined by the Chou–Fasman (1978a,b) predictive scheme to investigate this conformational duplicity. The sequence VELIRG, found in influenza neuraminidase (Elleman *et al.*, 1982), was found to have an α-helix structure (Varghese *et al.*, 1983). The same sequence found in the protein disk of tobacco mosaic virus (Anderer, 1963) was shown to have the β-sheet conformation (Bloomer *et al.*, 1978). On prediction of the secondary structure (Chou and Fasman, 1978a,b), $\langle P_\alpha \rangle = 1.07$ and $\langle P_\beta \rangle = 1.20$ were obtained, indicating a β-sheet conformation, which is an incorrect prediction. However, the prediction for the TMV coat protein, $\langle P_\beta \rangle = 1.21 > \langle P_\alpha \rangle = 1.12$, indicated a β-sheet conformation, which agrees with the x-ray structure.

Another sequence, NAAIRS, found in phosphofructokinase (Kolb *et al.*, 1980) was located in an α-helical segment. This same sequence in thermolysine (Titani *et al.*, 1982) was present in a β-sheet conformation (Holmes and Matthews, 1982). On predicting this sequence, with its contiguous residues on either side, the phosphofructokinase gave values of $\langle P_\beta \rangle = 1.14$, $\langle P_\alpha \rangle = 1.06$, an incorrect prediction. However, this sequence in thermolysine, with contiguous sequences, had $\langle P_\alpha \rangle = 1.03$ and $\langle P_\beta \rangle = 1.05$. Thus, the prediction could not distinguish between these structures. The most likely explanation for these questionable predictions is that the secondary structure of a sequence in an intact protein is strongly influenced by its neighboring amino acids as well as by the tertiary interactions encompassed in the domain in which it resides. Attempts are now being made to include the latter factor in secondary structural predictions (e.g., Chou–Fasman).

Schulz and Schirmer (1979) have discussed the merits of various predictive algorithms in great detail and have determined the relationship between the various quality indices used by the individual authors.

Busetta and Hospital (1982) analyzed the prediction of the secondary structures of 38 proteins of known structure, using the methods of Chou and Fasman (1974a,b) and Garnier *et*

al. (1978). The percentage of correctly predicted residues in the three states helix, extended, or turn (H, E, T) of the basic Garnier *et al.* (1978) method was 56.8%, whereas for the Chou and Fasman (1974a,b) model it was found to be 47%. By varying the decision constant DC_H, which corresponds to a proportion of H states, less than 0.2 or greater than 0.5, the percentage of correctly predicted states remains important at, respectively, 49.1% and 56.6%. The use of hydrophobic reinforcements (with $k_E1 = 1.5$) improves the basic procedure of Garnier *et al.* (58.4% correctly predicted states). The prediction level is clearly better for proteins with a single type of secondary structure (all α or all β) than for proteins of a mixed type (α + β or α/β). A previous knowledge of the protein type greatly improves the prediction level. They suggest that the efficiency of prediction could be improved by the use of the distribution of hydrophobic residues.

Kabsch and Sander (1983a) tested the three most widely used methods (Chou and Fasman 1974a,b; Garnier *et al.*, 1978; Lim, 1974a,b,c) for prediction of protein secondary structure from the amino acid sequence on 62 proteins of known structure using a new program package and data collection. They state that they have overcome the ambiguities in two of the best-known methods, the Chou and Fasman (1978a) and Lim (1974a–c) methods, and have also removed the variation in the definition of secondary structure given by crystallographers by their objective and accurate assignment of secondary structure by a pattern recognition algorithm (Kabsch and Sander, 1983b). They have now applied these "improved" methods to more than 10,000 residues. For the three-state definition of secondary structure (helix, sheet, loop/turn), the overall prediction accuracy for these newly defined secondary structures did not exceed 56% for the best of these (Lim, 1974a–c; Garnier *et al.*, 1978) and was only 50% for the most widely used method of Chou and Fasman (1978a). The Chou and Fasman ambiguities were overcome by selecting possible secondary structure segments such that the sum of the preference parameters over all chosen segments was maximal. The β-turn prediction was done separately using the Chou and Fasman (1979a) method. Ambiguities in the method of Lim (1974a) were overcome by a simplified iterative procedure for segment selection written by Lenstra (1977). The method of Robson and co-workers (Garnier *et al.*, 1978) was used as programmed by the authors.

Nishikawa (1983) also assessed the predictive accuracy of the same three methods. The predictive abilities of the three methods turn out to be almost at the same level but unexpectedly low, less than 55% measured by the three-state assessment (α, β, and coil) or less than 45% measured by the four-state assessment (α, β, turn, and coil). Wallace *et al.* (1986) have evaluated the validity of using predictive schemes developed for soluble proteins (Chou and Fasman, 1974a,b; Garnier *et al.*, 1978; Burgess *et al.*, 1974) for membrane proteins and have concluded that they are inappropriate for predicting the structure of membranes (15 examined). Only two of these membrane proteins, crambin and the reaction center from *Rhodopseudomonas viridis*, have had their structure determined by x-ray crystallography. The other membrane protein conformations were determined by physical-chemical techniques, and the interpretation of the data is open to question. Thus, the sample is very limited to draw such a definitive conclusion. Further discussion of this point can be found in Section VII.

Scheraga and co-workers (Burgess *et al.*, 1974; Tanaka and Scheraga, 1976a–c; Maxfield and Scheraga, 1976) have been severely critical of the Chou and Fasman (1974a,b) method. They have pointed out apparent ambiguities in some of the predictive rules, have suggested that the scheme is incomplete, have pointed out perceived misconceptions in their mathematical treatment, and have demonstrated that they cannot reproduce the claimed results.

In conclusion, it can be stated unequivocally that the original claims of accuracy in the predictability of the various methods of the secondary structure of proteins have not been found to be maintained in the laboratories of others.

C. Other Predictive Algorithms

Other predictive algorithms that have been published but not used extensively because of either their lesser predictability score or their newness are those of Scheraga and co-workers (Burgess *et al.*, 1974; Tanaka and Scheraga, 1976a,b,c; Maxfield and Scheraga, 1976) and Troitskii and Zav'yalov (1972). McLachlan (1977) developed a purely statistical theory that uses the observed tendencies of single amino acids and the lengths of typical helices and β sheets. Ptitsyn and Finkelstein (1983) developed a molecular theory of protein secondary structure that takes into account both local interactions inside each chain region and long-range interactions between different regions, incorporating all these interactions in a single Ising-like model. Local interactions are evaluated from the stereochemical theory describing the relative stabilities of α and β structures for different residues in synthetic polypeptides, and long-range effects are approximated by the interactions of each chain region with the averaged hydrophobic template (also see Ptitsyn, 1985). Busetta and Hospital (1982) studied ways of introducing different kinds of interactions that lead to protein folding into secondary structure predictions. The secondary structure prediction was based on the Garnier *et al.* (1978) method with hydrophobic reinforcement (Busetta and Hospital, 1982; Lim, 1974a–c). A previous knowledge of the protein type (e.g., α, α/β) greatly improves the accuracy of prediction.

Moult and James (1987) examined the feasibility of determining the conformation of segments of a polypeptide chain up to six residues in length in globular proteins by means of a systematic search through the possible conformations. Trial conformations are generated by using representative sets of ϕ, ψ, and χ angles that have been derived from an examination of the distribution of these angles in refined protein structures. A set of filters based on simple rules that protein structures obey was used to reduce the number of conformations to a manageable total. The most important filters are the maintenance of chain integrity and the avoidance of too-short van der Waals contacts with the rest of the protein and with other portions of the segment under construction. The electrostatic energy, including a solvent-screening term, and the exposed hydrophobic area are evaluated for each accepted conformation. The method was tested on two segments of chain in the trypsinlike enzyme from *Streptomyces griseus*. It was found that there is a wide spread of energies among the accepted conformations, and the lowest-energy ones have satisfactorily small root mean square deviations from the x-ray structure.

Bashford *et al.* (1987) examined 226 globin amino acid sequences for any unique features. These sequences, which only have two residues absolutely conserved and the residue identities of some pairs of sequences only 16%, were aligned by the use of structural data and analyzed by a new procedure. Although individual chains vary in size between 132 and 157 residues, deletions and insertions result in there being only 102 residue sites in common in all globins, forming six separate regions. Within the conserved regions 32 sites are highly hydrophobic. Another 32 sites are almost always occupied by charged polar or small nonpolar (Gly or Ala) residues, which occur on the protein surface. These six conserved regions and the residue restrictions that occur at the 66 sites within these regions were encoded into two "templates." One was based only on the sequences so far determined; the other was extended to include yet unobserved substitutions that seemed plausible on the basis of size, hydrophobicity, and polarity. Each of the 3286 nonglobin sequences in the data bank was examined by a computer program to see how closely it could be matched to these templates, and it was found that no nonglobin made an exact match to either template. Thus, the features of the globin sequences that are conserved and define its fold are essentially unique to that family.

Klein, DeLisi, and co-workers (Klein *et al.*, 1984; Klein, 1986), using the National Biomedical Research Foundation sequence data, found by discriminant analysis that the pro-

tein superfamilies cluster into six groups that can be distinguished on the basis of four variables characterizing amino acid composition and local sequence data. These variables are average hydrophobicity, net charge, sequence length, and periodic variation in hydrophobic residues along the chain. The clusters they distinguished were (1) globins, (2) chromosomal proteins, (3) contractile system proteins and respiratory proteins other than cytochromes, (4) enzyme inhibitors and toxins, (5) enzymes except hydrolases, and (6) all other proteins. The overall probability of correctly allocating a given protein to one of these functional groups was 0.76, with the allocation reliability being highest for globins (0.97) and for chromosomal proteins (0.93). They also found that approximately 53% of these protein sequences can be allocated to one of 26 functional classes, each of which can be characterized by the joint occurrence of four or fewer attributes. The attributes reflect collective physicochemical properties of the sequences in a class, ranging from simple characteristics of composition, such as average hydrophobicity and net charge, to amphilicity and the propensities of various residues to be in certain preferred conformations. These attributes permit 17 of the 26 groups to be filtered from all other proteins in the data base with a miscalculation error of less than 2%, and the remaining nine groups can be filtered with errors not exceeding 13%.

Argos *et al.* (1982) developed a prediction algorithm for membrane-bound proteins based on physical characteristics of the 20 amino acids and refined by comparison to the proposed bacteriorhodopsin structure. The scheme was devised to delineate likely membrane-buried regions in the primary sequences of protein known to interact with the lipid bilayer. Using the thus-calculated lipid-buried segments in several membrane-bound proteins allowed a hierarchic ranking of the 20 amino acids in their preference to be in lipid contact. A helical wheel analysis of the predicted regions suggested which helical faces are within the protein interior and which were in contact with the lipid bilayer. Such analyses should be viewed with caution, as the original bacteriorhodopsin structure used is a highly speculative one.

An interesting approach was developed by Manavalan, Ponnuswamy, and co-workers (Manavalan and Ponnuswamy, 1977, 1978; Ponnuswamy *et al.*, 1981) based on the "preferred environment" associated with each of the 20 amino acids. Previously one of the best parameters exploited in studies on protein structure has been the hydrophobic index of amino acids (Tanford, 1962; Jones, 1975; Rose, 1978; see Section III.B). However, Chothia (1976) noted that this parameter had a very poor correlation with the extent to which the residue is buried in the native folded-protein matrix. Although the hydrophobic index is a measure of the preference by a residue in a nonpolar environment, it does not reflect to the same extent the environment in protein crystals. This is because of the difference in the environment within the protein molecule and in the nonpolar solvent used in deriving the parameter. Manavalan and Ponnuswamy (1977) defined a parameter called "the surrounding hydrophobicity" for residues, a set of modified hydrophobic indices. This new set of parameters was found to correlate in a much more significant way with the buried/exposed behavior of the residues found in the protein matrix. They extended their studies to define hydrophobic domains, nucleation sites, loop sections, characteristic directionality of the chain (segments), and the depth of the residue from the surface. The phenomenon of protein folding was interpreted in terms of the enrichment of the hydrophobic environment, and an attempt was made to assign spatial positions for the residues from the centroid of the molecule, using statistical parameters.

Argos (1987b) utilized certain residue physical characteristics and the Dayhoff relatedness-odds amino acid exchange matrix (Dayhoff *et al.*, 1983) to provide sensitive criteria for detection of weak sequence homologies. The search procedure uses several residue probe lengths in comparing all possible segments of two protein sequences, and search plots are shown with peak values displayed over the entire search length. Alignments are automatically effected using the highest search matrix values without the necessity of gap penalties.

D. Chou–Fasman Algorithm

The Chou–Fasman algorithm for the prediction of the secondary structure of proteins and its application by the authors (Chou and Fasman, 1974a,b, 1975a,b, 1977a–c, 1978a,b, 1979a,b; Fasman and Chou, 1974; Schulz *et al.*, 1974a,b; Fasman *et al.*, 1976, 1977; Small *et al.*, 1977; Fasman, 1980, 1982, 1985, 1987; Rosenblatt *et al.*, 1980, 1981; Heber-Katz *et al.*, 1985; Nussbaum *et al.*, 1985; Ötvös *et al.*, 1988; Murphy *et al.*, 1988) have been employed in studies of conformational problems in the following areas: a guide to x-ray crystallographic studies; segments with potential for conformational changes; rationale for amino acid substitution in peptide synthesis of biologically active polypeptides; recognition of homologous conformations in analogous proteins from various species with sequence differences; conformational dependence of protein-binding sites to membranes, nucleic acids, etc.; and the understanding of the loss of biological activity through enzyme cleavage (Fasman and Chou, 1974c).

The Chou and Fasman (1974a,b; 1978a,b) method has attracted much attention because it is seemingly simple to use and can be utilized without a computer. A recent survey by ISI, Philadelphia, showed that it had been referenced over 1061 times; however, popularity does not guarantee efficacy. Although its simplicity has been attractive, it has been frequently severely criticized (Matthews, 1975; Kabsch and Sander, 1983a; Burgess and Scheraga, 1975; Maxfield and Scheraga, 1976; Lenstra, 1977; Busetta and Hospital, 1982; Nishikawa, 1983; Wallace *et al.*, 1986). Many of the criticisms are well founded, and many others seek perfection comparable to x-ray crystallography. The original intent (see above) was to aid in our understanding of structure–function relationships and not to substitute for x-ray crystallography. In early comparative evaluations with 11 other methods (Schultz *et al.*, 1974a,b; Schultz and Schirmer, 1979), the Chou–Fasman algorithm compared favorably and ranked near the most accurate in all secondary structure predictions, while the other methods had poorer records. The detailed comparison with other methods is discussed below.

Because the object of this predictive scheme was to open up new approaches to biochemical and molecular biological problems, it can be rated as having contributed significantly in this area (see reviews Fasman, 1985, 1987). The following examples show its wide utility and contributions to interesting biological problems. One of the early predictions was that made for the pancreatic trypsin inhibitor (Chou and Fasman, 1974b). This 58-amino-acid-residue polypeptide was predicted with 87% of the α-helical and 95% of the β residue agreeing with the x-ray-determined structure (Huber *et al.*, 1971). This also illustrated the fact that predictions can be made on shorter sequences than those used to assemble the parameters.

The predictive algorithm not only can locate where the secondary structural regions are but will detect those regions in proteins having both helical and β-forming potentials and therefore the potential for conformational changes. For example, the x-ray crystallographic studies of concanavalin A (Edelman *et al.*, 1972; Hardman and Ainsworth, 1972) showed only 2% helix in the native structure. However, 55% helicity can be induced in concanavalan A with 70% chloroethanol (McCubbin *et al.*, 1971). The predictive scheme (Chou and Fasman, 1974b) correctly located all 12 β-sheet regions in concanavalin A with only one overpredicted β region. In addition, it showed that a total of 47% of its residues in 13 regions also have α potential, although many of these had still higher β potentials. Similarly, elastase has 7% helicity as shown from x-ray diffraction (Shotton and Watson, 1970), but circular dichroism studies showed 35% helicity in sodium dodecyl sulfate (Visser and Blout, 1971). The predictive method (Chou and Fasman, 1974b) showed that there are 79 residues in 15 regions with helical potential, accounting for 33% helicity. Hence, the easily computed $\langle P_\alpha \rangle$ and $\langle P_\beta \rangle$ values for the α and β segments in proteins may assist in elucidating the regions potentially

capable of undergoing conformational change. It is interesting to note that the B1–7 region of insulin was predicted as β sheet (Chou and Fasman, 1974b) with $\langle P_\beta \rangle = 1.15 > \langle P_\alpha \rangle = 1.07$, in agreement with the x-ray data. Since B1–7 also has α potential, it is not surprising that this region was found to be helical in 4-Zn insulin in 6% NaCl (Bentley *et al.*, 1976).

The predictions of the conformation of glucagon, a hormone containing 29 amino acid residues, offers an excellent example of the potential of the method. Utilizing the conformational parameters for helix, β sheet, β turns, and random coil, Chou and Fasman (1975) predicted two conformational states for glucagon. They showed that the conformational sensitivity of glucagon may reside in residues 19–27, which have both α-helical potential ($\langle P_\alpha \rangle = 1.19$) and β-sheet potential ($\langle P_\beta \rangle = 1.25$). Thus, in predicted form *a*, residues 5–10 with $\langle P_\beta \rangle = 1.08 > \langle P_\alpha \rangle = 0.86$ adopt a β conformation, while residues 19–27 form a helical region (32% α, 21% β). In predicted form *b*, both regions, residues 5–10 and residues 19–27, are β sheets (0% α, 52% β). Circular dichroism studies (Srere and Brooks, 1969) of glucagon solutions (12·6 mg/ml) yield 33% α and 20% β, supportive of form *a*. Infrared studies of glucagon gels and fibrils (Gratzer *et al.*, 1967; Epand, 1971) have a predominant β conformation consistent with form *b*. In addition, three reverse β turns were predicted at 2–5, 10–13, and 15–18, suggesting that this small polypeptide has the potential to fold into a relatively compact structure. Thus, it appears that glucagon has different α and β conformations under different concentration conditions. Hence, residues 19–27 may be involved in an α → β transition.

The *in vivo* concentration of glucagon is probably too small to elicit the β conformation, but this conformational state may be induced on binding of glucagon to its receptor site. Because the conformational state of region 19–27 is sensitively balanced between α and β states, it was predicted that replacement of one or more residues of high β potential in this region with strong α formers would lock the conformation in the helical state. It is also feasible to lock the β conformation by suitable substitutions. Thus, the predictive scheme offers a working hypothesis whereby the structure of the biologically active hormone may be arrived at. The α → β transition of glucagon in solution has been followed by means of circular dichroism (Moran *et al.*, 1977), providing evidence for the potential conformational change predicted. Hruby *et al.* (1986) tested this hypothesis and studied the conformational aspects of glucagon agonists and antagonists by using synthetic analogues. Full analogues of glucagon designed to increase α-helical probabilities in the C-terminal region of glucagon lead to highly potent analogues. [Lys^{17}, Lys^{18}, Glu^{21}]Glucagon was 500% and 700% more potent than glucagon in the receptor-binding and adenylate cyclase assays, respectively. [Phe^{13}, Lys^{17}, Lys^{18}, Glu^{21}]Glucagon was also more potent than the native hormone.

Merrifield and co-workers (Murphy *et al.*, 1988) synthesized six new analogues of glucagon containing replacements at positions 19, 22, and 23. They were designed to study the correlation between predicted conformation in the 19–27 segment of the hormone and the conformation calculated from circular dichroism measurements and observed activation of adenylate cyclase in the liver membrane. The analogues were [Val^{19}]glucagon, [Val^{22}]glucagon, [Glu^{23}]glucagon, [Val^{19}, Glu^{23}]glucagon, [Glu^{22}, Glu^{23}]glucagon, and [Ala^{22}, Ala^{23}]-glucagon. The structures predicted for the 19–27 segment ranged from strongly α helical to weakly β sheet. The observed conformations varied as functions of amino acid composition, solvent, concentrations, pH, and temperature but did not correlate with prediction. There was, however, a correlation between predicted structure and activation of adenylate cyclase in rat liver membranes.

The conformational parameters P_α, P_β, and P_t are expedient for detecting regions in proteins with potential for conformational changes caused by mutations or changes in solvent conditions. The *lac* repressor–*lac* operator interaction of *Escherichia coli* provides an excel-

lent example of the specificity of protein binding to DNA (Bourgeois and Pfahl, 1976). The amino acid sequence of the *lac* repressor, a polypeptide subunit containing 347 amino acid residues, has been determined. Its secondary structure was predicted to contain 37% α helix and 35% β sheet (Chou *et al.*, 1975). The extensive β sheets predicted in the 215–324 region may be responsible for tetramer stabilization found in both the *lac* repressor and the core. These β sheets are almost devoid of charge and would have an extremely hydrophobic nucleus. There are 23 predicted β turns in the *lac* repressor, made up of 50% charged and polar residues (serine and threonine), which would be found on the surface, conferring solubility.

Examination of five *lac* repressor mutants yields significant information regarding conformational requirements for repressor function. Mutant AP46 has an $Ala^{53} \rightarrow Val^{53}$ replacement, causing a predicted α → β transition at residues 52–57 and a loss of repressor activity. Several amber mutants at Gln^{26} (Leu, Ser, Tyr) still cause repression, and no conformational change was predicted, in agreement with this observation. In mutant AP309 a $Ser^{16} \rightarrow Pro^{16}$ change is incurred with loss of biological activity, and a predicted β turn at 14–17 is lost by this mutation. Thus, in these examples it is possible to correlate biological activity with definite secondary structures, and a loss of activity results from induction of a conformational change by a mutation.

The β-turn conformation can now be predicted with the same degree of accuracy as the α-helical and β-sheet regions in proteins (Chou and Fasman, 1979a). The three-dimensional structure of proinsulin has been predicted (Snell and Smyth, 1975). The C-peptide sequences of ten mammalian species show a remarkable conservation of predicted conformation, with a β turn at residues 15–18 flanked by two helices. Data from 29 proteins (Chou and Fasman, 1978a) have shown that a high β-turn potential exists in the 12–17 regions for this series and, more importantly, in none outside of it (Chou and Fasman, 1979b). Although no biological role has been assigned to the C-peptide, the present prediction shows that the β-turn conservation in proinsulin is probably necessary for directing the proper folding of C-peptide helices, which possibly masks the receptor binding region of the hormone, thus making the precursor, proinsulin, inactive.

The designs of several biologically active polypeptides, based on predicted secondary structures, have been reported (Chakravarty *et al.*, 1973; Gutte *et al.*, 1979; Fukushima *et al.*, 1979; Moser *et al.*, 1983, 1987).

Gutte *et al.* (1979) published an interesting and bold application of the predictive scheme. A model of a neutral artificial 34-residue polypeptide with potential nucleic acid binding activity was synthesized. The conformation was designed to bind the nucleotide sequence of the anticodon of yeast $tRNA^{Phe}$, $m^2{}'GAA$. The backbone of this peptide chain was to fold around this ligand. The structure chosen for the 34-residue peptide contained a β strand, a reverse turn, an antiparallel β strand, a second reverse turn, and an α helix running parallel to the second β strand. The residues chosen were based on the parameters of the Chou and Fasman method (1978a). Specific amino acid side chains were to form salt bridges with the phosphate moieties, other hydrophobic aromatic rings allowed stacking interactions or intercalation with the bases of the trinucleotide ligand, and specific hydrogen bonds were also projected. The product and its covalent dimer showed strong interaction with cytidine phosphates and single-strand DNA. The dimer had considerable ribonuclease activity.

Kaiser and co-workers (Fukushima *et al.*, 1979) designed a synthetic amphiphilic helical docosapeptide with the surface properties of plasma apolipoprotein A-1 based on the Chou and Fasman parameters (1978a).

Gutte and co-workers (Moser *et al.*, 1983) designed and synthesized a hydrophobic 24-residue polypeptide that could potentially form a four-stranded antiparallel β sheet and bind the insecticide DDT. More recently, this group (Moser *et al.*, 1987) expressed the synthetic

gene for this artificial DDT-binding polypeptide (DBP) in *E. coli*. Recombinant and chemically synthesized DBP showed identical properties.

Recently there has been considerable interest in the manner in which precursor proteins are synthesized as prepro sequences and the manner in which they are transported across the membrane (Davies and Tai, 1980). Potts *et al.* (1982) have carefully studied the physiology, biosynthesis, and mechanism of action of the parathyroid hormone (PTH). Rosenblatt *et al.* (1979) have chemically synthesized the peptide representing the NH_2-terminal extension of prepro-PTH. The 30-amino-acid single-chain peptide contains the pre region (24 residues) and the prohormone specific hexapeptide of PTH. The pre sequence has also been called the signal or leader peptide, and it has been found to contain a common hydrophobic core in many proteins. This core may aid it in penetrating the membrane (Wickner, 1979). By the predictive method of Chou and Fasman (1978a), two secondary structures were found to be highly probable for the precursor sequence of prepro-PTH: (1) a form with high β-sheet content ($\alpha = 20\%$, $\beta = 57\%$), corresponding experimentally to the aqueous conformation, and (2) a form with high α-helix content and no β sheet ($\alpha = 83\%$, $\beta = 0\%$), corresponding to the observed conformation in a nonpolar environment. The circular dichroism spectrum of the prepro-PTH was examined in an aqueous buffer (pH 7.0) and was found to contain 27% α-helix and 43% β, whereas in a nonpolar solvent, hexafluoro-2-propanol, similar to an intramembrane environment, the polypeptide was found to contain 46% α helix and 0% β. This would support the mechanism by which leader sequences may facilitate passage of nascent preproteins across the lipid layer of the rough endoplasmic reticulum (RER) and into the RER cisternal space by assuming an α-helical conformation to aid in passing through the membrane (Wickner, 1979). The prediction also indicates the presence of a β turn at the end of the leader sequence, which may delineate the COOH terminus and the cleavage site of the leader sequence.

A closer examination of the conformational roles of signal peptides was made by Briggs and Gierasch (1984). The λ receptor protein wild-type and mutant signal peptides were synthesized, and conformational analyses were performed. Secretion of the *E. coli* λ receptor protein (LamB protein) appears from genetic evidence to be correlated with the predicted tendency of its signal sequence to adopt a α-helical conformation (Emr and Silhavy, 1983). They have isolated a strain of *E. coli* with 12 base-pair deletions in the region for the signal sequence of the λ receptor protein. This mutant synthesizes the LamB protein but is unable to export it to its normal location in the outer membrane. Two pseudorevertants containing point mutations that lead to amino acid substitutions in the shortened signal sequences are able to export and process the LamB protein. Emr and Silhavy (1983) report the results of the Chou and Fasman calculations on all four of the signal peptide sequences. The wild-type signal sequence is predicted to adopt a α-helical conformation. The deletion mutant (missing four amino acids) cannot nucleate a helix and is random. The two revertants were predicted to be capable of assuming an α helix. Briggs and Gierasch (1984) synthesized the 12- to 16-residue portions of the four signal peptides and examined their conformations by CD in both aqueous solvents and micellar systems (40 mM aqueous SDS). In aqueous media all signal peptides were largely in the random conformation. In 40 mM SDS the wild-type and pseudorevertant peptides contained considerable helical content (≃25%), whereas the deletion mutant signal peptide had little change in structure between SDS and water. The predictive method (Chou and Fasman, 1978a) had suggested that the signal sequences capable of assuming an α helix would be found to be exported, whereas nonhelical signal sequences would not. This hypothesis has been further confirmed by the synthesis of the signal peptides and analysis of their conformation by CD. This elegant work also illustrates the fact that the predictive scheme is applicable to membrane proteins. The internal milieu of a membrane is not significantly

different from that of the hydrophobic core of a globular protein. Thus, the conformational parameters are applicable to both environments.

An interesting application of the predictive method was the successful demonstration of the structural homology between proteins from widely differing sources that have been suspected of having a possible ancestral relationship. Goodman, Rivier, and colleagues (Pallai *et al.*, 1983) examined corticotropin-releasing factor (CRF) from the hypothalamus, sauvagine from the skin of a frog, and urotensin I from a teleost fish. These polypeptides have approximately 50% homology in amino acid sequences. The circular dichroism spectra in trifluoroethanol all have approximately the same structure, $\simeq$70–80% α helix. The predicted conformation is in good agreement with the measured value (e.g., CRF, CD 78% α; predicted, 78%). All three polypeptides possess a long internal helix, spanning about 25 residues, connected to a turn region to a COOH-terminal element that is an α-helix in CRF and urotensin I and a β sheet in sauvagine. Thus, based on secondary structure, the similarities and differences in biological activity could be rationalized.

The secondary and tertiary structures of interferon were predicted from four homologous amino acid sequences (Sternberg and Cohen, 1982). Several predictive methods were used, and although satisfactory agreement was lacking, four α helices were found to be important in the tertiary fold. A possible tertiary model for interferon was proposed in which the four α helices pack into a right-handed bundle similar to that observed in several other protein structures. This tertiary structure was obtained by using a helix-docking algorithm (Cohen *et al.*, 1979; Cohen and Sternberg, 1980a,b).

The three-dimensional structure of interleukin-2 has recently been solved to 3.0-Å resolution (Brandhuber *et al.*, 1987). Helices B, C, D, and F form an apparent antiparallel α-helical bundle that differs significantly from the classical four-helix bundle represented by cytochrome c', cytochrome b_{562}, and myohemerythrin (Richardson, 1981). The packing regions of the helices are shorter, involving only three to four turns of helix, whereas classical four-helix bundles usually have at least five turns in each helix. There are also three other helical regional (B′, A, and E). Therefore, the predictive result is encouraging but requires further modification and refinement.

Kaiser and co-workers (De Grado *et al.*, 1981) designed and synthesized a cytotoxic peptide with mellitinlike activity. The desired polypeptide of 26 residues was to possess an amphiphilic α helix, bind to phospholipid layers, and form monolayers as well as have a basic C-terminal hexapeptide. A synthetic peptide with sequence residues based on P_α and P_β values that was homologous to mellitin was produced, and its CD spectrum in aqueous media had a 69% α-helical content and formed a tetramer in solution as mellitin did. This synthetic polypeptide formed stable monolayers, caused hemolysis of erythrocytes, and disrupted phospholipid bilayers. Thus, this amphiphilic synthetic peptide appears to possess some of the properties necessary for cytotoxic behavior.

An approach to the design of peptide-hormone analogues in which amino acid substitutions are based on predicted effects on secondary structure was investigated. The structural requirements for analogues of the parathyroid hormone's binding domain in the region 25–34 were investigated for bioactivity, and their solution conformation was determined by circular dichroism (Nussbaum *et al.*, 1985). Biological activity of these analogues in the rat renal adenylate cyclase assay *in vitro* and binding affinity in a radioreceptor assay were threefold those of the unsubstituted PTH_{1-34}.

Hruby *et al.* (1986) and Merrifield and co-workers (Murphy *et al.*, 1988) studied the conformational dependence of the bioactivity of glucagon and found striking effects. Hruby *et al.* (1986) prepared a number of analogues of glucagon by total synthesis using the solid-phase

method of peptide synthesis. Analogues of the 29-amino-acid hormone designed to increase α-helical probabilities in the C-terminal region of glucagon led to highly potent analogues. [Lys^{17}, Lys^{18}, Glu^{21}]Glucagon was 500% and 700% more potent than native glucagon in the receptor-binding and adenylate cyclase assays, respectively. Glucagon has a $\langle P_\alpha \rangle_{19-27} = 1.18$, whereas [Lys^{17}, Lys^{18}, Glu^{21}]glucagon has a $\langle P_\alpha \rangle_{19-27} = 1.23$, thus having higher helical potential. [Phe^{13}, Lys^{18}, Glu^{21}]Glucagon was also more potent than the native hormone. The importance of the 10–13 residues of glucagon as a "hinge region" for correlating binding and transduction regions of glucagon was investigated. It was found that [Phe^{13}]glucagonamide is a potent analogue with full biological activity; [Phe^{10}]glucagon, although quite potent, is a partial agonist. Other derivatives were investigated for their agonist and antagonist potentials. Merrifield and co-workers (Murphy *et al.*, 1988) synthesized analogues of glucagon with amino acid replacements at positions 19, 22, and 23. They were designed to study the correlation among predicted conformation in the 19–27 segment of the hormone, the conformation calculated from circular dichroism measurements, and the observed activation of adenylate cyclase in the rat liver membrane. The observed conformations did not correlate well with prediction, but the predicted conformation and activation of adenylate cyclase in the rat liver membrane did correlate well.

Because a computer program was not published by the authors for the Chou–Fasman predictive method for the secondary structure of proteins (Chou and Fasman, 1974a,b), there have been many published procedures (approximately 20) as well as methods that have alterations and additions that are claimed to have improved the method. The first computerized method to appear was that of Argos *et al.* (1976). The known structures of about 40 proteins were compared to other predictive methods (see below). The accuracy of the predicted helices was found to be better than for β-sheet regions and turns. The amino-terminal half of the protein molecule was predicted with higher accuracy than was the carboxyl half. Dufton and Hider (1977) applied a modified Chou–Fasman algorithm to the prediction of the secondary structure of 57 snake venom toxins. A common distribution of secondary structure was detected throughout these toxins with a distinctive pattern of β sheet, α helix, and β bend for each toxin. The basic difference utilized in their scheme, relative to the original algorithm, was to employ straightforward multiplication instead of the arithmetic mean to find favorable segments.

Argos *et al.* (1978) applied their computer program (Argos *et al.*, 1977) with an expanded data base but found no improvement in the correctness of the predictions. The data base used was that obtained by Levitt (1978), who expanded the data base from 1939 residues (Chou and Fasman, 1974a) to a 5523-residue sample. Busetta and Hospital (1982) attempted to improve the efficiency of the Chou and Fasman (1974a,b) and the Garnier *et al.* (1978) procedures for the prediction of secondary structures by the introduction of the use of the distribution of hydrophobic residues. A set of 38 proteins was examined, and the ratio of residues in the helix versus helix plus extended was utilized, as previously demonstrated by Garnier *et al.* (1978), which produced slightly improved predictions. These authors noted that the prediction level is better for proteins with a single type of secondary structure (all α or all β) than for a mixed type (α + β or α/β). A BASIC microcomputer program for plotting the secondary structure of proteins was developed by Corrigan and Huang (1982) for use on an Apple II+. A modified version of the Chou and Fasman method (1974a,b) was constructed that included nucleation site determination through multiplication of conformational preferences as well as weighing factors to represent structurally stabilizing short-range interactions. They concluded that this method has nearly achieved its upper limit of prediction accuracy, although slight improvement through the use of stereochemical weighing factors and conformational parameters might

be possible. On analyzing the conformational parameters of Chou and Fasman (1978a), Charton and Charton (1983) found that for the α helix, coil, and turn parameters, steric effects were predominant, whereas for β-sheet parameters, intramolecular forces are dominant.

The need for a computer program to apply the Chou and Fasman (1978a, 1979a) prediction method was pointed out by Rawlings *et al.* (1983), and a DEC System 10 FORTRAN computer program was published and made available. Comparisons of their results for seven proteins were also made available with the predictive schemes of Nagano (1977) and Burgess *et al.* (1974). The results achieved for the seven proteins studied were excellent (between 80 and 96% correct). Novotny and Auffray (1984) described a computer program (written in FORTRAN, which runs on a Digital VAX 11/780 computer) that, given a nucleotide or amino acid sequence, outputs protein secondary structure prediction curves as well as hydrophobicity and charge-residue profiles (graphics output on several terminals, VT125, Tektronix 4010, GIGI). The program allows for cumulative averaging of properties (secondary structure propensities, hydrophobicity, and charge profiles) of several homologous primary structures, a concept shown to improve the predictive accuracy. The smoothing of hydrophobicity profiles used in this procedure was that of Rose and Roy (1980). This method was applied to human and murine histocompatability antigens of class I and II. A PASCAL microcomputer program for prediction of protein secondary structure, written for use on an Apple IIe or IIc, was published by Parrilla *et al.* (1986). This uses the Chou and Fasman (1974a,b) method with minor modifications and, in addition, performs an analysis of the hydropathic character of the residues for predicting of external/internal regions of the polypeptide chain. It also can search for probable glycosylation and phosphorylation sites. The hydropathic character is similar to that chosen by Kyte and Doolittle (1982). The sequence chosen for N-glycosylation is Asn-*X*-Thr/Ser (Mononen and Karjalainen, 1984), and the target for phosphorylation is Gly/Ile-Ser/Thr-Gly/Ala/Val-Lys/Arg (Schulz and Schirmer, 1979).

MESQ, a versatile and "user-friendly" software package dedicated to the analysis, display, and prediction of protein structure was published by Black and Glorioso (1986). Protein secondary structure is predicted using the parameters of Chou and Fasman (1978a,b). Zero-order sequence hydrophobicity is calculated with the use of four user-selectable sets of 20 amino acid side-chain polarity values obtained from Argos *et al.* (1982), von Heijne (1981a), Hopp and Woods (1981), and Kyte and Doolittle (1982). Both the α and β-strand hydrophobic moments are also calculated according to the method of Eisenberg *et al.* (1982a,b, 1984a,b). The program is written for the IBM microcomputer family with the extensive use of color graphics, graphic output, and structural "cartoons."

Gribskov *et al.* (1986) described a computer program for the analysis of protein secondary structure that produces both graphic and printed output. Structural predictions are based on the Chou and Fasman (1978a,b) and Garnier *et al.* (1978) methods, hydropathy analysis by the method of Kyte and Doolittle (1982), and a simplified method of hydrophobic moment analysis (Eisenberg *et al.*, 1984a,b) is included.

An interactive analysis of protein structure using a microcomputer spreadsheet, Lotus 1-2-3, was used to predict the α-helix, β-sheet, and hydrophobicity profiles of protein sequences (Vickery, 1987). The Chou and Fasman (1978a,b) empirical algorithm was used to predict secondary structure, and the Kyte and Doolittle (1982) method for the analysis of hydrophilicity/hydrophobicity was applied, using an IBM personal computer. Williams *et al.* (1987) incorporated information about neighboring residues participating in short- and medium-range interactions into predictions of protein secondary structure and, contrary to proposals by several authors, found that no improvement resulted. However, they obtained an 8% improvement for predictions of secondary structure based on the algorithm of Chou and Fasman (1974a,b) by eliminating many rules that include neighboring residue interactions and

choosing the best decision constants for structure assignments. Specifically, all rules described by Chou and Fasman (1978a,b) regarding α-helix and β-sheet nucleation, propagation, termination, breakers, boundaries, and overlapping regions were not used. This simplified method is claimed to yield a 57% correct assignment for three states α helix, β strand, and β turn.

Krchnak *et al.* (1987) describe a computer program (written in BASIC for an IBM personal computer) that predicts protein immunogenic determinants on the basis of the probability of β turns as specified in the Chou and Fasman (1978a,b) procedure.

Deléage *et al.* (1987) developed a computerized program (written for an Apple IIe microcomputer) for predicting the secondary structure of proteins from their amino acid sequence by closely following the scheme of Chou and Fasman (1978a,b). Some of the qualitative rules (Chou and Fasman, 1978a,b) have been converted to numerical scales to obtain unambiguous predictions. On testing 21 proteins with known three-dimensional structure, the percentage of correctly predicted amino acids was between 41 and 66% for a three-state (α helix, β sheet, and coil) description of protein secondary structure. Spectroscopic and structural properties of three homologous dimeric inhibitors of microbial origin, *Streptomyces* subtilisin inhibitor, alkaline proteinase inhibitor, and plasminostreptin, were examined by comparing hydropathy maps (method of Kyte and Doolittle, 1982).

Deléage and Roux (1987) described an algorithm for the prediction of protein secondary structure with a marked improvement of accuracy by taking into account the predicted class of the proteins. This "double predictive method" consists of first predicting the secondary structure from a new algorithm that uses parameters of the type described by Chou and Fasman (1978a,b) and then predicting the class of proteins from their amino acid composition (Nakashima *et al.*, 1986). The parameters were obtained from a 59-protein data base (Kabsch and Sander, 1983b). These two independent predictions allow one to optimize the parameters calculated over the secondary structure data base to provide a final improved prediction of secondary structure. When the procedure was tested on the 59 proteins in the data base, a 72% success in class prediction was obtained, and 61.3% of residues were correctly predicted for three states (α helix, β strand, and coil). This method shows great promise.

Ralph *et al.* (1987) have written a FORTRAN program, called PRSTRC for the prediction of the secondary structure of proteins. This modified Chou and Fasman (1978a,b) analysis carries out a running average of amino acid structure occurrence frequencies, utilizes a simple set of nucleation conditions, and in addition, allows the user control over nucleation threshold and cutoff parameters. The algorithm has also included the prediction of omega loops (Leszczynski and Rose, 1986) and includes a profile of charge distribution and a hydropathy profile (Rose *et al.*, 1985). The authors claim an overall improvement in the average correlation coefficient for α helices and β strands of 12% and 24%, respectively, over the Chou–Fasman algorithm as written by IntelliGenetics (1981). PRSTRC allows the user control of all threshold and cutoff values in the prediction scheme. Thus, if the x-ray structure is known, one may manipulate the values to optimize the accuracy. They state that this flexibility allows for optimization of secondary structure prediction for homologous proteins. Therefore, their figures for increased accuracy must be taken with caution, as it has not been demonstrated on a blind test with many proteins.

Lathrop *et al.* (1987) developed a computer program, ARIADNE, as a hierarchic pattern-directed inference system for protein structural analysis. Input to ARIADNE consists of the primary sequence, any secondary structure predictions, and patterns describing the structure of interest. The secondary structure, α helices, β strands, and β turns, were predicted by the program PRSTC (Ralph *et al.*, 1987) on the basis of the Chou and Fasman (1974b) conformational probabilities. The system identifies the optimal match between a given complex pattern descriptor and protein sequences annotated with various inferred properties by abstracting

Table I. Conformational Parameters for α-Helical and β-Sheet Residues

α-Residues	P_α		β-Residues	P_β	
		64 Proteins			
Glu	1.44 ± 0.06	H_α	Val	1.64 ± 0.07	
Ala	1.39 ± 0.05		Ile	1.57 ± 0.08	
Met	1.32 ± 0.11		Thr	1.33 ± 0.07	
Leu	1.30 ± 0.05		Tyr	1.31 ± 0.09	
Lys	1.21 ± 0.05	h_α	Trp	1.24 ± 0.14	
His	1.12 ± 0.08		Phe	1.23 ± 0.09	
Gln	1.12 ± 0.07		Leu	1.17 ± 0.06	
Phe	1.11 ± 0.07		Cys	1.07 ± 0.12	
Asp	1.06 ± 0.06		Met	1.01 ± 0.13	
Trp	1.03 ± 0.10	I_α	Gln	1.00 ± 0.09	
Arg	1.00 ± 0.07		Ser	0.94 ± 0.06	
Ile	0.99 ± 0.06	i_α	Arg	0.94 ± 0.09	
Val	0.97 ± 0.05		Gly	0.87 ± 0.05	
Cys	0.95 ± 0.09		His	0.83 ± 0.09	
Thr	0.78 ± 0.05		Ala	0.79 ± 0.05	
Asn	0.78 ± 0.06		Lys	0.73 ± 0.06	
Tyr	0.73 ± 0.06	b_α	Asp	0.66 ± 0.06	
Ser	0.72 ± 0.04		Asn	0.66 ± 0.06	
Gly	0.63 ± 0.04	B_α	Pro	0.62 ± 0.07	
Pro	0.55 ± 0.05		Glu	0.51 ± 0.06	
		24 Proteins			
Glu	1.51		Val	1.70	
Met	1.45		Ile	1.60	H_β
Ala	1.42		Tyr	1.47	
Leu	1.21	H_α	Phe	1.38	
Lys	1.16		Trp	1.37	
Phe	1.13		Leu	1.30	
Gln	1.11		Cys	1.19	h_β
Trp	1.08		Thr	1.19	
Ile	1.08	h_α	Gln	1.10	
Val	1.06		Met	1.05	
Asp	1.01	I_α	Arg	0.93	
His	1.00		Asn	0.89	i_β
Arg	0.98		His	0.87	
Thr	0.83	i_α	Ala	0.83	
Ser	0.77		Ser	0.75	
Cys	0.70		Gly	0.75	b_β
Tyr	0.69	b_α	Lys	0.74	
Asn	0.67		Pro	0.55	
Pro	0.57	B_α	Asp	0.54	B_β
Gly	0.57		Glu	0.37	

intermediate levels of structural organization. This method was applied to predict a common structural domain in amino acyl-tRNA synthetases (Webster *et al.*, 1987). The general utility and reliability of ARIADNE will have to await further tests on known structures.

E. Class Prediction

Chou (1979, 1980) analyzed the x-ray structures of 64 different proteins containing 11,444 residues in terms of their α-helical and β-sheet regions as well as their amino acid compositions. This sample represented a larger number of nonhomologous proteins than the Levitt (1978) sample (49 different proteins out of 66). It was found that the four distinct classes of proteins—α, β, α + β, and α/β (Levitt and Chothia, 1976; Schulz and Schirmer, 1979; Richardson, 1981)—had significantly different amino acid compositions. The sample had 19 α proteins (69% α, 1% β), 15 β proteins (15% β, 5% α), 16 α/β proteins (35% α, 23% β), and 14 α + β proteins (34% α, 21% β). The average residue chain length for the four classes of proteins varied in decreasing order: $\langle N_{\alpha/\beta} \rangle = 271 > \langle N_{\beta} \rangle = 185 > \langle N_{\alpha+\beta} \rangle = 135 > \langle N_{\alpha} \rangle > 129$. The overall secondary structural content in the 64 proteins was 35% α helical and 25% β sheet, which is very similar to the α and β content in α + β as well as α/β proteins. An earlier analysis based on 29 proteins (4741 residues) showed an average 38% helicity and 20% β content (Chou and Fasman, 1978b), whereas the more recent analysis (Levitt, 1978) found 31% α and 28% β in 66 proteins. It should be noted that of the 66 proteins analyzed by Levitt (1978), 17 were redundant in the sense that some proteins were independently solved halves of dimers and some proteins were the same but solved by different laboratories. Compared to the conformational parameters of Chou and Fasman (1978b), there were a few significant changes (Table I), and the relative order of the amino acids was relatively constant. A computerized algorithm was developed that could assign proteins to the correct structural class based on their amino acid compositions with 80% accuracy. Chou (1979) developed a new set of conformational parameters, P_{α} and P_{β}, for the amino acids in the four different classes of proteins (see Chapter 12 by Chou). With this new methodology, some previously incorrect predictions (Chou and Fasman, 1978b) were corrected (for greater details of the method, see Chapter 12).

Geisow and Roberts (1980) also reported that the amino acid preferences for secondary structure vary with the protein class. The classes were also determined by the algorithm described by Levitt and Greer (1977). The P parameters for the classes α, β, mixed α/β, and the global values were calculated. The global values found by Geisow and Roberts (1980) were similar but not identical to those of Chou and Fasman (1974a,b). Structurally important residues (hydrophobic, disulfide-bond-forming, and charged amino acids) have the largest index changes between classes. Previously it had been reported (Lifson and Sander, 1979) that the preference of amino acids for parallel or antiparallel β-strand arrangements differed substantially. The Chou and Fasman (1974a,b) procedure was slightly modified by Geisow and Roberts (1980) to incorporate the information provided by a periodic structure, and predictions were made with the new indices; significant improvements in the α proteins were obtained, but not with the β-proteins. Busett and Hospital (1982) incorporated the possible use of the distribution of the hydrophobic residues in both the Chou and Fasman (1978a,b) and Garnier *et al.* (1978) methods. The prediction level was improved for proteins of a single type of secondary structure (all α or all β) more than for the mixed type (α + β or α/β).

Klein and DeLisi (1986) used the multidimensional statistical technique of discriminant analysis to allocate amino acid sequences to one of four secondary structural classes: high α content, high β content, mixed α and β, and low content of ordered structure. The discrimination was based on four attributes: estimates of percentages of α and β structures (Chou and Fasman, 1974b, as modified by Dufton and Hider, 1977; and the method of Garnier *et al.*, 1978) and regular variations in the hydrophobic values of residues along the sequence occur-

ring with periods of 2 and 3.6 residues. The accuracy of the method, classifying 138 sequences, is 80% with no misallocations between α-rich and β-rich classes. Nakashima *et al.* (1986) analyzed the folding type of 135 proteins of known three-dimensional structure in terms of their amino acid composition. The amino acid composition of a protein was expressed as a point in a multidimensional space spanned with 20 axes, on which the corresponding contents of 20 amino acids in the protein are represented. The distribution pattern of proteins in this composition space was examined in relation to five folding types: α, β, α/β, α + β, and irregular type. The α, β, and α/β types are distinctly separated; however, the points representing proteins of the α + β and irregular types cannot be easily classified. The assignment of the folding type to five classes by this method gave an accuracy of 70%, and to four folding classes a reliability of 79%. The authors found ambiguity in assigning a unique type of some proteins by the method of Levitt and Chothia (1976). Therefore, a quantitative measure to define a folding type was adopted: α-type proteins, $\alpha > 15\%$ and $\beta < 10\%$; β-type proteins, $\alpha < 15\%$ and $\beta > 10\%$; α/β-type proteins, $\alpha > 15\%$ and $\beta > 10\%$ with dominantly parallel β sheets; α + β-type proteins, $\alpha > 15\%$ and $\beta > 10\%$ with predominantly antiparallel β sheets; and irregular-type proteins, $\alpha < 15\%$ and $\beta < 10\%$. The success of this procedure has an important implication, namely, that properties of single amino acids are, as an approximation, additive regardless of the sequence. The nature of a protein as a whole is, therefore, approximately determined by its amino acid composition.

Klein (1986) improved on the Nakashima *et al.* (1986) analysis by the use of discriminant analysis. Analysis by class—α, β, mixed (α/β or α + β), and irregular (four classes)—can be predicted with an 83% reliability, and to one of five classes (separate α/β and α + β) with 78% reliability. It was shown that optimal linear combinations of amino acid frequencies used as attributes give slightly better accuracy than attributes using secondary prediction methods for each residue.

Deléage and Roux (1987) have used the technique of Nakashima *et al.* (1986) for class prediction and utilized this in a modified Chou and Fasman (1974a, 1978a) algorithm, called a "double prediction method," for the prediction of secondary structure of proteins from their amino acid sequence (see Chapter 12). When tested on 59 proteins, the methods yielded a 72% success in class prediction, with 61.3% of residues correctly predicted for three states (α helix, β sheet, and coil). Ponder and Richards (1987) approached the folding problem, i.e., structural class prediction, from another point of view, that previously expressed by Drexler (1980): "What sequences are compatible with a given structure?" Ponder and Richards (1987) have elaborated on the Drexler suggestion through the development of tertiary templates for the various classes of protein domains. Blundell and Sternberg (1985) had previously suggested the phrase "tertiary template" to describe a peptide that is characteristic of a particular tertiary structure or fold and in which selected sequence positions have a specified composition.

Ponder and Richards' (1987) use of the term tertiary template is an extension of the above definition to the entire core of a protein domain and to the development of the templates through stereochemical considerations rather than statistical inference. They assume that each class of protein has a core structure that is defined by internal residues and that external, solvent-contacting residues contributing to the stability of the structure are of primary importance to function but do not determine the architecture of the core portions of the polypeptide chain. An algorithm has been developed to supply a list of permitted sequences of internal residues compatible with a known core structure. This list is referred to as the tertiary template for that structure. In general, the positions in the template are not sequentially adjacent and are distributed through the polypeptide chain. The template is derived using the fixed positions of the main-chain and β-carbon atoms in the test structure and selected stereochemical rules. Two packing criteria were used: avoidance of steric overlap and complete filling of available space. The program also notes potential polar group interactions and disulfide bonds as well as

possible burial of formal charges. Central to the algorithm is the 64-side-chain rotamer library. Preliminary tests make it appear likely that templates prepared from the currently known core structures will be able to discriminate between these structures and be useful in deciding whether a sequence of unknown tertiary structure fits any of the known core classes and, if a fit is found, how the sequence should be aligned in three dimensions to fit the core of that class.

Sheridan *et al.* (1985) found that the amino acid composition and hydrophobicity patterns of protein domains correlate with structures. They correlated the sequence and tertiary structure for 212 domains from globular proteins and polypeptides. The sequence of each domain is described as a set of 25 features: the mole percentage of 20 amino acids, the number of residues in a domain, and the abundance of four simple patterns in the hydrophobicity profile of the sequence. Pattern recognition methods were applied to find the two axes through the 35-dimensional sequence–feature space that best discriminate, respectively, predominantly α-helix domains from predominantly β-strand domains and parallel α/β domains from other domains. The domains were further divided into two categories based on whether the Cys content is above (Cys-rich) or below (normal) 4.5%. They found the secondary structure vector for the subset of Cys-rich domains points in a significantly different direction than the equivalent vector for the normal domains. Thus, Cys-rich and normal domains are best treated separately. On projection of the secondary structure vectors onto the plane containing the origin of the feature space, it is seen that α, β and parallel domains cluster in a plane, with the β cluster partially overlapping the parallel cluster. They could correctly predict the structural class with 83% accuracy.

McGregor *et al.* (1987) studied the relationship between side-chain conformation (dihedral angles) and secondary structure in globular proteins. Sixty-one proteins solved to a resolution of 2 Å or better were analyzed. The strongest feature observed was that the χ_1 distribution (rotation around the C_α–C_β bond) for most side chains in an α helix showed the absence of the g^- ($\chi_1 = 60°$) conformation and a shift towards the t (180°) conformation when compared to the non-α/β structures. The exceptions to this tendency were for short polar side chains that form hydrogen bonds with the main chain, which prefers g^+ (300°). Shifts in the χ_1 preferences for residues in the β sheet were observed.

VI. PREDICTION OF TERTIARY STRUCTURE

The future prospects for the prediction of the tertiary structure of proteins are unlimited. The opportunity for the design and construction of new proteins, e.g., vaccines (Mutter, 1985), drugs, herbicides, and pesticides (Blundell *et al.*, 1986b, 1987), await a better understanding of the protein-folding problem (Ghelis and Yon, 1982; Jaenicke, 1984). The advent of recombinant DNA techniques has led to an explosion of information and techniques for the construction of these new hybrid molecules. Protein engineering will be added to the armament of the protein chemist, making possible studies, such as of enzyme mechanisms, at a level previously thought impossible.

There are three broad approaches presently utilized for the prediction of the tertiary structure of a polypeptide (Ponder and Richards, 1987): (1) use of sequence homology with peptides of known three-dimensional structure; (2) prediction of secondary structure units followed by the assembly of these units into a compact structure; and (3) use of empirical energy functions *ab initio* to derive the tertiary structure of minimum potential energy. The first two rely heavily on the data base of structures determined by x-ray crystallography, but the third, in principle, does not. The first procedure uses rules of structure only by implication. The second uses them quantitatively, with the goal of a rough outline of the structure. The third relies on quantitative calculations for evaluation of trial structures.

In 1975, the characterization of the tertiary structure in globular proteins was still elementary (Kuntz, 1975). The approach that showed promise had been the identification of "folding domains" with larger proteins (Wetlaufer and Ristow, 1973; Liljas and Rossman, 1974; Rossman and Liljas, 1974). The "distance plots" method for the location of domains showed promise (Phillips, 1970; Rossman and Liljas, 1974; Liljas and Rossman, 1974; Kuntz, 1975). These "distance plots" are graphs of C_α–C_α distances plotted against residue number, with contour lines drawn at fixed interatomic distances. Kuntz (1975) pointed out that these maps lead to an ordering of tertiary structural features such as distorted three-dimensional "superhelical" structures that are principal constituents of folding domains.

By 1980, with the structures of ~80 globular proteins known to atomic resolution, it became evident that the tertiary fold is largely determined by the packing of α helices and/or β strands (Levitt and Chothia, 1976; Chothia and Janin, 1982; Richardson, 1976, 1977; Sternberg and Thornton, 1976, 1977a–c; Chothia *et al.*, 1977; Richmond and Richards, 1978). One of three motifs was generally found: a stacked pair of β sheets (β/β); α helices packed against a predominantly parallel β sheet (α/β); or an assembly of α helices (α/α). These motifs satisfy the hydrogen-bonding requirements of buried main-chain nitrogen and oxygen atoms while shielding a substantial fraction of the nonpolar atoms from solvent (Chothia and Janin, 1975). Taylor and Thornton (1984) have devised a procedure that further assists in recognizing these supersecondary structures in proteins.

The supersecondary structures are the main components of the domains that, when assembled, constitute the three-dimensional conformations of proteins. An interesting paper by Dill (1985) discusses the theory of the folding and stability of globular proteins. A theory was developed for the folding of proteins to the globular and soluble state using lattice statistical mechanics. Folding is assumed to be driven by the association of solvophobic monomers to avoid solvent and is opposed by the chain configurational entropy. The theory predicts a phase transition as a function of temperature and solvent character. Molecules that are too short or too long or that have too few solvophobic residues are predicted not to fold. Globular molecules should have a largely solvophobic core, but there is an entropic tendency for some residues to be "out of place," particularly in small molecules. For long chains, molecules comprised of globular domains are predicted to be thermodynamically more stable than spherical molecules. The number of accessible conformations in the globular state is calculated to be an exceedingly small fraction of the number available to the random coil. Thus, as the molecular weights of proteins increase, there is a high probability that they will become divided into domains rather than increasing in size of a single unit.

A. Combinatorial Approach

Cohen, Sternberg, and co-workers (Cohen *et al.*, 1979, 1980, 1981, 1983; Cohen and Sternberg, 1980a; Edwards *et al.*, 1987) have developed a stepwise method, termed the combinatorial approach, for predicting the three-dimensional structure of a protein from its amino acid sequence (for reviews see Nemethy and Scheraga, 1979; Sternberg and Thornton, 1978; Schulz and Schirmer, 1979). There are three stages in this procedure: (1) predict the regular secondary structures, now possible with up to 80% accuracy; (2) pack the α helices and β strands into an approximate native fold; (3) use simplified energy calculations (Levitt, 1976; Kuntz *et al.*, 1976; Robson and Osguthorpe, 1979) to refine the fold into the native structure. Because it has been shown (Hagler and Honig, 1978; Cohen and Sternberg, 1980b) that structures predicted solely by simplified energy calculations are not significantly better than random models for a compact globular protein, the combinatorial approach appears to have many advantages. The failure of the energy calculations presumably stems from the difficulties

in modeling protein–solvent interactions, the use of analytical functions to approximate the chemical potential and compounding of these errors in the computed gradient, and because the energy surface has multiple minima, which makes it nearly impossible to locate the global minimum.

In the combinatorial approach it is generally assumed that step 1, the secondary structure prediction, has been successful, and attention is mainly directed to step 2, the docking of secondary structures into a nativelike three-dimensional structure. Initially, a list of trial structures is generated by packing all combinations of the α helices and β strands. The native fold will be in this list, and structures are eliminated that violate stereochemical rules when applied to the packing of α helices in myoglobin (Richmond and Richards, 1978; Cohen *et al.*, 1979; Cohen and Sternberg, 1980a). In predicting the structure of myoglobin, over 10^8 trial structures were generated by docking together hydrophobic patches on the surface of the α helices. However, only 20 folds did not violate the steric and connectivity constraints. The addition of two distance constraints obtained from experimental data on heme binding further restricted the number of allowed structures to two. The relative arrangements of α helices in one of these structures closely resembled that in the native structure. The root-mean-square deviation (Cohen and Sternberg, 1980b) between predicted structure and the native protein was 4.3 Å. With this same combinatorial approach, an algorithm has been written to predict β-sheet structures in proteins, e.g., β_2-microglobulin and an HLA-B7 antigen fragment (Cohen *et al.*, 1980).

The analysis and prediction of the structural motifs in the glycolytic enzymes were reviewed by Sternberg *et al.* (1981). Ten of the 13 enzymes in this pathway have been studied by x-ray crystallography. It is shown that all the enzyme structures are variations and extensions of a basic theme of a many-stranded (four to nine), predominantly or totally parallel β sheet that is shielded from solvent by α helices (i.e., α/β structure). There were strong structural similarities between the domains of some but not all enzymes. In particular, the dinucleotide-bonding fold of lactate dehydrogenase and the β barrel of triose phosphate isomerase are found in other domains. General rules governing the topology and packing of α helices against β sheet provided a basis for the combinatorial prediction of the tertiary fold of glycolytic domains from their amino acid sequence and observed secondary structure. The prediction algorithm demonstrates that there are severe restrictions on the number of possible structures. However, these restrictions do not fully explain some of the remarkable structural similarities between different enzymes that probably result from evolution from a common ancestor.

Taylor and Thornton (1984) devised a procedure to recognize supersecondary structure in protein sequences. The term supersecondary structure was introduced by Rao and Rossmann (1973) to describe larger continuous folds found in proteins. An idealized template derived from known supersecondary structure was used to locate probable sites by matching secondary structure probability profiles. The method was applied to the identification of βαβ units in β/α type proteins with 75% accuracy. The location of supersecondary structure was then used to refine the original Garnier *et al.* (1978) secondary structure prediction, resulting in an 8.8% improvement, which correctly assigned 83% of secondary structure elements in 14 proteins. Slight modifications of the Garnier *et al.* (1978) method were suggested, producing a more accurate identification of protein class and a better prediction of β/α-type proteins. A method for the incorporation of hydrophobic information into the prediction was also described. Taylor (1984) published an algorithm to compare secondary structure predictions rather than percentages of all residues correctly assigned. This procedure is more important if the tertiary structural prediction is the subsequent goal.

Vonderviszt *et al.* (1986) developed a simple approach to domain border prediction,

relying only on the amino acid sequence. Statistically determined sequential and association preference data of amino acids were combined to generate short-range preference profiles along the polypeptide chains. Domain boundaries correlate with the minima of preference profiles, but some false minima also exist. Preferences of pairs of amino acids and amino acids separated from each other by one residue were obtained. Statistically determined association potentials (E_{ij}) of Narayana and Argos (1984) were used to describe interaction between amino acids.

Hones *et al.* (1987) tried a manual prediction of glucose dehydrogenase based on the hydropathic nature of the internal β sheet and the amphilic character of external helices. The overall homology of primary structure between this enzyme and lactate dehyrogenase was low, and independent predictions of secondary structure produced different patterns of β strands and α helices. However, studies on physicochemical and chemical modification indicated similarity of structure. This method led to the identification of analogues of all the β strands present in lactate dehydrogenase with one exception.

Cohen and co-workers (Hurle *et al.*, 1987) have applied their combinatorial approach to tertiary structural prediction and combined it with circular dichroism measurements to establish the class of the protein. Hurle *et al.* (1987) used vacuum circular dichroism to assign the α subunit of tryptophan synthetase to the α/β class of supersecondary structure. The two-domain structure of the α subunit was assumed based on the work of Miles *et al.* (1982) and Beasty and Matthews (1985), which eliminated consideration of a barrel structure and focused attention on a β-sheet structure. With the algorithm of Cohen *et al.* (1983) a secondary structure was predicted. By the use of other algorithms, the final structure was predicted to have a parallel β sheet flanked on both sides by α helices. This did not agree with the subsequently obtained x-ray structure, which was shown to be a α/β barrel (D. Davies and C. Hyde, quoted in Hurle *et al.*, 1987). Cohen *et al.* (1986b) applied their scheme to predict the three-dimensional structure (Cohen *et al.*, 1979; Cohen and Sternberg, 1979) of interleukin-2. This combinatorial approach generated 3.9×10^4 structures. Of these, 27 satisfied steric constraints and maintained the connectivity bridge between Cys^{58} and Cys^{105}. This allowed the structure to be placed into five structural categories, which contained the right-handed cylinders for a fourfold α-helical bundle. Circular dichroism measurements placed the protein into the α-helical class. Thus, the most plausible structure for interleukin-2 is a right-handed fourfold α-helical bundle.

The three-dimensional structure of interleukin-2 has subsequently been solved to a 3.0-Å resolution (Brandhuber *et al.*, 1987). Four helices form an apparent antiparallel helical bundle that differs significantly from the classical four-helix bundle represented by several proteins (Richardson, 1981). The packing of regions of the helices is shorter, involving only three to four turns of the helix, whereas classical four-helix bundles usually have at least five turns in each helix. There are also three other helical regions.

Billeter *et al.* (1987b) have applied the new method of constrained optimization, known as the "ellipsoid algorithm" (Shor, 1977), to the problem of the docking of two molecules. This method is efficient with respect to computer time and is robust when dealing with nonconvex problems. Its ability to make large steps, especially at the beginning of the optimization, and the fact that only one violated constraint is used in any iteration makes it a powerful tool for problems that involve many local minima and for which no good starting points are available (e.g., Ecker and Kupferschmid, 1982). Although the maximum number of variables that can be used has not yet been determined, problems with up to 55 torsion-angle variables have been solved (Billeter *et al.*, 1987a). Billeter *et al.* (1987b) used this method to explore it as a tool for determining sterically acceptable interactions between two molecules. These interactions are described by constraints on intermolecular distances. To specify the relative orientation of the two molecules, a new set of variables had to be introduced, which

represents the set of all possible rotations in a vector space as required by the ellipsoid algorithm. Applications discussed include the docking of two macromolecules and the formation of an enzyme–inhibitor complex. Previously, Kuntz *et al.* (1982) had designed an algorithm that matches the geometry of a binding site to that of a ligand, docking a rigid ligand to a rigid receptor. This method was recently adapted (DesJarlais *et al.*, 1986) to allow for some flexibility in the ligand by docking fragments of the ligand and screening the docked fragments for subsets that can be rejoined to yield the complete molecule with acceptable stereochemistry.

B. β Sheets

The β sheets observed in globular proteins exhibit an extraordinary diversity of structural forms. Salemme (1983) has reviewed the structural properties of protein β sheets and has presented a clear and incisive review of this secondary protein structure. The classical flat β-sheet arrangement originally described by Pauling and Corey (1951) is in contrast to the great variety of twisted and curved surfaces found in protein β sheets. Salemme (1983) reviewed the operative forces and constraints that produce different twisted β-sheet geometries. Briefly, the following factors contribute to the complexity of β sheets. Because the repeating subunits of the chains are chiral, they tend to assume minimum-energy conformations whose effect is to twist the polypeptide chains away from the twofold helical conformation that produces flat sheets. This in turn results in the introduction of energetically unfavorable distortions into the interchain hydrogen bonds. The final configuration of the sheet represents an energetic compromise between optimizing the conformational energy of the polypeptide chains and preserving the interchain hydrogen bonds (Weatherford and Salemme, 1979). The final geometry of the sheet therefore depends on the specific features of interchain hydrogen bonding that constrain the possible ways in which the polypeptide chain can twist. The great variety of sheet geometries observed in globular proteins reflect alternate ways of reaching this compromise. Because of differences in parallel and antiparellel sheet symmetry properties, the effects of chain twisting can produce quite different structural geometries under the corresponding different constraints imposed by hydrogen bonds.

Chothia (1973) showed that, in general, the β sheet with a right-handed twist, when viewed along the polypeptide chain direction, has a lower free energy than sheets that are straight or have a left-handed twist. Ananthanarayanan and Bandekar (1976) attempted to predict the β regions in 16 globular proteins by applying the one-dimensional Ising model of Lifson and Roig (1961). The parameters for the theory were derived from the statistical data on globular proteins given by Chou and Fasman (1974a,b). Results obtained compared favorably to those from other methods, but it was pointed out that not considering the long-range interactions in their method and in other methods based on short-range interactions would make these methods incomplete and incapable of being uniformly applicable to all proteins. Finkelstein *et al.* (1970) showed that the domains constructed from antiparallel β structures can have only an extremely limited set of topologies depending primarily on the number of β portions and the localization of the β hairpin initiating the formation of the protein structure. Ptitsyn *et al.* (1979) proposed a folding mechanism for β proteins on physically reasonable assumptions. According to this mechanism, the folding pathway and final protein topology depend only on the total number of β strands and on the localization of the initiating complex in the given protein chain. This reduces the number of possible topologies for β protein from 10^2 or 10^8 (depending on the number of β strands) to only a few. Lifson and Sander (1979) determined the frequencies for the 20 amino acids separately for antiparallel and parallel arrangements of strands.

Parallel (β_P) and antiparallel (β_A) arrangements of strands in a sheet differ in the hydrogen pattern between strands and in the type of chain connectivity they allow: short reverse-turn connections for β_A and longer crossover connections for β_P (Levitt and Chothia, 1976; Richardson, 1977; Sternberg and Thornton, 1977a–c). The distinction between the two arrangements results in strikingly different sets of preference parameters, including some of the largest values reported so far for any substructure. These authors suggested the use of these new parameters; beyond secondary structure prediction, the different preferences for β_A and β_P may aid in predicting the tertiary interaction between strands. Lifson and Sander (1980a,b) also determined the frequency of occurrence of nearest-neighbor residue pairs on adjacent antiparallel and parallel strands in 30 known protein structures. These were studied by statistical methods. The largest and most significant correlations were: Ser/Thr (1.9 ± 0.3), Ile/Val (1.7 ± 0.3), and Lys/Arg/Asp-Gln (1.8 ± 0.3) in β_A and Ile/Leu (1.9 ± 0.4) in β_P. The pair Gly/Gly never occurs in any β sheet.

Previous attempts at analyzing the tertiary structure of β sheets by statistical means have been of two types: (1) analysis of sheet topology and (2) analysis of the amino acid content of β sheets. Schulz and Schirmer (1974), Richardson (1976, 1977), Levitt and Chothia (1976), and Sternberg and Thornton (1977a,b) have analyzed known β sheets as to length, direction, and number of strands, ordering of strands within the sheet, length, type, and handedness of crossover conditions, and statistical significance of the occurrences of folding units containing β strands. Sternberg and Thornton (1977c) observed that the most hydrophobic strands tend to occur at the center of the sheet and put forward the hypothesis of the hydrophobic ordering of strands. Von Heijne and Blomberg (1977, 1978) analyzed pair correlations among hydrophobic neutral and polar classes of residues. They found that interstrand pairs between residues of the same class occur more often than expected by random chance. They concluded from their observations "that inter- and intrastrand nearest neighbor interactions of a rather unspecific character are responsible for the main stabilizing forces in β sheets." Lifson and Sander (1980a) challenged this conclusion and report pair correlations to the level of specific individual recognition among the most frequent amino acid residues. Less frequent residues are grouped according to their resemblance in size, structure, polarity, and genetic exchangeability (Sander and Schulz, 1979). For antiparallel strands, with a database of 788 residue pairs, the statistical analysis of pair correlations involves seven individual residues, five groups of two residues each, and one group of three residues. For parallel strands, with a data base of only 263 pairs, the groups are more extensive.

Chothia and Janin (1982) studied the packing of β-sheet-to-β-sheet in a family of proteins that are formed by two β-pleated sheets packed face to face. Concanavalin A, plastocyanin, α-crystallin, superoxide dismutase, prealbumin, and the immunoglobulin domains are representative members of this family. They found that when β-pleated sheets pack face to face in proteins, the angle between the strand directions of the two β sheets is near $-30°$. They propose a simple model that shows how the relative orientation of two packed β sheets is a consequence of (1) the rows of side chains at the interface being approximately aligned and (2) the β sheet having a right-handed twist. The special amino acid composition of residues at the β-sheet-to-β-sheet interfaces makes the contact surfaces essentially smooth and hydrophobic. Cohen *et al.* (1981) examined the tertiary structure of ten β sandwiches formed by face-to-face packing of two primarily antiparallel β sheets. They formed a well-defined structural class with the following features: (1) a standard packing geometry with the two twisted β sheets separated by 8.3 to 10.3 Å and rotated counterclockwise by 20° to 50° and (2) common values and positions for changes in solvent-accessible contact area during the condensation of β strands → β sheets → β sandwich. Sheet packing produces anticomplementary patterns of area changes. (3) The sheet–sheet interface has a bilayer structure when medium-

sized residues in the two sheets stack, but there is interdigitation of large and small residues in different sheets; and (4) the twisted nature of the β sheet explains both the left-handed rotation between the two sheets and the observed anticomplementary pattern of area changes.

These observations have been incorporated into a computer algorithm to predict the tertiary fold of β sandwiches from primary and secondary structure. Salemme and Weatherford (1981a,b; Salemme, 1981) studied the conformational and geometric properties of idealized β sheets. In a series of three papers, the parallel, antiparallel and mixed, and isotropically stressed β sheets were studied. Comparison of observed parallel β structure with the idealized models showed that to the extent that the observed structures are regularly hydrogen bonded, they are closely approximated by the models. The long-range geometric configuration of twisted sheets was shown to be primarily an equilibrium between the forces that cause the individual polypeptide chains to twist in order to minimize their local conformational potential energies and the requirements for interchain hydrogen bonding, which generally resist the introduction of twist into the sheet. Antiparallel structures possess conformational degrees of freedom that allow them to assume a greater diversity of spatial configurations than occur in parallel sheets.

Hol *et al.* (1981) demonstrated that as a result of the regular arrangement of peptide dipoles in secondary structure segments and low effective dielectric constant in hydrophobic cores, the electrostatic energy of a protein is very sensitive to the relative orientations of the segments. Evidence was provided that the alignment of secondary structure dipoles is significant in determining the three-dimensional structure of globular proteins. In globular proteins containing only β sheet there is an overwhelming tendency to align in an antiparallel manner, and all peptide dipole moments cancel each other, with no residual dipole in either direction. In a parallel arrangement, the components of the dipole moment parallel to the strand direction interact unfavorably with each other. This effect in the antiparallel strand may only be ~0.4 kcal/mole for a pair of short β strands containing four residues. This gives a difference in electrostatic energy of ~0.8 kcal/mole between parallel and antiparallel arrangements in three strands, which is smaller than the energy difference for α helices, but it may nevertheless be significant. Chou *et al.* (1982) minimized the energy of two- and three-chain antiparallel and parallel β sheets. All computed minimum-energy β sheets were found to have a right-handed twist, as observed in proteins. As in the case of the right-handed α helices, it is the intrastrand nonbonded interaction energy that plays a key role in forcing β sheets of L-amino acids to adopt a right-handed twist. The minimized energies of parallel β sheets are considerably higher than those of the corresponding antiparallel β sheet, indicating that parallel β sheets are intrinsically less stable. The energy difference between antiparallel and parallel β sheets arises from closer packing of the chains and a more favorable alignment of the peptide dipoles in the antiparallel structures.

Chothia and Lesk (1982a,b) have studied the family of small copper-containing proteins plastocyanin and azurin, which are active in the electron transport systems of plants and bacteria. Both proteins contain two β sheets packed face to face. Using computed superpositions of the structures, they aligned the sequences, identified homologous positions, and studied how the structures have changed as a result of mutations. The residues in the vicinity of the copper-bonding site show minimal amino acid substitution and form almost identical structures, whereas other portions are more variable in sequence and structure. Buried residues tend to maintain their hydrophobic character, but mutations change their values. Eleven immunoglobulin domains whose x-ray structures were known were also studied. Mutations caused (1) displacements and rotations of the β sheets relative to each other up to 2 Å and 20°, (2) lateral insertions of side chains from extended loops into the interface regions to compensate for reductions in the volume of β-sheet residues, (3) insertion of a residue into a strand to

form a β bulge, (4) local changes in conformation, and, only rarely, (5) complementarity in adjacent mutations. The mutations of interior residues are accommodated through substantial structural changes consistent with the preservation of their function.

Chothia and Janin (1982) distinguished two classes of β-sheet-to-β-sheet packing in globular proteins. Both classes have β sheets with the usual right-handed twist packed face to face. In orthogonal β-sheet packings, the strand directions of the different β sheets are 90° to each other. Twisted sheets in this orientation have anticomplementary surfaces: one pair of diagonally opposite corners in β sheets is very close, and the other pairs of corners splay apart. At the close corner, the β sheets are usually covalently connected, connected by a β bend. Contacts between the β sheets occur along the diagonal joining the close corners. They involve about one quarter of the β-sheet residues, and two-thirds of them are Val, Ile, or Leu. In aligned β-sheet packings, the angle between the strand's directions of the packed β-sheets is ~30°. In this orientation, the twisted β-sheet surfaces are complementary. Novotny *et al.* (1984) presented a twisted hyperboloid (strophoid) as a model of β barrels in proteins. With a least-squares fitting procedure, polypeptide backbones of one parallel and seven antiparallel β barrels were approximated with various curved surfaces. Although the hyperboloid gave better approximations to all the β-barrel backbones than the ellipsoid, elliptical cylinder, or catenoid, the best approximations were obtained with a novel surface, a twisted hyperboloid (strophoid).

Chothia *et al.* (1985) studied the domain association in immunoglobulin molecules, which have variable domains. Approximately three quarters of the interfaces between VL and VH domains are formed by packing of VL (variable light chains) and VH (variable heavy chains) β sheets in the conserved "framework," and one quarter from contacts between the hypervariable regions. The β sheets that form the interface have edge strands that are strongly twisted (coiled) by β bulges. As a result, the edge strands fold back over their own β sheet at two diagonally opposite corners. When the VL and VH domains pack together, residues from these edge strands form the central part of the interface and give what Chothia *et al.* called a three-layer packing; i.e., there is a third layer composed of side chains inserted between the two backbone side-chain layers that are usually in contact. This three-layer packing is different from previously described β-sheet packings. The 12 residues that form the central part of the three observed VL–VH packings are absolutely or very strongly conserved in all immunoglobulin sequences. This strongly suggests that this structure is a general mode for the association of VL and VH domains and that three-layer packing plays a central role in forming the antibody-combining site.

Garratt *et al.* (1985) applied the prediction algorithm of Garnier *et al.* (1978) to 16 proteins whose structures are dominated by β sheet. Comparisons of the predicted structures with those defined by Kabsch and Sander (1983a,b) showed that for β-sheet residues, the quality of prediction falls off markedly with increasing residue accessibility. Two subclasses (internal and external) of β residues have been distinguished on the basis of hydrogen-bonding patterns, and the distribution of amino acid types within each subclass was found to be different. Thus, the Chou and Fasman (1974a,b) P_β-type parameters for these previously indistinguished states have been derived.

Getzoff *et al.* (1986) presented a qualitative computer graphics approach to the characterization of forces important to the assembly of β domains that should have general utility for examining protein interactions and assembly. In their approach, the nature of the molecular surface buried by the domain contacts, the specificity of the residue-to-residue interactions, and the identity of electrostatic, hydrophobic, and hydrophilic interactions were elucidated. These techniques were applied to the β-barrel domains of Cu,Zn-superoxide dismutase, immunoglobulin Fab, and tomato bushy stunt virus protein. Strong β-domain interactions (identified

by their biochemical integrity) apparently result from chemical, electrostatic, and shape complementarity of the molecular surfaces buried from interaction with solvent molecules. Electrostatic forces appear to be important in both stabilizing and destabilizing specific contacts.

Taylor (1986a) described a pattern-matching procedure, based on fitting templates to the sequence, that allows general structural constraints to be imposed on the patterns identified. The templates correspond to structurally conserved regions of the sequence and were initially derived from a small number of related sequences whose tertiary structures are known. The templates are then made more representative by aligning other sequences of unknown structure. Two alignments were built up containing 100 immunoglobulin variable-domain sequences and 85 constant-domain sequences, respectively. From each of these extended alignments, templates were generated to represent features conserved in all the sequences. These consisted mainly of patterns of hydrophobicity associated with β structure. For structurally conserved β strands with no conserved features, templates based on general secondary structure prediction principles were used to identify their possible locations. The specificity of the templates was demonstrated by their ability to identify the conserved features in known immunoglobulin and immunoglobulin-related sequences but not in other nonimmunologic sequences.

C. Packing of α Helices (α/α)

One of the first attempts to fold secondary structures into the native tertiary fold was that of Ptitsyn and Rashin (1975), who predicted the folding of the α helices of myoglobin. Using the crystallographic assignments of α helices, they docked the α helices to bury the hydrophobic residues maximally. One of the most favorable structures obtained in this manner coincided in rough resolution with the native tertiary fold. Richmond and Richards (1978) also examined the helical packing in sperm whale myoglobin. The approach of two helices along the contact normal connecting their axes produced solvent-exclusive effects at a distance of about 6 Å from the final position. The solvent-excluded area formed in such an interaction site is equivalent to a large hydrophobic contribution to the free energy of association. Helices of close-packed spheres form useful approximations to actual peptide helices. On the basis of the geometry implied by the close-packed-sphere helix, an algorithm was proposed for picking potentially strong helix–helix sites in the peptides of known sequence. When combined with preliminary secondary structure predictions, this algorithm might usefully restructure the search for these specific types of contact in the docking portion of a general fold program.

Efimov (1979) likewise considered the packing of α helices in globins. It was shown that close packing of hydrophobic side chains on the surface of an individual α helix or on the surface of the bihelical structure is obtained at a certain combination of rotational isomers. This allowed the prediction of rotational isomers in α-helical regions of proteins. Cohen *et al.* (1979) devised a computer program to fold a peptide chain consisting solely of helical segments and connecting links of known length using myoglobin as the example. This was an extension of the earlier work of Ptitsyn and Rashin (1975) and Richmond and Richards (1978). The helices are paired according to the list of potential sites, with each helix paired at least once. The list of pairs is then examined geometrically, with two filters being used: (1) lengths of connecting links must be equal to or greater than the end-to-end distances of helices, and (2) nonpaired helices must not collide. Adjustment of parameters reduced the final number of possible structures to 20, one of which closely resembles the actual distribution of helices in myoglobin. Cohen and Sternberg (1980a) added two new filters to their predictive method (Richmond and Richards, 1979). The use of chemical information to constrain the distal–proximal histidine separation aided in the prediction of the structure of myoglobin. Out of 20

structures, only two very similar structures satisfied this additional filter. The two remaining structures had root-mean-square deviations of 4.48 and 4.53 Å from the crystal structure. Cohen and Sternberg (1980b) reported a method and quantified the significance of obtaining a specific root-mean-square deviation when folding proteins of different molecular weight. The average root-mean-square deviation was found to be proportional to the number of residues, and this correlation was explained by a mathematical model.

The analysis of the pattern of residue-to-residue contacts at the interface of 50 helix-to-helix packings observed in ten proteins of known structure was carried out by Chothia *et al.* (1981). This analysis supported a model for helix-to-helix packing in which the ridges and grooves on the helix surface intercalate. These ridges are formed by rows of residues whose separation in sequence is usually four, occasionally three, and rarely one. The model explains the observed phenomenon of packings whose interhelical angle is ~50°. Of the 50 packings, 38 agree with the model, and the general features of another ten packings were described by an extension to the model in which ridges can pack across each other if a small side chain occurs at the place where they cross. Before this paper there were only three other detailed models for helix-to-helix packing, those described by Crick (1953), by Efimov (1977, 1979), and by Richmond and Richards (1978). Crick's (1953) knobs-into-holes model for helix packing was developed as part of the coiled-coil model for α-keratin. Efimov (1977, 1979) has described two models for helix packing: the polar and the apolar. In the apolar model, side chains are clustered on one face of a helix by residues i, $i + 4$, and $i + 8$ having a *trans* conformation, $\chi \simeq 180°$, and residues $i + 1$ and $i + 5$ having a *gauche* conformation, $\chi_1 \simeq -60°$. If side chains are represented by spheres, he shows that the packing of two such helices gives the torsion angle between helices, Ω, values of +30°, −30°, and 90°. Efimov (1979) states that this model (Efimov, 1977) coincides in principle with that described by Crick (1953), by Chothia *et al.* (1977), and by Richmond and Richards (1978).

A domain of four α helices packed into a right-handed bundle has been proposed to be a recurrent motif in protein structure (Sternberg and Cohen, 1982; Sheridan *et al.*, 1982; Efimov, 1982a). The fourfold α-helical structure had been previously suggested (Argos *et al.*, 1977; Weber and Salemme, 1980) to be similar to that observed in several proteins, e.g., the disk of tobacco mosaic virus (Bloomer *et al.*, 1978). Sternberg and Cohen (1982) predicted the secondary and tertiary structures of four interferons from homologous amino sequences. Three methods of predicting the secondary structure were used (Lim, 1974a–c; Chou and Fasman, 1978a; Garnier *et al.*, 1978) and gave differing results, but all of them suggested four α helices. The prediction of Hayes (1980) for the interferons also agreed in principle with that of Sternberg and Cohen (1982). The algorithm for docking α helices into a tertiary fold (Richmond and Richards, 1978; Cohen *et al.*, 1979; Cohen and Sternberg, 1980a,b), which was slightly modified, substantiated the proposed right-handed four-helical bundle model. One model was shown to be compatible with the known disulfide linkages in interferon. It should be noted that there are instances of left-handed bundles formed from four α helices (Blow *et al.*, 1977; Matthews *et al.*, 1971; Remington *et al.*, 1978; McLachlan *et al.*, 1980).

Sheridan *et al.* (1982) developed a simple dipole model for estimation of the electrostatic interactions between α helices in the protein tertiary structural motif of an array of four closely packed α helices. It was found that, for the proteins examined (cytochrome c', hemerythrin, myohemerythrin, cytochrome b_{562}, and a T4 phage lysozyme domain), their common anti-parallel arrangement of adjacent helices confers a stabilization of 5–7 kcal/mole. In contrast, a similarly packed array of parallel helices is relatively destabilized by 20 kcal/mol. Hol *et al.* (1981) has calculated the electrostatic interactions between the peptide groups of both α helices and β sheet in various proteins and demonstrated that secondary structures pack in a manner to provide significantly favorable electrostatic energy. Efimov (1982a,b) demonstrated that four-

helix complexes can form two "mirror-symmetric" structures. A consideration of hydrophobic interactions, hydrogen bonds, and salt bonds is not sufficient for an unequivocal choice of one of these two structures. It was shown that severe stereochemical restrictions on the packing of the α helices are imposed by the length of the interhelical regions, i.e., the constrictions.

Murzin and Finkelstein (1983) considered the packing of α-helical portions in globules with a single monobound nucleus. A scheme for the description of α-helical complexes by a system of "longitudinal" and "transverse" interhelical contacts was proposed. It was shown that the closed helical globules are well described by quasispherical polyhedra with the ends of the α helices serving as their apices and the axes of the α helices and the lines of contacts between their ends as ribs. It was found that the cross pieces between the helical portions run along certain ribs of such polyhedra. A novel supersecondary structure, referred to as an αα corner, was described by Efimov (1984). The αα corner is formed by two consecutive α helices packed approximately crosswise and connected by two or more peptide units. It was shown that the amino acid sequences coding for the αα corners have a strictly definite order of hydrophobic, hydrophilic, and Gly residues. A hypothesis was suggested that the αα corner can be an embryo of protein folding.

Cohen and Kuntz (1987) applied a series of heuristic algorithms to the sequence of human growth hormone. A family of five structures that are genetically right-handed fourfold α-helical bundles was found from an investigation of 10^8 structures. A plausible receptor binding site was suggested. Circular dichroism studies showed a secondary structure consisting of 45–50% α helix and no β structure, thus placing this protein in the α/α class of proteins. The turn algorithm of Cohen *et al.* (1986a,b) using the patterns for α/α domains (AA-TURNS; Cohen *et al.*, 1982) identified 12 turns. The key positions for hydrophobic residues that favor helicity (Cohen *et al.*, 1982; Richmond and Richards, 1978) were also obtained. Helical boundaries were defined by the delimit methods (Cohen *et al.*, 1983). Helix–helix interactions, which were divided into three classes by the helix-packing algorithm of Cohen *et al.* (1979), and helix docking sites were located by the method of Richmond and Richards (1978). This list was further processed using the steric restriction of Cohen *et al.* (1979). The interaction site central residue, interaction class, and secondary structure provide the input for the combinatorial helix-packing algorithm of Cohen *et al.* (1979). Distance restraints were used to further eliminate possible structures (Havel *et al.*, 1983; Cohen and Sternberg, 1980a,b). The surface and volume characteristics of proteins have indicated that proteins do not have a high surface-area-to-volume ratio (Richards, 1977). Thus, 119 of the structures generated have high surface-area-to-volume ratios and long loop excursions, so these were eliminated. The remaining 67 structures fall into the fourfold helical bundle motif. Five of the remaining structures have the right-handed topological preference (Weber and Salemme, 1980) and were kept and have small variations. The application of the helix dipole for stability (Shoemaker *et al.*, 1987) further assisted in establishing the connectivity of the four helices.

D. Amphipathic α Helices and β Strands: Dipole Moments and Electrostatic Interactions

Segrest and Feldman (1977) presented a computer analysis of the general occurrence of amphipathic helix patterns in proteins with known sequence. An α helix is considered amphipathic when it can be divided into separate polar and nonpolar faces. A specific distribution of charged residues along the polar face is frequently observed. A minimal residue length of 11 was set for selection of amphipathic helices. The *Protein Sequence Data Tape* (1972) plus *Supplemental* (1973), National Biomedical Research Foundation, Georgetown University

Medical Center, Washington, D.C. 20007, was used. A total of 649 amino acid sequences had two distinct properties that could be quantified: the number of ion pairs and the degree of hydrophobicity of the potentially nonpolar face. Positively charged lysyl and arginyl residues occurred at the interface between the polar and nonpolar "faces" of each amphipathic sequence. Amphipathic helices were found to occur with much greater frequency in known lipid-binding proteins than in proteins in general.

Fourier analysis of the hydrophobicities (Kyte and Doolittle, 1982; Eisenberg *et al.*, 1982a,b) of the acetylcholine receptor subunits by Finer-Moore and Stroud (1984) revealed regions of amphipathic secondary structure. Prediction of a consensus secondary structure based on this analysis and on empirical prediction methods (Garnier *et al.*, 1978) for the sequence external to the bilayer ($\simeq$75%) and separate evidence for oriented helices in the transmembrane regions led to a testable hypothesis about how the ion channel is formed and might function. Previously McLachlan and Karn (1983) had discussed the use of Fourier analysis as a means of identifying periodic properties of amino acid sequences. The periodicity of the amphipathicity was found to be exactly that expected for an α helix at least 30, and up to 50, amino acids in length. The predicted overall secondary structure of the acetylcholine receptor has 27% β sheet and 44% α helix, in rough agreement with both circular dichroism studies [29% β, 34% α (Moore *et al.*, 1974)] and Raman spectroscopy studies [34% β, 25% α helix, 14% disordered helix (Chang *et al.*, 1983)]. They proposed a mode of assembly of the acetylcholine receptor complex from its subunits in which the transmembrane helices, H1–H3, are hydrophobic and could be inserted into the bilayer during synthesis. As the subunits associate, the COOH-terminal hydrophobic helix H4 and amphipathic helix A could fold into the bilayer looped together. Thus, a hydrophilic lipid-free channel could be formed between subunits from residues that would be unstable in the individual subunit's interface with lipids. Correct ion pairing could encode correct multisubunit assembly.

Krebs and Phillips (1984) showed that the amphiphilicity of an α-helical segment in a protein may be quantitated by calculating its mean helical hydrophobic moment (μ_H) (Eisenberg *et al.*, 1982a,b). For proteins whose hydrophobic interactions with interfaces are mediated by α helices, the surface pressure exerted at the air–water interface correlates with the product ($\mu_H \cdot F$), where μ_H is the mean helical hydrophobic moment averaged over all helices in the entire molecule and F is the fraction of α helix in the protein. Knowledge of μ_H permits a description of amphipathic α helices and their surface activities at the air–water interface of serum apolipoproteins, surface-seeking peptides, and globular water-soluble proteins. Jähnig and co-workers (Vogel and Jähnig, 1986; Vogel *et al.*, 1985) estimated the secondary structure of the lactose permease of *Escherichia coli* reconstituted in lipid membranes by Raman spectroscopy. They obtained an estimate of ~70% α-helix content, β-strand content below 10%, and a β-turn contribution of 15%. About one third of the residues in α helices and most other residues are exposed to water. Then, by a method for structural prediction that accounts for amphipathic helices (Eisenberg *et al.*, 1984a,b; Finer-Moore and Stroud, 1984), ten membrane-spanning helices were predicted that are either hydrophobic (Kyte and Doolittle, 1982) or amphipathic. These are expected to form an outer ring of helices in the membrane, the interior of which would be made of residues that are predominantly hydrophilic and, by analogy to sugar-binding proteins, suited to provide the sugar-binding site. The β-turn profile (Chou and Fasman, 1978a) was also utilized.

Vogel and Jähnig (1986) also determined the secondary structure of porin, maltoporin, and OmpA protein reconstituted in lipid membranes by Raman spectroscopy. The three proteins have similar structures consisting of 50 to 60% β strand, about 20% β turn, and less that 15% α helix. By a method for structural prediction that accounts for amphipathic β strands, folding models were developed for porin and for a segment of OmpA protein incorporated into a membrane. In the model, the OmpA fragment consists of eight amphipathic

membrane-spanning β strands that form a β barrel. Similarly, porin is folded into ten amphipathic membrane-spanning β strands that are located at the surface of the trimer towards the lipids and eight predominantly hydrophilic strands in the interior. The search for stretches of 20 amino acid residues that would form a membrane-spanning α helix used the Kyte and Doolittle (1982) method. To ascertain β-strand amphipathicity, the method for elaboration of amphipathic helices (Eisenberg *et al.*, 1984a,b; Finer-Moore and Stroud, 1984; Vogel *et al.*, 1985) was modified. The secondary structure of residues exposed to water and the β turns was predicted by the Chou and Fasman (1978a) algorithm.

Cornette *et al.* (1987) developed computational techniques for detecting amphipathic structures in proteins. Protein segments that form amphipathic α helices have periodic variation in the hydrophobicity value of the residues along the segment, with a 3.6-residue-per-cycle period characteristic of an α helix. They compared the usual method for detecting periodicity based on a discrete Fourier transform with a method based on a least-squares fit of a harmonic sequence to a sequence of hydrophobicity values. The analogue to the usual Fourier transform power spectrum is the "least-squares power spectrum," the sum of squares accounted for in fitting a sinusoid of given frequency to a sequence of hydrophobicity values. The sum of the spectra of the α helices in their data base peaked at 97.5°, and approximately 50% of the helices can be accounted for in this peak. Thus, approximately 50% of the α helices appear to be amphipathic, and, of those that are, the dominant frequency at 97.5° rather than 100° indicates that the helix is slightly more open than previously thought, with the number of residues per turn closer to 3.7 than 3.6. This extra openness was examined in crystallographic data and was shown to be associated with the C terminus of the helix. The α amphipathic index, the key quantity in their analysis, measures the fraction of the total spectral area that is under the 97.5 peak and is characteristic of hydrophobicity scales that are consistent for different sets of helices. They developed an optimized hydrophobicity scale that maximizes the amphipathicity index and has a correlation of 0.85 or higher with nine previously published scales. Although the scale is optimal only for predicting α amphipathicity, it also ranks high in identifying β amphipathicity and in distinguishing interior from exterior residues in proteins.

The possibility of certain residues exerting helix-forming influence in either the COOH-terminal or NH_2-terminal direction preferentially was suggested by the nonrandom distribution of amino acid residues between the two ends of helical regions in globular proteins (Cook, 1967; Ptitsyn, 1969). Robson and Pain (1972) confirmed that this directional effect exists and characterized this effect for individual residues and its dependence on distance along the polypeptide chain. A study of primary sequence–conformation relationships by information theory analysis involved the estimation of an information function, $I(S_j; R_{j+m})R_j$ (Robson and Pain, 1971). This describes the information in residue R_{j+m} at position $j + m$ (where $-8 \leq m \leq +8$) concerning the conformation S_j of residue R_j and will include any directional effect, since the value of the function will be dependent on the sign of m. The function $i(S_j; R_{j+m})$ is the information in the residue at position $j + m$ concerning the configuration per average-type residue, which could be at position i. The analysis of 11 proteins showed which residues showed well-defined characteristics: i.e., Ala, Leu, and Ile exhibit a tendency to support helix formation in both directions; Pro disrupts helices in the NH_2-terminal direction and, interestingly, has a definite tendency to potentiate helix formation in the COOH-terminal direction; Asp and Glu show strong tendencies to helix disruption in the NH_2-terminal direction; Glu exhibits a strong directional influence in that it supports helix formation in the COOH-terminal direction, whereas the influence of Asp is of marginal significance; basic amino acids Lys and marginally His and Arg show tendencies to support helix formation in the NH_2-terminal direction. Other groups have also studied the tendency of certain side chains to cluster at the end of helices (Lewis *et al.*, 1971; Crawford *et al.*, 1973).

Lewis and Bradbury (1974) found the attractive and repulsive electrostatic interactions of

the *i*th residue with its neighbors $i \pm 1$, 2, 3, 4, and 7 to be helix-breaking if there was more than one net repulsion. Maxfield and Scheraga (1975) studied the effect of neighboring charges on the helix-forming ability of charged amino acids in proteins. They found helix-disruptive effects of some charged residues at the $i \pm 4$ position, whereas helix-stabilizing effects were found with the oppositely charged residues at $i \pm 2$ and $i \pm 3$.

Chou and Fasman (1974a) further characterized this directional propensity and evaluated the frequency of helical boundary residues in 15 proteins. Their results were in general agreement with those of Robson and Pain (1972); however, there were several major differences: Arg, His, and Lys were all found with the highest frequency at the COOH terminus of helices, and Pro, Asp, and Glu were found to have the highest frequency at the NH_2 terminus. Blagdon and Goodman (1975) proposed that helix initiation could occur at the end of helices, i.e., terminal initiation by polar groups and by turns, and this would have a directionality because of the distribution of charge discussed above. The significance of the fairly strong electric fields produced by electric macrodipoles of the α helix and by ionic charges in stabilizing proteins has been recognized for some years (Wada, 1976; Perutz, 1980; Hol *et al.*, 1978; van Duijen *et al.*, 1979). These ideas have been incorporated into the α-helix dipole hypothesis. Wada (1976) discussed the α helix as an electric macrodipole and its importance in protein structure. Hol *et al.* (1978) discussed in detail the role of the α-helix dipole for the properties of proteins. In an α helix the peptide dipoles are aligned nearly parallel to the helix axis, and the axial component is 97% of the dipole moment. Theoretical considerations indicate an increase in the peptide dipole moment by polarization related to hydrogen bonds in the α helix, yielding a value of up to 5 D. The α helix has a considerable electric field, which runs from the C terminus to the N terminus. The strength of the field increases up to a helix length of about 10 Å (two turns), whereafter further elongation has only a marginal effect.

Hol *et al.* (1981) applied the concept of peptide dipoles to protein folding. As a result of the regular arrangement of peptide dipoles in secondary structure segments and the low effective dielectric constant in hydrophobic cores, the electrostatic energy of a protein is very sensitive to the relative orientation of the segments. Evidence was provided that the alignment of secondary structure dipoles is significant in determining the three-dimensional structure of globular proteins. Model calculation of helix–helix interactions showed that electrostatic energy between the backbone dipoles of helices was maximum for antiparallel helices rather that parallel helices. In proteins with alternating α and β structures the α–β dipolar interaction is favorable when helices and strands are antiparallel.

Sheridan and Allen (1980) presented electrostatic potential maps of α helices of various lengths using a point-charge model. It was shown that the potential field of the helix can be mimicked by two equal and opposite charges, one at each terminus. The magnitude of these equivalent charges reaches its limiting value of ±0.2 to 0.3 electron at a helix length of approximately seven to ten residues. Hydration, the presence of counterions, and solvent dielectric would no doubt reduce the influence of both charged residues and the helix dipole, the former probably to a greater extent. In order to have the helix dipole potential as the major electrostatic influence on substrate or coenzyme binding site, a protein must sequester these sites within crevices shielded from solvent and at the same time keep any charged residues at a reasonably large distance from the site. If these conditions are met, anions will find stability at the N termini of helices. On the other hand, at points somewhat removed from binding-site crevices, the greater electrostatic influence is very likely to be the charged residues on the protein surface. That is, helices will stabilize the presence of anions already in the binding site but will not draw them in from a distance.

Wada and Nakamura (1981) analyzed the distribution of the distance separating the ionic charges (ionized groups and the apparent charges at the termini of the α helices) for more that

44,000 charge pairs in 14 proteins. Their results show that charges in the proteins are, on average, surrounded by charges of the opposite sign. Previous calculations of the electric potential near the helix termini have shown that the electrostatic effect of the α-helix dipole is equivalent to the effect of one half of a positive unit charge at the N terminus of the α helix and one half of a negative charge at the C terminus (Hol *et al.*, 1978; Sheridan and Allen, 1980). The macrodipole of the α helix has been found to have the same order of contribution to stabilizing the native protein conformation as ionized groups. These electrostatic interactions have been stated to be involved at active sites of functional proteins in binding substrate or a coenzyme or in enhancing enzymatic reaction rates (Perutz, 1980; Hol *et al.*, 1978; van Duijen *et al.*, 1979; Sheridan and Allen, 1980). Perutz *et al.* (1985), in discussing the pK_a values of two His residues in human hemoglobin, invoke the dipole of the helices to explain the abnormal values. In HbCO A, His, FG(97)β has a pK_a of 7.8 compared to the pK_a value of about 6.6 characteristic of free His at the surface of proteins. This high pK_a is accounted for by its interaction with the negative pole at the C terminus of helices F and FG. It corresponds to a free energy change of the same order as that observed in the interaction of histidines with carboxylate ions and confirms the strongly dipolar character of α helices, which manifests itself even when they lie on the surface of the protein.

Thornton and Sibanda (1983) analyzed the structural, dynamic, and functional aspects of amino- and carboxyl-terminal regions of proteins of known structure. Terminal regions are usually located on the surfaces of the protein, accessible to solvent, and are often flexible. There is a significant preference for terminal regions in single-domain proteins to be in close proximity. This partially reflects the compact globular nature of proteins, but the preference for spatial proximity is stronger in native proteins than in randomly generated structures. In addition, in multidomain and multisubunit proteins, it was found that the terminal regions are commonly involved in the interface between domains and subdomains.

The role of electrostatic interactions is complicated and not readily described analytically. However, efforts have been made to analyze and calculate their contribution to protein folding and stability. As an example, Rogers and Sternberg (1984) have discussed the different dielectric models applied to the packing of α helices. The effect of the α-helix dipole in stabilizing the tertiary structure of globular proteins was examined using three of the commonly used dielectric models. These are (1) the uniform dielectric model, (2) the distance-dependent dielectric model, and (3) the cavity dielectric model. They found the cavity model to be the most reasonable since it attempts to describe the markedly different dielectric responses of the solvent and the protein. The protein is set at a low continuous dielectric value, and the solvent is set at a high continuous dielectric value. It was found that for the cavity model of the dielectric, the calculated interaction energy between two helices is strongly dependent on how exposed the helix termini are to solvent. For helices with exposed termini, the calculations using the cavity model yielded electrostatic interaction energies that were lower by an order of magnitude than those using the uniform dielectric model.

Several authors have attempted to calculate the contribution of electrostatic potentials in an enzyme active site. Gilson and Honig (1987) used experimental studies on the serine protease subtilisin to compare calculated electrostatic interactions between two specific sites on a protein (Klapper *et al.*, 1986). The extent of agreement between the theoretical and experimental results suggested that the continuum solvent model used in the calculation reproduces the essential feature of the interaction. Sternberg *et al.* (1987) have shown that the algorithm of Warwicker and Watson (1982), which uses classical electrostatics and considers both the charge position and the shape of the molecule, can be used to model several pK_a shifts in subtilisin. These pK_a shifts were produced by site-directed mutagenesis. Quiocho *et al.* (1987) examined three highly refined atomic structures of periplasmic binding proteins. Hy-

drogen bonds, acting primarily through backbone peptide units, are mainly responsible for the involvement of the positively charged Arg^{151} residue in the ligand site of the arabinose-binding protein, for the association between the sulfate-binding protein and the completely buried sulfate dianion, and the formation of the complex of the Leu/Ile/Val binding protein with the Leu zwitterion. They propose a general mechanism in which the isolated charges on various buried, desolvated ionic groups are stabilized by the polarized peptide units.

Van Belle *et al.* (1987) described calculations implementing molecular mechanics and molecular dynamics simulation procedures on crystal structures of crambin, liver alcohol dehydrogenase, and ribonuclease T_1. Evaluation of the contribution of polarizability of the protein matrix to electrostatic energies, local fields, dipole moments of peptide groups, and secondary structure elements in the polypeptide chain was carried out. The results confirmed that induced dipole moments in proteins provide important stabilizing contributions to electrostatic energies and that these contributions cannot be mimicked by the usual approximations where either a continuum dielectric constant or a distance-dependent dielectric function is used. It was found that an induced protein dipole appreciably affects the magnitude and direction of local electrostatic fields in a manner that is strongly influenced by the microscopic environment in the protein. The induction effects from surrounding protein atoms tend on average to increase peptide dipoles and helix macrodipoles by about 16%, again reflecting electrostatic stabilization by the protein matrix, and show that (at least in the α/β domain of alcohol dehydrogenase) the contributions of side chains to this stabilization is significant.

The role of electrostatics and calculations of electrostatic interactions have been reviewed in depth (Rogers and Sternberg, 1984; Warshel and Russell, 1984; Matthew, 1985; Matthew and Gurd, 1986a,b; Honig *et al.*, 1986; Rogers, 1986).

The overall stability of a protein structure is considered to be the sum of contributions from electrostatic interactions, hydrogen bonding, and van der Waals and hydrophobic interactions (solvent exclusion). The role of solvent accessibility and the effective p*K* values for ionizable groups has been studied for many years and plays a role in enzyme catalysis, etc. The generation and utilization of electrical potentials is a central function of biological membranes. These topics are not discussed in this chapter, and the reader is referred to Chapter 8 by N. K. Rogers.

E. Packing of α Helices and β Pleated Sheets (α/β)

Chothia *et al.* (1977) presented simple models that describe the rules for almost all the packing that occurs between and among α helices and pleated sheets. These packing rules, together with the primary and secondary structure, are the major determinants of the three-dimensional structure of proteins. The two principles that dominate the way secondary structures associate are (1) that residues that become buried in the interior of a protein close-pack and occupy a volume similar to that which they occupy in crystals of their amino acids and (2) that associated secondary structures retain a conformation close to the minimum-free-energy conformations of the isolated secondary structures. Efimov (1977) also considered the stereochemistry of α helices and the β structure in globular proteins. Side-chain packing of hydrophobic amino acids was represented in a rough approximation in the form of spheres. Principles of close packing were applied for formation of close-packed layers of hydrophobic side chains on the surface of α helices and β structure.

Janin and Chothia (1980) proposed a model for the packing of α helices on β sheets in α/β proteins. It involved the association of two smooth surfaces with complementary twists: the surface of a regular β sheet with a right-handed twist and the helix face formed by two rows

of residues, i, $i + 4$, $i + 8$, etc., and $i + 1$, $i + 5$, $i + 9$, etc. The model required the helix areas to be parallel to the β strands and contacts to be limited to the residues mentioned above. It also follows from the model that when two helices pack against each other as well, their axes should be at an angle of about $-40°$, compatible with only one class of helix–helix interaction. To substantiate their model, they examined helix–sheet and helix–helix contacts in eight proteins of known three-dimensional structure. The contact regions have a typical amino composition, especially in the β sheets, that is very rich in Val and Ile, two residues that contribute to the formation of a smooth surface. Thus, the requirement of helix–sheet packing sets severe restrictions on the configuration and on the amino acid composition of the secondary structures in α/β proteins.

Sternberg and Cohen (1982) and Sternberg *et al.* (1982) applied their combinatorial predictive approach to the docking of α helices and β strands into a tertiary fold. They point out that several problems still remain before a general predictive scheme using their combinatorial docking of secondary structure can be developed. (1) Secondary prediction is not yet accurate enough to provide the starting assignment required by the scheme (Schulz and Schirmer, 1979), and it was suggested that an improvement of secondary structure prediction might be to have some feedback from a tertiary docking algorithm. (2) Another problem is that many proteins are divided into spatially distinct regions, known as domains (Wodak and Janin, 1981), and an algorithm is required to locate the link between domains starting from the amino acid sequence. At present the combinatorial approach is a reduced list of possible structures, and methods are required to identify the correct fold and discard the alternatives.

Sternberg *et al.* (1982) studied the packing of α helices and β strands in six α/β proteins (e.g., flavodoxin) and developed a computer algorithm to predict the tertiary structure of an α/β protein from its amino acid sequence and actual assignment of secondary structure. The packing of an individual α helix against a β sheet generally involves two adjacent rows ±4 rows on nonpolar residues on the α helix. The pattern of interacting β-sheet residues results from the twisted nature of the sheet surface and the attendant rotation of side chains. In general, the α helix sits 10 Å above the sheet and lies parallel to the strand direction. The prediction follows a combinatorial approach. First, a list of possible β-sheet structures (10^6 to 10^{14}) is constructed by the generation of all β-sheet topologies and β-strand assignments. This list is reduced by constraints of topology and the location of nonpolar residues to mediate the sheet–helix packing and then rank ordered on the extent of hydrogen bonding. This algorithm was uniformly applied to 16 α/β domains in 13 proteins.

Thornton and Chakauya (1982) analyzed a sample of 70 nonhomologous proteins whose terminal residues had been classified either as α or β. It was found that the resulting distribution of the helical, β-strand, and coil conformations is significantly different for the amino and carboxy terminals. The amino terminal preferentially adopts an extended β strand, whereas the carboxyl terminal is usually helical. The observed difference derives from the α/β proteins in which the helix and strand alternate along the sequence, suggesting that the origin of this preference lies in the special structural topology of α/β proteins and the βα unit. Taylor and Thornton (1983) analyzed 62 βαβ units of known sequence and structure from 18 different proteins, which allowed them to calculate the average lengths of helix, strand, and coil regions and constructed an "ideal" βαβ unit. The observed βαβ sequences were scaled (expanded or contracted) to produce a maximum correspondence with the ideal sequence. The βαβ unit having its two strands lying adjacent in the β sheet was more precisely defined than the βαβ unit having nonadjacent strands.

The secondary structure was predicted by the method of Garnier *et al.* (1978). The predicted scheme is based on finding the best position along the protein sequence for a scaled

ideal βαβ template. The application of this basic method to the location of βαβs and the refinement of secondary structure prediction is as follows. (1) A standard secondary structure prediction method (e.g., Chou and Fasman, 1978a; Garnier *et al.*, 1978) is used to generate the helical (P_α), extended (P_β), or turn/coil (P_t) probability profiles from the sequence. (2) Additional information on the distribution of hydrophobic residues, as specifically required in the βαβ unit, is added to these probabilities (Cohen *et al.*, 1982). Typically, this contribution is the same order of magnitude as the original P_α and P_β (Schulz and Schirmer, 1979). (3) The ideal βαβ template, which can expand or contract, is fitted to the β, α, and turn profiles for each position, on the sequence and the maximum F is found (the F value, the goodness of fit, is obtained as a product of the area under the predicted secondary structure probability curves for each type of structure in the region where it corresponds to the idealized sequence). The fits having the highest F that do not cause β and α regions to overlap are taken as the most likely βαβ locations. (4) For the refinement of the secondary structure prediction, the P_α, P_β, and P_t profiles are weighed by F (in the Garnier *et al.*, 1978, method, this is equivalent to a local change in decision constant), and a new secondary structure prediction is obtained. When this approach was applied to 16 βαβ proteins of known sequence and structure, 80% of adjacent-stranded βαβ units and 50% of the nonadjacent βαβs were correctly located. An average improvement of 7.5% was observed. The overall accuracy of 67% can be analyzed for each structural component.

Cohen *et al.* (1983) attempted to suggest a route to more productive secondary structure assignments by means of a physical-model approach. The general strategy was adopted from the "expert systems" formulation of artificial intelligence, which consists of stating explicit hypotheses or rules about the system, providing a way of evaluating the validity of the rules, and developing a set of higher-order rules to resolve conflicts. The rules were simple lists of generalized amino acid sequences or patterns that are associated via the physical model with specific secondary structures. The model is a natural evolution of the suggestions of Lim (1974a–c), Nagano (1973), and others. The use of artificial intelligence technology offers novel departures from earlier efforts: (1) the algorithms are completely defined; (2) useful results can be obtained even when a unique assignment is not possible; and (3) the entire process is refinable so that new information or insights can be added and tested at any stage. The method was applied to α/β proteins. Turns were identified very accurately (98%) by simultaneously considering hydrophobicity and the ideal spacing of turns throughout the sequence. Although the assignment for any one sequence is not unique, at least one of the assignments bears a close resemblance to the native structure. The algorithm successfully divided proteins into two classes: α/β and non-α/β. The accuracy of the secondary structure assignments in the α/β class is sufficient to provide useful input for tertiary structure assignments.

Taylor and Thornton (1984) described a procedure to recognize supersecondary structure in protein sequences (see above). When applied to the identification of βαβ units in β/α type proteins, it was found to have 75% accuracy. Wierenga *et al.* (1986) derived an amino acid sequence "fingerprint" that has been used to test if a particular sequence will fold into a βαβ unit with ADP-binding properties. The fingerprint, constructed from analysis of five proteins, is a set of 11 rules describing the type of amino acid that should occur at a specific position in a peptide fragment. The total length of this fingerprint varies between 29 and 31 residues. When the data were checked against all possible sequences in a data base (PIR: Barker *et al.*, 1984), it appeared that every peptide that exactly follows this fingerprint does indeed fold into an ADP-binding βαβ unit.

Edwards *et al.* (1987) examined the structural and sequence patterns in the loops of βαβ units (See Section V.A.).

F. Prediction of Protein Conformation by Minimum-Energy Calculations

The earliest attempts to determine the structure of proteins by means of minimum-energy calculations were those of Ramachandran *et al.* (1963) and Liquori and co-workers (DeSantis *et al.*, 1965), who showed that the peptide unit can only adopt certain conformations. Nonbonding interactions were considered, and ϕ,ψ plots were constructed, which were later improved by semiempirical calculations (Lewis *et al.*, 1973b), to predict preferred regions. Scheraga and co-workers have contributed significantly to this area and have frequently reviewed the field (Scheraga, 1968, 1971, 1985; Nemethy and Scheraga, 1977). The assumption in all these calculations is that a protein folds so as to minimize the free energy of the system, and many investigators (e.g., Levitt and Warshel, 1975; Nemethy and Scheraga, 1977; McCammon *et al.*, 1977) have developed potential functions to describe the energy surface of the polypeptide chain. Chain folding is simulated computationally to find the energy minimum. Alternatively, conformational space is probed from a starting point by integrating the equations of motion over time.

There are two major problems in this approach that have yet to be effectively overcome: (1) decisions on the appropriate parameterization of the energy, including derivation of the values of the parameters, and (2) the multiple-minimum problem on the global energy surface (Ponder and Richards, 1987). Much research is presently directed toward solving these problems, and substantial progress is being made (e.g., Weiner *et al.*, 1984). However, the approximations required to model the computationally complex electrostatic and solvent effects properly are major stumbling blocks. This problem of the free energy of folding must be solved if we are to understand protein structure based on first principles. It must be noted that it has not yet been determined unambiguously whether the minimum corresponding to the native protein is the one of lowest energy. If the native conformation were not the global minimum, this would imply that certain conformations are kinetically inaccessible because of very high potential energy barriers. To date, no evidence is available that contradicts the hypothesis that the native conformation is the one of the lowest free energy. Although progress is being made, energy-minimization schemes have failed to predict chain folding accurately (Hagler and Honig, 1978; Cohen and Sternberg, 1980a,b). Because the literature covering this area is vast, no attempt is made to summarize it, but only significant trends are mentioned. An excellent in-depth review can be found in Chapter 7 by Mackay, Cross, and Hagler.

The methods of attempting to predict tertiary structure have been of two types. (1) The Monte Carlo approach of the Scheraga school, discussed in detail by Nemethy and Scheraga (1977), samples the conformational energy surface at random, and the emphasis is on the prediction of long-range residue-to-residue contact when the secondary structure is well predicted in advance or is constrained closely to the observed secondary structure. Further, it assumes that the native structure of a protein is that of least energy, whereas it might be the lowest-energy conformation that the protein can reach from an open conformation in reasonable time. (2) Folding procedures, on the other hand, take some account of the latter possibility but, in order to reduce the computation time to manageable proportions, use a very simplified representation of the molecular structure of the polypeptide chain (Levitt and Warshel, 1975; Levitt, 1976; Kuntz *et al.*, 1976). This simplification and other assumptions made in the latter methods have left them open to the criticism that they may have led to fortuitous approximate agreement with the observed structure (Robson, 1974; Nemethy and Scheraga, 1977; Hagler and Honig, 1978). Further, they simplify the backbone structure to such a degree that they cannot be used to provide detailed information about changes in secondary structure during folding (Robson and Osguthorpe, 1979).

Scheraga and co-workers (Ponnuswamy *et al.*, 1973) considered the role of medium-

range interactions in proteins. The energies of oligopeptide segments of lysozyme were minimized with respect to the dihedral angles of the central residue. As the length of the oligopeptide segment increased up to a nonapeptide, the low-energy conformation became that observed in the x-ray structure in most cases. This finding suggests that although short-range interactions appear to play the dominant role in determining the conformation of an amino acid residue in a protein, the additional interactions required to stabilize the conformation uniquely may be only of medium range, i.e., those within a nonapeptide, and longer-range interactions may be of considerably less importance.

Levitt and Warshel (1975; Levitt, 1976) developed a simple representation of protein conformations that was used together with energy minimization and normal-mode thermalization to simulate protein folding. Under certain conditions, the method succeeded in "renaturing" bovine pancreatic trypsin inhibitor from an open-chain conformation into a folded conformation close to that of the native molecule. Levitt (1976) described the representation in some detail and tested the methods extensively under a variety of different conditions. The use of the concept of time-averaged forces, introduced by Levitt and Warshel (1975), was shown to simplify conformational energy calculations on globular proteins. A detailed description is given of the simplified molecular geometry, the parameterization of suitable force fields, the best energy-minimization procedure, and the techniques for escaping from local minima. Extensive tests of the method on the native conformation of pancreatic trypsin inhibitor showed that this simplification worked well in representing the stable native conformation of this globular protein. The original early optimism for the application of minimization of the total potential energy of a polypeptide to lead to the observed equilibrium structure has not overcome the two major problems of (1) choice of appropriate parameterization of the energy and (2) the multiple-minimum problem on the global potential energy surface (Ponder and Richards, 1987).

Kuntz *et al.* (1976) described a method for calculating the tertiary structure of proteins given their amino acid sequence. The algorithm involves locally minimizing an energylike expression as a function of the Cartesian coordinates of the C_β of all residues. Although the approximation to the true polypeptide geometry and conformational energies is extremely approximate, quite respectable results were obtained for the small proteins subredoxin and trypsin inhibitor, where the root-mean-square errors were as low as 4.0 Å and 4.7 Å, respectively. This paper is illustrative of the then-developing interest in extremely simplified models to predict the structure of small globular proteins (e.g., Levitt and Warshel, 1975; Ptitsyn and Rashin, 1975; Burgess and Scheraga, 1975).

Robson and Osguthorpe (1979) proposed a new model and parameters for the computer simulation of protein folding that satisfied requirements for a fully automatic simulation of protein folding. The major improvement over previous methods was to retain a more realistic and complete representation of the protein backbone and to reduce the number of variables by coupling their behavior. When applied to the folding of pancreatic trypsin inhibitor, a root-mean-square fit of 6 Å with good secondary structure was obtained. This method allowed a more detailed examination of secondary structure transitions during protein folding. By observing changes in secondary structure during the simulated folding, the authors showed that α helices and extended chain regions predicted at the outset or formed early in the simulation are conserved and that certain residues are crucial as flexible hinge points to bring the secondary structure together in order to achieve tertiary packing. Hinge points were found in general not to contain Gly. This made an important distinction between a "reverse-turn region," for which Gly is statistically a strong candidate, and a hinge point in the protein backbone.

Finney *et al.* (1980) used detailed hydrogen-bonding, surface exposure, internal environment, and solvent interaction calculations on several proteins in conjunction with data from quantum mechanical hydrogen-bonding studies to estimate various contributions to the free energy of folding and discussed their likely significance. A picture emerged of globular proteins as extremely well-fitted jigsaw puzzles in which no single driving force dominated the marginal stability of the native conformation. Rather, the folded structure is seen as the result of a complex global maximization of several strongly interacting driving forces. In particular, the need to maintain very efficient internal hydrogen bonding and the role of the solvent as a hydrogen-bond sink were stressed as strong constraints on the (incomplete) maximization of hydrophobic effects. The possible significance of internal-dipole–induced-dipole interaction was tentatively discussed.

Pincus and Klausner (1982) predicted the three-dimensional structure of the signal sequence for murine κ light chain using conformational energy calculations. The strategies employed to calculate the conformation involved judicious combinations of the local minima for simple peptides to construct longer peptide chains. A simple method was presented for combining the local minima for the component di- and tripeptides of the leader sequence that allows construction of long peptide chains. These calculations, based on tested and reliable potential energy functions, employ a novel global search technique to identify the lowest-energy structures for a hexadecapeptide signal sequence, Glu-Thr-Asp-Thr-(Leu$_3$-Trp-Val)$_2$-Pro-Gly. It has been found that the core hydrophobic sequence, Leu$_3$-Trp-Val-Leu, adopts an α-helical conformation that is terminated by chain-reversal conformations for the four residues Trp-Val-Pro-Gly. The amino-terminal four residues adopt a low-energy conformation that is fully compatible with the succeeding α helix. The immediate neighboring sequence, Asp-Thr, exists in a single lowest-energy, double-equatorial conformation, whereas the first two residues, Glu-Thr, can adopt a variety of low-energy conformations. The calculations arrive at a highly structured and specific model for the conformation of the leader sequence that is compatible with experimental data.

Levitt (1983) calculated the nativelike folded conformations of bovine pancreatic trypsin inhibitor protein by searching for conformations with the lowest possible potential energy. Twenty-five random starting structures were subjected to soft-atom-restrained energy minimization with respect to both torsion angles and the atomic Cartesian coordinates. The restraints used to limit the search included the three disulfide bridges and the 16 main-chain hydrogen bonds that define the secondary structure. Novel features of the methods used included soft atoms to make restrained energy minimization work, writing numbers to classify chain threadings, and molecular dynamics followed by energy minimization to anneal the conformations and reduce the energies further. The best conformation calculated had a root-mean-square deviation of only 3 Å and showed the same special threading found in the x-ray structure.

Weiner *et al.* (1984) presented the development of a force field for simulation of nucleic acids and proteins. The approach began by obtaining equilibrium bond lengths and angles from microwave, neutron diffraction, and prior molecular mechanical calculations, torsional constants from microwave, NMR, and molecular mechanical studies, nonbonded parameters from crystal-packing calculations, and atomic charges from the fit of a partial-charge model to electrostatic potentials calculated by *ab initio* quantum mechanical theory. For proteins they focused on ϕ,ψ maps of Gly and Ala dipeptides, hydrogen-bonding interactions involving various protein polar groups, and energy refinement calculations on insulin. The authors discuss the improvements over previous attempts and have delineated areas for further improvement.

Robson and Platt (1986) made a reappraisal of the interatomic potential functions for protein structure calculations using the all-atom approximation (except CH, CH_2, and CH_3, which are treated as "united atoms"). The potential functions are somewhat novel in form and consistent with more efficient and robust folding algorithms. In addition, the potentials are calibrated for the rigid geometry approximation, since use of fixed standard bond lengths and valence angles (and fixed transplanar peptide groups) reduced the number of conformational variables and saves a great deal of computer time. Although these algorithms demand the use of potential functions of this special type, these functions can be readily implemented in more classical programs for the conformational analysis of proteins. They were calibrated or tested against a large body of experimental data, including (1) extended-basis-set *ab initio* quantum mechanical calculations, (2) nuclear magnetic resonance spectroscopic data and dipole moment data for di- and oligopeptides, (3) characteristic ratio data for random-coil homopolypeptides, (4) extensive data from peptide solubility studies, and (5) experimental structures of polyalanine fibers and globular proteins.

Brucceroleri and Karplus (1987) developed a procedure, CONGEN, for uniformly sampling the conformational space of short polypeptide segments in proteins. This method of homology modeling greatly simplifies the problem of tertiary structure prediction. Because the time required for this sampling grows exponentially with the number of residues, parameters were introduced to limit the conformational space that has to be explored. This was done by using the empirical energy function of CHARMM (Brooks *et al.*, 1983) and truncating the search when conformations of grossly unfavorable energies were sampled. Tests were made to determine control parameters that optimize the search without excluding important configurations. When applied to known protein structures, the resulting procedure is generally capable of generating conformations where the lowest-energy conformation matches the known structure with a root-mean-square deviation of 1 Å.

Bash *et al.* (1987) tackled the fundamental problem in chemistry and biochemistry of understanding the role of solvation in determining molecular properties. Recent advances in statistical mechanical theory and molecular dynamics can be used to solve this problem with the aid of supercomputers. These advances permit the free energies of solvation of all chemical classes of amino acid side chains, four nucleic acid bases, and other organic molecules to be calculated. The effect of a site-specific mutation on the stability of trypsin was predicted. The results are in good agreement with available experiments.

Although progress is being made, the formidable task of calculating the tertiary structure of proteins requires a great effort in future research.

G. Expert Systems

Robson and co-workers (Robson *et al.*, 1987; Fishleigh *et al.*, 1987) have studied rationales for an expert-system approach to the interpretation of protein sequence data. Fishleigh *et al.* (1987) analyzed the sequence of the epidermal growth factor receptor to identify regions of potential structural and functional importance. This receptor protein was analyzed using a variety of established methods and novel procedures developed for the study of weak internal and external homologies and for the use of homologous sequences in the prediction of secondary and supersecondary structures. First, a composition scan of sequences was done to locate consensus sets for posttranslational modification, identifying features such as transmembrane sections and analyzing regions of unusual amino acid composition. This was followed by searches for internal and external homologues and the prediction of the secondary structure of the protein of interest, with any homologous sequences identified. Information gained from the composition scan and the study of homologues can then be used to refine the

secondary structure prediction, which, in conjunction with any homology data, may permit the development of a model for the gross structures of the protein.

An additional technique enables the division of the protein sequence into sections that may represent current or ancestral exons, collectively referred to as paraexonic fragments. The Cys-rich regions were analyzed in terms of their separations along the sequence. The methods of Garnier *et al.* (1978) and Lim (1974a–c) were used for secondary structure prediction.

Robson *et al.* (1987) developed a suite of programs, named LUCIFER (Logical Use of Conformational Information and Fast Energy Routines) for the conformational study of drugs, proteins, and other biomolecules. The suite explicitly seeks to carry out global energy minimization as rapidly as possible (Robson and Platt, 1986) but yet puts emphasis on the rational use of external data to overcome any deficiencies. The external data are of three basic types. The first come from the scanning of data bases to identify primary, secondary, supersecondary, and tertiary structural homologies. Second, experimental data from physical chemists and pharmacologists are analyzed and exploited. Such data are exemplified by intergroup distances from NMR spectroscopy, infrared spectral data, characteristic ratios of polymers, net dipole moments, and pharmacological potencies of structural analogues. Third, the Imman graphics system developed by their group provides a vehicle for more qualitative judgments.

The value of this system has been illustrated in the modeling of immunoglobulins against electron microscopic and other experimental data (Pumphrey, 1986a,b; Robson and Garnier, 1986). This method was described for chloramphenicol acetyltransferase (CAT). The first step, homology detection, utilized the method of Fishleigh *et al.* (1987), that of Garnier *et al.* (1978), and a computerized version of the Lim method (Protein Resource Identification, National Biomedical Research Foundation, Georgetown University Medical Center, Washington, DC, 1986) for the secondary structure prediction. New procedures were also described for identification of secondary structure template similarity on the basis of theorem-proof algebra and for rapid evaluation of the hydrophobic packing of a protein conformation. These methods are discussed in relation to the molecular-graphics modeling of CAT, where it is assumed that there was a tertiary structure relationship between CAT and cat muscle pyruvate kinases. An attractive feature of the structure was that its hydrophobic radius of gyration was very close to the value that would have been expected for a protein of this size. The hydrophobic radius of gyration of a protein is a criterion by which nonnativelike conformations may be rapidly identified. Thirty-four proteins in the range of 40 to 500 residues were analyzed to obtain the hydrophobic radii of gyration for use in comparative studies. Although the authors state that LUCIFER is of proven worth for the study of biological peptides and for protein modeling against homologues, its applicability to *de novo* protein structure prediction is untested. A preliminary study of an avian pancreatic polypeptide is described and classified as both informational and promising.

In an excellent review on "Folding and Association of Proteins," Jaenicke (1987) gives a somber evaluation of the prospects for a more accurate prediction of the tertiary structure of proteins. There is still much work to be done.

VII. PREDICTION OF MEMBRANE STRUCTURE: METHODS OF PREDICTION

Membrane proteins are a diverse group of proteins that play an important role in modulating the activity of many of the cell's functions. With the advent of cloning methodologies and the ease of DNA sequencing, the availability of membrane proteins, previously obtainable in minute amounts, has made their study feasible. Of great interest is their mode of function. The diversity of proteins residing in the lipid bilayer of the membrane is immense, as is their

function from receptors to active enzymes. Those presently under study include bacteriorhodopsin, acetylcholine receptor, ATPases, T-cell receptor, and the sodium channel.

The method of prediction for membrane proteins that is utilized most frequently is that of Kyte and Doolittle (1982). This simple method for displaying the hydropathic character of a protein has been used to define the sequences that have a high probability of being embedded in the nonpolar membrane environment. This agrees with the first report by Capaldi and Vanderkooi (1972) of the low polarity of many membrane proteins. With membrane-bound proteins, the portions of their sequences that are located within the lipid bilayer are clearly delineated by large uninterrupted areas on the hydrophilic side of the midpoint line in the well-known plot of the hydropathic index versus sequence number. Kyte and Doolittle (1982) devised their own hydropathy scale, and a computer program, SOAP, assigns the appropriate hydropathic value to each residue in a given amino acid sequence and then successively sums those values, starting at the amino terminal, within overlapping segments displayed from each other by one residue. Although a segment of any size can be chosen, ordinarily spans of 7, 9, 11, or 13 were employed, odd numbers being used so that a given sum could be plotted above the middle residue of the segment. One of the novel features of the approach is that membrane-spanning segments can be identified and distinguished from sequences that merely pass through the interior of a protein. Kyte and Doolittle (1982) have carefully avoided associating the hydrophobic segments with any particular secondary structure, i.e., α helix or β strand. However, unfortunately, in the literature, it has been generally accepted and stated that these segments represent α-helical segments. It is possible, of course, that they do represent helical sections, and they probably are helical, but the Kyte and Doolittle method cannot be used to verify the presence of a helical segment in a bilayer. The β sheets are also extremely hydrophobic. On examination of P_α and P_β values (Chou and Fasman, 1974a) together with the hydropathy scale of Kyte and Doolittle (1982), it can be seen that β sheets on average have a higher average hydrophobicity that do α helices. As was recently shown, porin, which spans the outer membranes of *Escherichia coli* and forms voltage-dependent transmembrane channels, has a conformation dominated by β structure (Paul and Rosenbusch, 1985).

Paul and Rosenbusch (1985) have pointed out that conventional methods of secondary structure prediction for bacteriorhodopsin have been ambiguous; however, when they applied the Chou and Fasman (1978a,b, 1979a,b) method of utilizing β turns as well, they could rationalize the hydrophobic segments as being α helices immersed in the bilayer. This subject is discussed in detail later in this chapter.

Argos *et al.* (1982) developed an algorithm based on physical characteristics of the 20 amino acids and refined by comparison to the proposed bacteriorhodopsin structure to delineate likely membrane-buried regions in the primary sequences of proteins known to interact with the lipid bilayer. Application of the method to the sequence of the carboxyl-terminal one third of bovine rhodopsin predicted a membrane-buried helical hairpin structure. With the use of lipid-buried segments in bacteriorhodopsin as well as regions predicted by the algorithm in other membrane-bound proteins, a hierarchic ranking of the 20 amino acids in order of their preferences to be in lipid contact was calculated. A helical wheel analysis of the predicted regions suggested which helical faces are within the protein interior and which are in contact with the lipid bilayer. Nine arbitrarily chosen physical parameters that may have some relevance to conformational preference were used. However, a more serious criticism of the method is that it is based on a highly speculative structure of bacteriorhodopsin and is therefore suspect. Several other membrane proteins—glycophorin A, cytochrome b_5, cytochrome *c*, ATP synthase, M13 coat protein, precasein, proteolipid, and porin—were also predicted by this approach by Argos *et al.* (1982).

Mohana-Rao *et al.* (1983), using the algorithm of Argos *et al.* (1982), predicted the hydrophobic helical spans in the thylakoid membrane protein (TMP). A second method was used to indicate the number of possible helices in TMP as well as the N terminus of each helix. Curve segments from the helical spans of the vertebrate photoreceptor protein rhodopsin (RHO), a five-parameter plot, assumed to be another seven-helical protein like bacteriorhodopsin, were compared to all segments of the comparable TMP plot, and correlation coefficients were calculated at each value. The helical wheels were calculated using a table of membrane-buried helical preference values. Thus, it was concluded that the three proteins, thylakoid protein, bovine rhodopsin, and bacteriorhodopsin, are all seven-helical bundles, and it was hypothesized that this uniquely stable arrangement was evolved by convergent evolution and will be found frequently in membrane proteins. The consecutive assumptions leading to this conclusion cast doubts on this conclusion. Flinta *et al.* (1983) challenged the often-claimed proposition that transmembrane helices show "sidedness" in the distribution of polar and hydrophobic residues. However, an analysis of the statistical distribution of polar residues in randomly generated helices shows that the degree of bias commonly observed in real helices is far from statistically significant. Thus, it is concluded that a "patchy" distribution of residues in such helices should be interpreted with great care.

Eisenberg *et al.* (1984a,b; Eisenberg, 1984) published an algorithm that identifies α helices involved in the interactions of membrane proteins with lipid bilayers and distinguishes them from helices in soluble proteins. The membrane-associated helices are then classified with the aid of the hydrophobic moment plot, on which the hydrophobic moment of each helix is plotted as a function of its hydrophobicity. The magnitude of the hydrophobic moment measures the amphilicity of the helix (and hence its tendency to seek a surface between hydrophobic and hydrophilic phases), and the hydrophobicity measures its affinity for the membrane interior. Segments of membrane proteins in α helices tend to fall in one of three regions of a hydrophobic moment plot: (1) monomeric transmembrane anchors (class I HCA transmembrane sequences) lie in a region of highest hydrophobicity and smallest hydrophobic moment; (2) helices presumed to be paired (such as the transmembrane M segments of surface immunoglobulins) and helices that are bundled together in membranes (such as bacteriorhodopsin) fall in the adjacent region with higher hydrophobic moment and smaller hydrophobicity; and (3) helices from surface-seeking proteins (such as melittin) fall in the region with still higher hydrophobic moment. The α helices from globular proteins mainly fall in a region of lower mean hydrophobicity and hydrophobic moment. This procedure demonstrated that the sequences of diphtheria toxin may have four transmembrane helices and a surface-seeking helix in fragment B, the moiety known to have a transmembrane function.

The method of Finer-Moore and Stroud (1984), based on a Fourier analysis of hydrophobicities, revealed that the subunit sequences of the acetylcholine receptor have sequences of amphipathic secondary structures. Prediction of a consensus secondary structure based on this analysis and on an empirical prediction method leads to a testable hypothesis about how the ion channel is formed and might function. The acetylcholine receptor (AcChR), is a ≃250-kDa complex of five homologous glycoprotein subunits in stoichiometry $\alpha_2\beta\delta\gamma$. The local periodicity in the hydrophobicities of the primary sequences of all subunits was quantitated by Fourier analysis in the manner of McLachlan and Karn (1983). Fourier analysis was used for amphipathic secondary structure correlation together with a secondary structure prediction method for the ≃75% sequence external to the bilayer and separate evidence for oriented helices in the transmembrane regions to develop an overall structural scheme. Amphipathic spectra $I_k(\nu)$ (power spectra of hydrophobicities) were computed for stretches of 25 residues in length according to

$$I_k(\nu) = \Sigma_j = {}^{k+12}_{k-12} (h_j - h_k) \exp(2\pi \cdot j \cdot \nu)^2$$

where h_j is the hydrophobicity of residue j (kcal/mole) from a consensus set of values developed by Eisenberg *et al.* (1982a,b), and h_k is the average hydrophobicity of the 25 residues from $k - 12$ to $k + 12$. Hydrophobicity plots (Kyte and Doolittle, 1982) H_i were generated by

$$H_i = (1/7)\, \Sigma_j = {}^{i+3}_{i-3} h_j$$

Secondary structure prediction for extramembrane regions used the Garnier *et al.* (1978) algorithm. Sequence alignment of the AcChR subunits was carried out by hydrophobicity correlation by the comparison matrix method of McLachlan (1971). The amphipathic power spectrum $I_k(\nu)$ for each sequence was plotted as a two-dimensional contoured map with frequency ν (residues^{-1}) as abscissa and residue number k as ordinate. A prominent feature is the intense peak at $\nu = 1/3.5$ residues between $k = 412$ and 470. This periodicity is exactly that expected for an α helix, and the length and intensity of the peak strongly suggest an amphipathic α helix up to at least 30, possibly up to 58, amino acids in length. The overall secondary structure has 27% β sheet and 44% α helix, in agreement with circular dichroism studies (Moore *et al.*, 1974) and Raman spectroscopy (Chang *et al.*, 1983). A detailed discussion of the Finer-Moore and Stroud predictive method can be found in Chapter 19 by Finer-Moore, Bazan, Rubin, and Stroud.

Gray and Matthews (1984) surveyed the known protein structures and found that approximately 70% of serine residues and at least 85% (potentially 100%) of threonine residues in helices make hydrogen bonds to carboxyl oxygen atoms in the preceding turn of the helix. The high frequency of intrahelical hydrogen bonding is of particular significance for intrinsic membrane-bound proteins that form transmembrane helices. Hydrogen bonding within a helix provides a way for serine, threonine, and cysteine to satisfy their hydrogen-bonding potential, permitting such residues to occur in helices buried within a hydrophobic milieu. Honig and Hubbell (1984) estimated the free energies of transfer of ionized amino acid side chains in water to both their ion-paired and neutral hydrogen-bonded states in low-dielectric media. The difference between the two free energies corresponds to the proton transfer free energy in a "salt bridge" formed between acidic and basic groups (i.e., Lys and Glu residues). Dielectrics of 80 (H_2O) and 1 (vacuum) were used in the calculation. Their results suggest that it costs approximately 10–16 kcal/mole to transfer a salt bridge from water to a medium of $\epsilon = 2$–4 in ionized or neutral form. The proton transfer energy is thus approximately zero. The tendency of salt bridges to form additional hydrogen bonds in real proteins suggests that the ion pair will be present in most biological systems. Kuhn and Leigh (1985) developed a statistical technique for predicting transmembrane segments of membrane proteins from their amino acid sequences. A propensity scale was derived from the frequency of occurrence of amino acids in transmembrane fragments. Those values were compared to the hydropathy scale of Kyte and Doolittle (1982) and the signal sequence helical potential scales of Argos *et al.* (1982). The difficulty of such an approach is that the sequences termed transmembrane are rough estimates, as these are not definitely known.

Klein *et al.* (1986) used discriminant analysis to classify membrane proteins precisely as integral or peripheral and to estimate the odds that the classification is correct. On 102 membrane proteins from the National Biomedical Research Foundation, it was found that the discrimination between integral and peripheral membrane proteins can be achieved with 99% reliability. Hydrophobic segments of integral membranes can also be distinguished from interior segments of globular soluble proteins with better that 95% reliability. A procedure was

also proposed for determining boundaries of membrane-spanning segments, and it was applied to several integral membrane proteins. From the limited data available, the residues at the boundaries of a membrane-spanning segment are predictable to within the error in the concept of a boundary. As a specific indication of resolution, seven membrane-spanning segments of bacteriorhodopsin were resolved with no information other than sequence, and the predicted boundary residues agree with the experimental data on proteolytic cleavage sites. A computer program in FORTRAN for prediction of membrane-spanning segments is available from the sources listed in the appendices at the end of the text.

Argos and Mohana-Rao (1985) applied their predictive scheme (Argos *et al.*, 1982; Mohana-Rao *et al.*, 1983) to five functionally distinct lipid-bound proteins whose exonic structure is known in an attempt to shed light on the mechanism and etiology of splice junctions. They found that the splice junctions largely map to the predicted surface segments and that the number of junctions correlates with the length of the surface spans in the five proteins.

Jähnig and co-workers (Vogel *et al.*, 1985; Vogel and Jähnig, 1986) employed a structural predictive method that takes into account the amphipathic helices. The method of Kyte and Doolittle (1982), based only on hydrophobicity, underestimates the number of membrane-spanning helices. To predict amphipathic α helices, the authors represent the amphipathy as real space instead of Fourier space as in the analysis of the hydrophobicity, $H_\alpha(i)$ (Eisenberg *et al.*, 1984a,b; Finer-Moore and Stroud, 1984). Turns are also predicted by the Chou and Fasman (1978a) method. A structural model of lactose permease is proposed in which the ten membrane-spanning helices are expected to form an outer ring of helices in the membrane. The interior of the ring would be made up of residues that are predominantly hydrophilic and, by analogy to sugar-binding proteins, suited to provide the sugar-binding site. Vogel and Jähnig (1986) predicted the outer-membrane proteins of *Escherichia coli*, porin, maltoporin, and OmpA protein. Using an adaption of their method for structural prediction of amphipathic α helices, they predicted the amphipathic β strands in the membrane. In the model suggested, the OmpA fragment consists of eight amphipathic membrane-spanning β strands that form a β barrel. This agrees with the structure determined by Raman spectroscopy, which estimated 50 to 60% β strands, about 20% β turn, and less than 15% α helix. Similarly porin is folded into ten amphipathic membrane-spanning β strands that are located at the surface of the trimer towards the lipids and eight predominantly hydrophobic strands in the interior.

Edmonds (1985) has calculated the interaction energies between hydrophobic α-helical sections that span membranes, which are known to possess large electric dipole moments. These interaction energies, which include screening effects, remain comparable with a typical thermal energy of kT up to a separation of 20 Å. In addition, it is shown that, solely because of its dipole moment, an α helix that completely spans the membrane has an energy up to $5kT$ lower than one that terminates within the membrane width. The paper also describes the electrical interaction of the charge structure of a membrane channel and the protein helices that surround the pore. The gating charge transfer that is measured when a voltage-sensitive ion channel switches means that the dipole moment of the ion channel changes. This in turn results in a change in the radial forces that act between the pore and helices that surround it. A change in these radial forces, which tend to open or close the pore, constitutes an electrically silent gating mechanism that must necessarily act subsequent to the gating charge transfer. The gating mechanism could consist of a radial translation of the neighboring proteins or of their axial rotation under the influence of the torque that would act on a pair of approximately equidistant but oppositely directed α helices. An attempt to calculate the interaction energy of a typical pore and a single α helix spanning a membrane results in an energy of many times kT.

Engelman *et al.* (1986) reviewed the literature on identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins. They consider the arguments that support

the notion that helical structure will be a dominant motif in integral membrane protein organization, introduce the problem of suitable scaling of amino acids in terms of their polar and nonpolar characteristics, and discuss further the use of such scales in prediction of protein structure. They conclude that a suitable scale and protocol can lead to the successful identification of transmembrane helical structures in integral membrane proteins. The authors discuss the hydrophobicity scale they have developed over the past several years (Engelman *et al.*, 1981; Engelman and Steitz, 1981, 1984; Steitz *et al.*, 1982) in which the nonpolar properties of the amino acids, as they exist in a helix, were calculated using a semitheoretical approach that combines separate experimental values for the polar and nonpolar characteristics of groups in the amino acid chain. The development is similar to von Heijne's earlier work (von Heijne, 1981a,b), but with important differences. Their table of transfer energies for amino acid side chains in α-helical polypeptides includes hydrophobic and hydrophilic components of the transfer of amino acid side chains from water to a nonaqueous environment of dielectric 2 (GSE scale, Goldman, Engelman, Steitz). The GSE scale is used as a point of reference to compare with other scales (e.g., von Heijne, 1981b; von Heijne and Blomberg, 1979; Kyte and Doolittle, 1982; Rose *et al.*, 1985; Guy, 1985).

Using the GSE scale in sequence analysis, they tested their prediction of the photosynthetic reaction center of *Rhodopseudomonas viridis,* whose high-resolution structure is known (Deisenhofer *et al.*, 1984, 1985; Chapter 2 herein). The macromolecular assembly consists of four polypeptide chains: two of these (L and M) are globular integral membrane proteins, and one (H) is an anchored membrane protein. The crystal structure shows a region in which bundles of helices transverse an apparently nonpolar region. All putative membrane-spanning helices observed in the crystal are predicted from the hydrophobicity analyses of the sequences. Four helices each were suggested in both the L and M chains, and a fifth helix is possible, but of marginal significance and not corresponding to a transmembrane helix in the structure. The H-subunit profile suggests a single transmembrane helix. Eleven helices were assigned and are somewhat shorter that those actually observed (e.g., subunit M, helix A, predicted 52–71, observed 52–78). This method fails to predict the β-sheet structure found in porin (Kleffel *et al.*, 1985; Paul and Rosenbusch, 1985).

Engelman and Steitz (1981) proposed that the initial event for either secretion of protein across or insertion into membranes is the spontaneous penetration of the hydrophobic portion of the bilayer by a helical hairpin. The major proposals of this model are the following. (1) Energetic considerations of polypeptide structures in a nonpolar lipid environment as compared with an aqueous environment have led to the conclusion that only α and 3_{10} helices will be observed in the hydrophobic interiors of membranes. (2) During protein synthesis, the nascent polypeptide chain folds in the aqueous environment to form an antiparallel pair of helices, each of which is ~20 residues long. (3) The helical hairpin partitions into the membrane if the free energy arising from burying hydrophobic helical surfaces exceeds the free energy cost of burying potentially charged and hydrogen-bonding side groups. (4) Globular membrane proteins will be formed by the insertion of several pairs of helical hairpins, which are expected to be the fundamental unit of membrane protein folding. (5) In secreted proteins, the hydrophobic leader peptide forms one of these two helices and functions to pull polar portions of the secreted protein into the membrane as the second helix of the hairpin. (6) Insertion of the helical hairpin into the bilayer initiates secretion if the second helix is polar, and secretion of the newly synthesized protein continues until or unless a hydrophobic segment is encountered. (7) Alternatively, if both helices are hydrophobic, the hairpin will simply remain inserted in the membrane.

Steitz *et al.* (1982) wrote a computer program to analyze the amino acid sequence of secreted and membrane proteins in order to estimate quantitatively whether insertion of mem-

brane into lipid bilayers can be expected to be spontaneous on thermodynamic grounds and also to establish the probable topology of membrane proteins. Amino acid sequences were analyzed for the free energy of burying an α helix of definable length, usually 21 amino acid residues long. A 21-residue probe helix was removed down the sequence, and the free energy of burying each successive 21-residue helix was calculated and plotted. Bacteriorhodopsin and glycophorin were calculated in this manner. Approximately seven helices were found for bacteriorhodopsin.

Wallace *et al.* (1986) evaluated the accuracy of several predictive schemes in predicting the secondary structure of 15 integral membrane proteins and membrane-spanning polypeptides. Statistical analyses (χ^2) indicated a less than 0.5% correlation between the net predicted secondary structures and experimental results. The authors conclude that predictive schemes using soluble protein bases are inappropriate for the prediction of membrane protein folding. The methods evaluated were the Chou and Fasman (1974a,b), Garnier *et al.* (1978), and Burgess *et al.* (1974) algorithms, and a smaller number of proteins were analyzed by the Kyte and Doolittle (1982) and Engelman–Goldman–Steitz (1986) methods. This conclusion can be serious questioned, as only two of the 15 proteins evaluated have had their structure determined by x-ray crystallography. The remaining proteins' structures were determined by a combination of electron microscopy, Raman spectroscopy, infrared spectroscopy, and circular dichroism, methods whose accuracy has been frequently questioned.

Mohana-Rao and Argos (1986) developed a conformational preference parameter to predict helices in integral membrane proteins. Five parameters were found to be most suitable for predicting the presence or absence of hydrophobic stretches: hydration potential (Wolfenden *et al.*, 1979), the free energy of transfer for a given residue in a helix in aqueous medium to a helix in a nonpolar phase (von Heijne, 1981a), polarity (Trewhella *et al.*, 1983), bulk (Trewhella *et al.*, 1983), and turn conformational preferences (Levitt, 1978). The method was described in detail in an earlier paper (Argos *et al.*, 1982). The proteins chosen were those having at least two transmembrane helices. A total of 256 membrane-buried helices were predicted in 49 integral membrane proteins. The Chou–Fasman membrane-buried helix preference parameter for a particular amino acid is defined as the ratio of its composition in predicted helices to its composition in all sequence regions of the integral membrane proteins. When applied to the L and M subunits of *Rhodopseudomonas sphaeroides,* five helices were predicted, in agreement with the three-dimensional x-ray crystal structure. Data on signal sequences and amino acid exchanges in membrane proteins were also analyzed.

Cornette *et al.* (1987) developed a computational technique for detecting amphipathic structures in proteins. They optimized hydrophobicity scales and used these with a method based on a least-squares fit of a harmonic sequence to a sequence of hydrophobicity values. They termed this the "least-squares power spectrum." The sum of the spectra of the α helices in their data base peaked at 97.5°, and approximately 50% of the helices can account for this peak. Amphipathic α helices in their natural state have a periodic variation in the hydrophobicity values of the residues along the segment, with a 3.6-residue-per-cycle period characteristic of an α helix. Thus, approximately 50% of the α helices appear to be amphipathic, and of those that are, the dominant frequency at 97.5° rather than 100° indicates that the helix is slightly more open than previously thought, with the number of residues per turn closer to 3.7 than 3.6. Although the scale is optimal only for predicting α amphipathicity, it also ranks high in identifying β amphipathicity and distinguishing interior from exterior residues in a protein (see further discussion on amphiphilic helices in Section VI.D).

Furois-Corbin and Pullman (1987) undertook theoretical studies on the packing properties of α helices and on their ability to form conducting bundles using minimization techniques. Packages of two and five α helices containing leucines on their faces of contact and made

otherwise of alanine were studied. Such bundles were compared to pure poly(L-alanine) packages. They concluded that the essential packing properties were conserved, with near antiparallelism and a preponderance of nonbonded interactions. Helical packing is different when leucine is included, and substitution of serines for the alanine lying on the inner wall has little effect on interhelix packing.

VIII. PREDICTED MEMBRANE STRUCTURES

A. Bacteriorhodopsin

Bacteriorhodopsin is the transmembrane protein found in *Halobacterium halobium*. The purple membrane, which contains bacteriorhodopsin, is present in a number of extremely halophilic bacteria and catalyzes the light-driven proton translocation from the inside to the outside of the cell membrane. This generates an electrochemical gradient, which is used by the cell for the synthesis of ATP. Bacteriorhodopsin contains one retinal molecule per protein molecule, and this is linked as a Schiff base to the ϵ-amino group of Lys^{216}. This protein is probably the most studied membrane protein to date. Henderson and Unwin (1975) were able to deduce an electron density map at 7-Å resolution by an electron diffraction technique. From this is was concluded that bacteriorhodopsin forms a continuum of seven α helices, each of which spans the membrane and is largely embedded in it. Ovchinnikov *et al.* (1979) proposed the first model of bacteriorhodopsin based on studies of limited proteolysis. There were seven transmembrane helices accounting for 207 residues, which yielded a total of 83% α helix. This model was slightly altered when the antigenic structure and topography of bacteriorhodopsin were probed (Ovchinnikov *et al.*, 1985).

Engelman *et al.* (1980) attempted to fit the amino acid sequence of bacteriorhodopsin to the three-dimensional map of the molecule. Seven segments of the sequence were selected as being probable transmembrane α helices. The complete amino acid sequence had been previously determined (Ovchinnikov *et al.*, 1979; Gerber *et al.*, 1979; Walker *et al.*, 1979). There were 5040 ways of fitting these seven segments into the seven regions of helical density in the map, and these were evaluated on the basis of the criteria of connectivity of the nonhelical link regions, charge neutralization, and total scattering density per helix. The seven helices, with between 24 and 28 residues each, contained a total of 178 amino acids, which gave a helical content of 72%. Steitz *et al.* (1982) further confirmed the analysis of Engelman *et al.* (1980). The α helices appear to have their hydrophobic sides facing the nonpolar lipid moiety, whereas more hydrophilic portions are in the interior of the molecule (Engelman and Zaccai, 1980). Ovchinnikov (1982) combined the sites of enzyme proteolysis with the hydrophobicity profile (Rose and Roy, 1980) and the β-turn prediction (Chou and Fasman, 1979a) to revise the probable disposition of the polypeptide in the bilayer.

Huang *et al.* (1982), using a photosensitive *m*-diazirinophenyl analogue of retinal bound to bacterioopsin at Lys^{216}, regenerated a chromphore with λ_{max} at 470 nm. Photolysis of the complex at 365 nm resulted in covalent cross linking of the retinal analogue to the bacterioopsin in a >30% yield. Investigation of the sites of cross linking between the ^3H-labeled retinal analogue and the protein showed the peptide fragment (amino acid residues 190–248) to be the main radioactively labeled product. Stepwise Edman degradation showed Ser^{193} and Glu^{194} to be the predominant sites of cross linking. These results show that the chromophore in bacteriorhodopsin is inclined towards helix 6 and towards the exterior of the cell. These data provided information on the approximate angle that the chromophore makes with the plane of the membrane and required a modification of the then-current secondary structural model for

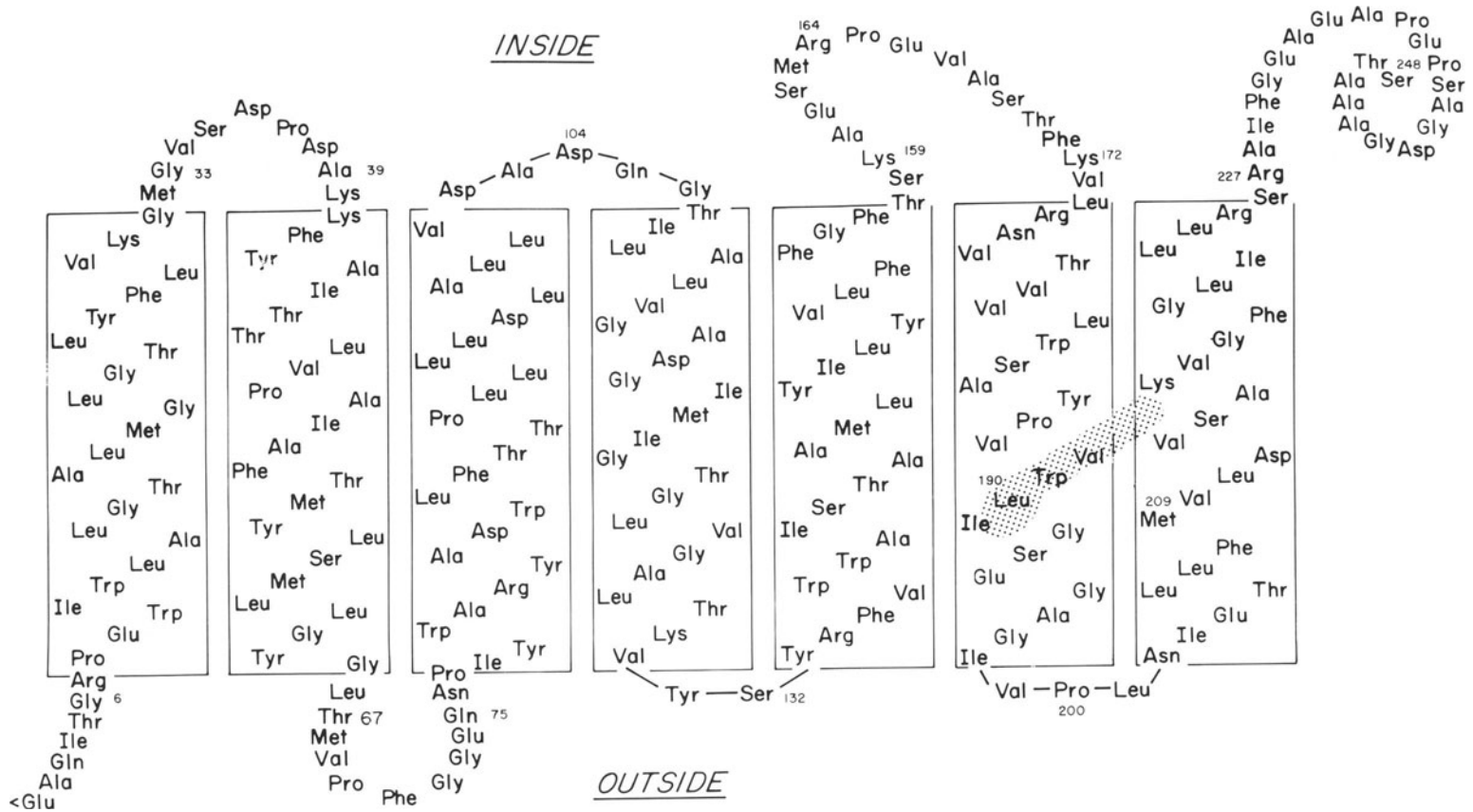

Figure 2. A revised arrangement of the polypeptide chain of bacteriorhodopsin across the membrane based on the work of Huang *et al.* (1982) and the proposal of Steitz *et al.* (1982). The shaded area represents the retinal analogue II forming a Schiff base with Lys^{216} at one end and cross links with amino acids 193 and 194 at the distant end.

bacteriorhodopsin (Engelman *et al.*, 1980). Modifications of the structures of helices 6 and 7 and the loop connecting helices 5 and 6 were required. Thus, if the overall size and structure of helix 7 in the original model of Engelman *et al.* (1980) is assumed to be correct, then Lys^{216}, the site of attachment of retinal, is positioned close to the center of the bilayer. To span the full length of the retinal analogue II (15 Å) between the Schiff base linkage and the cross-linking sites, amino acids 193 and 194, the latter amino acids must be placed within the layer, as shown in Fig. 2, rather than in the loop connecting helices 6 and 7, as in the original model. These seven helices vary in length from 23 to 29 residues, having a total of 173 residues in helices, or a 72% helical structure. Trewhella *et al.* (1983, 1984), using neutron diffraction data, assigned specific amino acid sequences to specific regions in the sequences in α helices. The seven helices had been shortened to contain either 20 or 21 amino acid residues, for a total of 57% helical content.

Paul and Rosenbusch (1985) found that the available predictive methods utilized for membrane proteins did not do well for the protein porin, which spans the outer membrane of *Escherichia coli*. By other criteria, this protein had a large β-strand component. They proposed to use the β turn as an indication of chain reversal and to avoid the hydrophobicity parameters. Such a model for porin was in good agreement with all experimental data. Application to the paradigm of hydrophobic membrane proteins, bacteriorhodopsin, revealed a pattern consistent with its α-helical folding, yielding seven helical segments. With the Chou–Fasman (1979a,b) algorithm for predicting β turns, it was found that the β turns appear between the proposed α-helical segments.

Wallace *et al.* (1986) evaluated three methods for the prediction of membrane protein secondary structure: the Chou–Fasman (1974a,b, 1978a,b) (CF), Garnier *et al.* (1978) (G), and Burgess *et al.* (1974) (B) algorithms. For bacteriorhodopsin, the experimental values of 74–83% α, 0–9% β sheet were quoted, having been determined by image reconstruction, infrared spectroscopy, x-ray diffraction, and circular dichroism. The predicted values obtained were: CF, 39–53% α, 22–28% β; G, 53% α, 41% β; B, 9% α, 26% β. It was thus concluded that these methods inadequately predict membrane protein structure. However, the methodologies used, because of the degree of resolution obtained with them, cannot be accepted with any degree of certainty, and this conclusion can be questioned. The predicted values for the photochemical reaction center of *Rhodopseudomonas viridis*, whose structure has been determined by x-ray crystallography (see Deisenhofer *et al.*, Chapter 2), were not satisfactory: L chain, experimental, 61% α, 21% β; CF, 33–44% α, 12–20% β, G, 19% α, 34% β; B, 3% α, 33% β; M chain, experimental, 65% α; CF, 16–33% α, 21–32% β; B, 11% α, 18% β; H chain, membrane, experimental, 100% α; CF, 8–19% α, 62–73% β; G, 28% α, 56% β; B, 54% α, 35% β (Wallace *et al.*, 1986).

Wu *et al.* (1982) examined lipid-induced conformations of peptide hormones and concluded that the α helix is the preferred conformation, as determined by circular dichroism, when the predicted conformation has potential for both α and β forms. These authors applied the Chou–Fasman (1974a,b, 1978a,b) algorithm to predict the secondary structure of bacteriorhodopsin and found that 60% β would predominate, with 20% α and 10% β turn. The six potential turns occurred at the seven ordered segments. Thus, it was concluded that the potential parameters for the amino acids might vary in a lipid environment.

The above predictions of bacteriorhodopsin used the conformational parameters based on 29 proteins. On the basis of the 29 data base and the 64 data base (Chou, Chapter 12), the results obtained were as follows: 29 data base, 16% α, 57% β; 64 data base, 21% α, 48% β (G. D. Fasman, unpublished data). Thus, there was a small improvement from using the larger data base.

The conformation of bacteriorhodopsin is as yet undetermined. Extensive studies using

circular dichroism to determine the secondary structure of bacteriorhodopsin have been carried out. Studies of Dencher and Heyn (1978) showed that nonionic detergents caused dissociation of bacteriorhodopsin into monomers without a change in the retinal environment. They also showed that light and dark adaptation causes large absorbance changes. Octyglucoside and Triton 100-X caused changes in the circular dichroism spectrum with time, with a disappearance of the exciton bands at 318 and 560 nm. However, no change in secondary structure of the protein was observed in the 200 to 250-nm region. Nabedryk *et al.* (1985) summarized the many circular dichroism studies on bacteriorhodopsin. The helical content is reported as 45–83% (Beecher and Cassim, 1976; Long *et al.*, 1977; Jap *et al.*, 1983; Mao and Wallace, 1984). The value obtained depends on the state of the bacteriorhodopsin. The intact purple membrane, the solubilized bacteriorhodopsin, and bacteriorhodopsin reconstituted in lipid vesicles all yield different helical contents. This is probably because of both difference in secondary structure and, in large part, absorption flattening and scattering, to various degrees, in each of these preparations.

Jap *et al.* (1983) considered the electron microscopy study of Hayward and Stroud (1981), which reported possible β structure in bacteriorhodopsin, and interpreted their circular dichroism and infrared data as having β structure. They suggested a model for bacteriorhodopsin consisting of only five α-helical rods (50%) and four β sheets of 11 amino acids (20% β). Nabedryk *et al.* (1985) interpreted their circular dichroism and polarized infrared studies to indicate 74 ± 5% transmembrane helices and detected no significant contribution of β strands running perpendicular to the membrane plane. Their data agree with the reports of Mao and Wallace (1984) and the model of Trewhella *et al.* (1983). The studies of Glaeser and Jap (Jap and Kong, 1986; Glaeser and Jap, 1985; Jap *et al.*, 1983) find that by using the theoretical analysis and empirical estimation of Gordon and Holzworth (1971) and the data of Mao and Wallace (1984), there is no major absorption flattening for submicrometer-sized membranes. In agreement with the reports of Mao and Wallace (1984), Khorana's laboratory, using circular dichroism, reports that bleaching of delipidated bacteriorhodopsin dissolved in deoxycholate caused a 25% drop in the $[\theta]_{222}$ (London and Khorana, 1982), thus indicating a significant conformational change in protein structure. Huang *et al.* (1981) report a helical content of 50% in SDS. Popot *et al.* (1987) report a helical content of 58% by circular dichroism measurements of a reconstituted, cleaved, two-chain assembly of bacteriorhodopsin, whereas in lipid vesicles a helical content of 75–80% was found. Thus, a wide divergence in helical content of bacteriorhodopsin is reported. Differences are probably related to the preparative methods, particle size, the degree of association of the puple membrane, and the methodology used to deconvolute the circular dichroism curves to obtain the secondary structure.

B. Acetylcholine Receptor

Another thoroughly studied membrane protein is the nicotinic acetylcholine receptor protein (AcChR). The AcChR complex is comprised of five structurally similar subunits (two α, one β, one γ, and one δ) that stack next to each other so that a channel forms between the subunits when the AcChR is in the open conformation. This assembly contains both the binding site for the neurotransmitter and the cation-gating unit. Smythies (1980) applied the Chou–Fasman (1978a) predictive scheme to the amino acid sequence of the N-terminal segment of the α subunit. Two α helices cross linked by four ionically bound complementary amino acids (Arg/Lys to Glu) were suggested as the secondary structure.

Guy (1981) modeled the protein structure of agonist, competitive antagonist, and snake-neurotoxin-binding sites using the first 54 residues of AcChR α subunit from *Torpedo califor-*

nica. These models were based on the premise that the N-terminal portion of the subunits form the outermost extracellular surface of the AcChR and that agonists bind to this portion. The structure proposed was based on a comparison of the secondary structure predicted by the methods of Chou and Fasman (1978a) (using the Levitt, 1978, data base) and Lim (1974a–c). The tertiary structures of the α subunit were developed by requiring that it bind strongly to the crystal structures of the snake neurotoxins α cobra toxin and erabutoxin b. Finally, the quaternary structures were developed on the bases of interactions among the proposed tertiary structure (Devillers-Thiery *et al.*, 1983). A 1350-base-pair-long cDNA was cloned by Devillers-Thiery *et al.* (1983) and sequenced to yield the amino acid sequence (437 residues) of the α subunit of the receptor. Four regions of high hydrophobicity were assumed to be helical regions, and a very approximate model was constructed. Kosower (1983a,b, 1987) predicted six transmembrane segments of the α subunit of the AcChR from the amino acid sequence as determined by Toda *et al.* (1982) on the basis of hydrophobic sequences (Kyte and Doolittle, 1982). The choice of ion channel elements is based on the theory of single-group rotation (Kosower, 1982, 1983a).

Noda *et al.* (1983) reported the primary structure of the γ-subunit precursor of the AcChR. Comparison with the four subunits ($\alpha_2\beta\gamma\delta$) revealed marked homology. The hydrophobicity profiles (Hopp and Woods, 1981) as well as the predicted secondary structures (Chou and Fasman, 1978a,b) were used. As with the predicted α subunit (Noda *et al.*, 1982), based on both the hydrophilicity profile (Hopp and Woods, 1981) and the predicted secondary structure (Chou and Fasman, 1978a,b), each subunit has four strongly hydrophobic regions, which may represent regions with secondary structure that may be transmembrane segments or may be involved in intersubunit interaction. Claudio *et al.* (1983) also obtained the sequence of the γ subunit and predicted the secondary structure on the basis of hydropathicity profiles (Kyte and Doolittle, 1982; Hopp and Woods, 1981). Four stretches of hydrophobic residues were located and were attributed to membrane-spanning regions. Finer-Moore and Stroud (1984) developed an amphipathic analysis of membrane proteins and applied it to acetylcholine (see discussion of method in Section VI.D). On the basis of the prediction of amphipathic α helices, hydrophobicity plots, and the secondary structure as predicted by the Garnier *et al.* (1978) method, a model of the subunits of acetylcholine receptor subunits was devised. Stroud and Finer-Moore (1985) have reviewed the work on the acetylcholine receptor.

In summary, the predictive scheme of Claudio *et al.* (1983), Noda *et al.* (1983), and Devillers-Thiery *et al.* (1983) proposed four hydrophobic transmembrane domains. A fifth amphipathic transmembrane domain was proposed by Finer-Moore and Stroud (1984) and Guy (1981). Immunologic techniques suggest two additional amphipathic domains (Criado *et al.*, 1985) to account for the transmembrane orientation of various sequences.

Brisson and Unwin (1985) resolved an electron microscopic image and determined that the acetylcholine receptor had five membrane-spanning subunits, which were shown to be at pentagonally symmetrical positions around a channel over a large fraction of their length. The channel consists of a wide synaptic portion and a narrow portion extending through the membrane into the interior of the cell. Thus, the predictive schemes of Finer-Moore and Stroud (1984) and of Guy (1981) appear to give the most accurate representation of the structure.

Dixon *et al.* (1986) cloned the gene and cDNA for the mammalian β-adrenergic receptor (βAR). This receptor is known to modulate adenylate cyclase activity and consists of a catalytic moiety and regulatory guanine nucleotide-binding proteins, which provide the effector mechanism for the intracellular actions of many hormones and drugs. Analysis of the amino acid sequence of βAR indicates a significant amino acid homology with bovine rhodopsin and suggests that, like rhodopsin, βAR possesses multiple membrane-spanning regions.

Hydropathicity profiles of the βAR amino acid sequences were produced using the analysis of Hopp and Woods (1981) and were similar to the rhodopsins, of which bacteriorhodopsin is known to contain seven membrane-spanning helices (Engelman *et al.*, 1982).

Numa and co-workers (Kubo *et al.*, 1986) cloned, sequenced, and expressed the complementary cDNA encoding the muscarinic acetylcholine receptor. The muscarinic receptor was found to be homologous with the β-adrenergic receptor and rhodopsin in both amino acid sequences and suggested transmembrane topography. The Kyte and Doolittle (1982) method of expressing the hydropathicity was utilized.

Peralta *et al.* (1987) determined the amino acid sequence of the 466 amino acids comprising the M_2 muscarinic acetylcholine receptor. This receptor is predicted to have seven membrane-spanning regions distinguished by the disposition of a large cytoplasmic domain. The atrial muscarinic receptor is distinct from the cerebral muscarinic receptor gene product, sharing only 38% overall amino acid homology and possessing a completely nonhomologous large cytoplasmic domain. The hydrophobicity, as determined by the Kyte and Doolittle (1982) method, suggested six transmembrane regions with a seventh region of less pronounced hydropathicity. Bonner *et al.* (1987) isolated cDNAs for three different muscarinic acetylcholine receptors from a rat cerebral cortex library, and the cloned receptors were expressed in mammalian cells. This gene family provided a new basis for evaluating the diversity of muscarinic mechanisms in the nervous system. Hydropathicity profiles of these receptor sequences indicated the presence of seven transmembrane domains.

Greeningloh *et al.* (1987) have demonstrated that the ligand-binding subunit of an inhibitory central neurotransmitter receptor, the glycine receptor, shares homology with the nicotinic acetylcholine receptor polypeptide family. The primary structure of the 48K glycine receptor was deduced from its cDNA. Analysis for regional hydropathicity (Hopp and Woods, 1981) revealed four hydrophobic segments long enough to form transmembrane α helices, which is similar to that of the nicotinic acetylcholine receptor subunits. Recent reviews covering the acetylcholine receptor area have appeared (McCarthy *et al.*, 1986; Hucho, 1986), and a discussion of the family of receptors coupled to guanine nucleotide regulatory proteins, which include the β-adrenergic receptor, covers the area in great detail.

C. ATPases

Sodium- and potassium-dependent ATPase [$(NA^+ + K^+)$-ATPase], which is responsible for the active transport of Na^+ and K^+, is distributed universally among animal cell membranes and consists of two subunits, α and β. The larger α subunit with a relative molecular mass (M_r) of 84,000–120,000 is thought to have the catalytic role. Numa and coworkers (Kawakami *et al.*, 1985) determined the primary sequence of the α subunit of the *Torpedo californica* $(Na^+ + K^+)$-ATPase. Analysis by Kawakami *et al.* (1985) of the amino acid sequence for local hydrophobicity (Kyte and Doolittle, 1982) and the predicted secondary structure (Chou and Fasman, 1978b) suggests the presence of at least six transmembrane segments, presumably α-helical structures. Shull *et al.* (1985) characterized the complementary DNA for the catalytic subunit of the sheep kidney $(Na^+ + K^+)$-ATPase. The 1016-amino-acid sequence was analyzed by the Kyte and Doolittle (1982) procedure to establish the hydrophobic sequences, and eight major hydrophobic sequences (H1–H8), ranging in size from 17 to 29 amino acids, were found, which are the likely transverse membrane regions.

MacLennon *et al.* (1985) deduced the amino acid sequence from the cDNA sequence of the $(Ca^+ + Mg^+)$-dependent ATPase from rabbit muscle sarcoplasmic reticulum. They propose that the protein has three cytoplasmic domains joined to a set of ten transmembrane helices by a narrow pentahelical stalk. The homology between the amino acid sequence of the

extramembranous segments of the Ca^+-ATPase and the K^+-ATPase of *Escherichia coli* was previously pointed out by Shull *et al.* (1985). There was ~50% homology in some limiting regions. MacLennan *et al.* (1985) identified the transmembrane sequences by their hydrophobicity and applied standard methods only to the extramembranous regions. These regions are mainly on the cytoplasmic regions of the membrane and are separated by transmembrane hydrophobic hairpins, which are located by the polarity plot. This plot gives results similar to those of the Kyte and Doolittle (1982) plot. An earlier prediction of the Ca^+-ATPase (Allen *et al.*, 1980) produced an ambiguous answer using the method of McLachlan (1978); therefore, the present authors utilized the procedure of Taylor and Thornton (1984), which resolved some of the ambiguities in priority of strand and helix in the earlier prediction.

The segments connecting the globular region to the membrane are of particular interest, as they are likely to contain the Ca^{2+}-binding region and to transmit movements from the phosphorylation and nucleotide domains that bring about ion translocation. Brandl *et al.* (1986) demonstrated that rabbit genomic DNA contains two genes that encode Ca^{2+}-ATPase of fast twitch and slow twitch (and cardiac) sarcoplasmic reticulum, respectively. The deduced amino acid sequences of the products of the two genes are highly conserved in putative Ca^{2+}-binding regions, in sectors leading from cytoplasmic domains into transmembrane domains, and in transmembrane helices. The assignment of the secondary structure was that previously described by Brandl *et al.* (1986). Ovchinnikov *et al.* (1986) determined the primary structure and spatial organization of pig kidney ($Na^+ + K^+$)-ATPase. The cDNA complementary to pig kidney RNAs coding for α and β subunits of ($Na^+ + K^+$)-ATPase were cloned and sequenced. Selected tryptic hydrolysis of the α subunit within the membrane-bound enzyme and trypsin hydrolysis of the immobilized isolated β subunit were also performed. The mature α and β subunits contain 1016 and 302 amino acids, respectively. Hydropathicity profiles were determined by two methods (Ovchinnikov, 1982; Capaldi and Vanderkooi, 1972; Kyte and Doolittle, 1982). Eleven regions were found that could serve as intramembrane segments. However, comparison of the above data and those on amino acid sequences of the hydrophilic peptides of the α subunit (Arzamazova *et al.*, 1985) allowed a selection of transmembrane segments. They assumed that there are five probable transmembrane fragments localized inside the membrane, with the remaining two, shorter and less hydrophobic, being exposed outside.

The plasma membrane ATPase of *Neurospora crassa* was deduced from genomic and cDNA sequences by Hagler *et al.* (1986). It contains a protein of 920 amino acids possessing as many as eight transmembrane segments. The *Neurospora* ATPase shows a significant amino acid sequence homology with the ($Na^+ + K^+$)- and Ca^{2+}-transporting ATPases of animal cells, particularly in regions that appear to be involved in ATP binding and hydrolysis. The method of Engelman *et al.* (1986) was used to determine the hydropathicity, and eight transmembrane segments were predicted.

Vogel *et al.* (1986) mapped the ATP substrate site in the epidermal growth factor (EGF-K). Sequence alignment of four protein kinases, cGMP-dependent protein kinase (cGK), cAMP-dependent protein kinase (cAK), the γ-subunit of phosphorylase b kinase (γ), and the EGF receptor (EGF-K) indicated that only a few residues are strictly conserved. The segment 1–11 contains three invariant Gly residues and was predicted to adopt an α-helical structure (Nagano, 1974; Maxfield and Scheraga, 1976; Robson and Suzuki, 1976; Chou and Fasman, 1978a). Such an element is consistently found in all nucleotide-binding sites, including various dehydrogenases (Sternberg and Taylor, 1984). The secondary structure of segment 2, including Cys^{157}, was analyzed by various predictive models (Nagano, 1974; Maxfield and Scheraga, 1976; Robson and Suzuki, 1976; Chou and Fasman, 1978a; von Heijne, 1981a,b), including a hydropathy plot using a similar approach to that outlined by Taylor and Thornton

(1984). Turns for the EGF receptor were predicted using several methods (Maxfield and Scheraga, 1976; Robson and Suzuki, 1976; Chou and Fasman, 1978a; Sternberg and Taylor, 1984). Kanazawa *et al.* (1985) cloned a mutant gene of the γ subunit of H^+-tranlocating ATPase from *Escherichia coli,* mutant NR70, and found that there were seven amino acids deleted from the amino-terminal portion. This deletion resulted in the loss of a hydrophobic domain (23–28) as determined by a Kyte and Doolittle (1982) hydrophobicity plot. This caused total loss of the assembly of F_1 on the membrane.

The plasma membrane ATPase of plants and fungi is a hydrogen pump. The protein gradient generated by the enzyme drives the active transport of nutrients by H^+–symport. In addition, the external acidification in plants and the internal alkalinization in fungi, both resulting from activation of the H^+ pump, have been proposed to mediate growth responses. The ATPase has a relative molecular mass similar to those of the $(Na^+ + K^+)$- and Ca^{2+}-ATPases of animal cells and, like those proteins, forms an aspartylphosphate intermediate. Serrano *et al.* (1986) cloned, mapped, and sequenced the gene encoding the yeast plasma membrane ATPase (PMAI). The strong homology between the amino acid sequences encoded by PMAI and those of $(Na^+ + K^+)$-, Na^+-, K^+-, and Ca^{2+}-ATPases is consistent with the notion that the family of cation pumps that form a phosphorylated intermediate evolved from a common ancestral ATPase. The hydrophobicity profiles of $(Na^+ + K^+)$-ATPase, Ca^{2+}-ATPase, and K^+-ATPase are very similar to that of the H^+-ATPase as determined by the Kyte and Doolittle (1982) and Eisenberg *et al.* (1984a,b) methods. A possible ten-trans-membrane-segment structure was proposed.

D. T-Cell Receptor

Hedrick *et al.* (1984a) have isolated a cDNA clone, TM86, that represents a species of mRNA that is expressed in T cells but not in B cells, encodes a membrane-associated protein, and is rearranged in T cells. Hedrick *et al.* (1984b) have analyzed clone TM86 and also several cross-reacting clones isolated from a thymocyte cDNA library by nucleotide sequencing. The data reveal a remarkable resemblance between the amino acid sequences of the protein encoded by the mRNAs represented by the TM86 and related cDNA clones and the immunoglobulin proteins of B cells. An analysis of the amino acid sequences was carried out to predict the secondary structure by the method of Kyte and Doolittle (1982). The hydrophobicity plot indicates a number of interesting features. (1) The 86T1 protein has the alternating hydrophobic–hydrophilic stretches characteristic of globular proteins (Kyte and Doolittle, 1982). (2) The predicted leader polypeptide occurs in a very adequate hydrophobic environment, and, most importantly, it predicts a possible transmembrane region spanning the end of the 86T1 sequence, followed by a string of positive charges (Lys-Arg-Lys), which are characteristic of the cystoplasmic portion of a number of lymphocyte cell-surface markers (Kabat *et al.*, 1983).

Barth *et al.* (1985) found ten different V_β gene segments when the sequences of 15 variable (V_β) genes of the mouse T-cell receptor were examined. Both the T-cell receptor and the immunoglobulins are heterogeneous cell-surface glycoproteins that can recognize many antigens (e.g., Allison *et al.*, 1982). T-cell receptor molecules are composed of α and β chains, each of which, like the immunoglobulin chains, is divided into variable (V) and constant (C) regions (e.g., Allison *et al.*, 1982). The V regions together form the antigen-binding domain. Like the immunoglobulin gene, they are divided into separate V_β, diversity (D_β), and joining (J_β) gene segments that are assembled by recombination during T-cell development to form a V_β that is associated with either of two constant ($C_\beta 1$ and $C_\beta 2$) genes. Barth *et al.* (1985) have compared the V_β and immunoglobulin V segments by analyzing them

for two properties believed to reflect important structural features of these molecules: the distribution of β-pleated-sheet-forming potential (Chou and Fasman, 1978b) and the predicted hydrophobicity profile (Kyte and Doolittle, 1982). The analysis showed nearly identical results for mouse V_β, V_H, and V_κ segments. This agrees with work of Patten *et al.* (1984). The authors conclude that the T-cell receptor and immunoglobulin molecules fold into comparable tertiary structure.

Arden *et al.* (1985) analyzed 19 complementary DNA clones encoding the α chain of the T-cell antigen receptor derived from thymic transcripts. The primary sequence was analyzed by several methods for primary structural patterns. The variability plot of Wu and Kabat (1970) examines the variation of each residue position between members of a set of similar sequences. In immunoglobulin and β-chain V regions, two regions of relative hypervariability were noted, with a third hypervariability region positioned at the junction of the V, D, and J gene segments. The Kyte and Doolittle (1982) hydropathicity and the Chou and Fasman (1974a,b) method, which determines the relative potential for β-pleated-sheet formation were used. Both methods showed that the V_α, V_β, V_H, and V_κ sequences are very similar. These data suggest that the T-cell antigen receptor and immunoglobulin molecules have similar structure.

Tillinghast *et al.* (1986) characterized the variability of the expressed T-cell receptor β-chain repertoire and compared this variability to the known murine β-chain repertoire. No features were found that distinguish human V_β genes from their murine counterparts. Variability among the 15 human V_β genes was analyzed by the Wu and Kabat (Kabat *et al.*, 1983) variability plot. Twelve human V_β genes were also analyzed with the hydrophobicity algorithm of Hopp and Woods (1981). The areas of highest variability on the Wu–Kabat analysis correspond, in general, with the hydrophilic regions of the molecule, consistent with these regions being available to interact with ligand. The hydrophilic plot shows conservation of several hydrophobic regions that are thought to be involved in the formation of the common tertiary structure shared among all T-cell receptor chains as well as members of the immunoglobulin supergene family (e.g., Novotny and Haber, 1985). Schiffer *et al.* (1986) examined the β-chain variable regions of human, mouse, and rabbit T-cell receptors for antigen and were able to classify them into two subgroups by the application of the variability plots of Wu and Kabat (1970).

E. Sodium Channel

The sodium channel is a membrane protein that mediates the voltage-dependent modulation of the sodium ion permeability of electrically excitable membranes (Catterall, 1984). Numa and co-workers (Noda *et al.*, 1984) cloned and sequenced the cDNA for the *Electrophorus electricus* electroplax sodium channel. Analysis of the derived protein sequence indicated a protein consisting of 1820 amino acids that exhibited four repeated homologous units, determined by homology matrix comparison (Toh *et al.*, 1983), that are presumably oriented in a pseudosymmetric fashion across the membrane. Each homology unit contains a unique segment with clustered positively charged residues, which may be involved in the gating structure, possibly in conjunction with negatively charged residues clustered elsewhere. The amino acid sequence of the sodium channel protein was analyzed for predicted secondary structure (Chou and Fasman, 1978b) and for local hydropathy (Kyte and Doolittle, 1982). The segments that can form α helix or β sheet extend over a wide range of the polypeptide chain.

The hydropathy profile along the polypeptide chain indicates that each repetitive homology unit contains five hydrophobic segments (S1, S2, S3, S5, and S6) at equivalent positions. In addition, each unit contains a characteristic segment with strong positive charge (S4) located

between segments S3 and S5. Segments S5 and S6 in each repeat correspond to highly hydrophobic regions with predicted secondary structures that comprise a continuous stretch of 24–38 uncharged amino acid residues including many nonpolar residues. These hydrophobic regions are flanked on both sides by charged residues. Because such a sequence is characteristic of transmembrane protein segments, segments S5 and S6 most probably transverse the membrane, forming α-helical structures. Segments S1, S2, and S3 in each repeat represent hydrophobic regions with predicted secondary structure consisting of 18–28 amino acid residues that are largely nonpolar but include a few charged residues. Negatively charged residues predominate in segments S1 and S3, whereas both positively charged and negatively charged residues are present in segment S2, so that this segment generally has no net charge. If an α-helical structure is assumed for these segments, the charged side chains in them are clustered largely on one side of the helix, the opposite side being occupied mostly by nonpolar side chains. Thus, segments S1, S2, and S3 can also span the membrane, forming amphipathic α-helical structures. Segment S4 in each repeat represents a positively charged region with predicted secondary structure comprising 21 amino acid residues, including five to seven Arg or Lys residues, Arg being predominant. This segment exhibits a unique structural feature in that the Arg or Lys residues are located every third position. The residues intervening between these basic residues are mostly nonpolar. If an α-helical structure is assumed for this segment, the positively charged side chains would be distributed along a spiral line, making a three quarters to one turn. Alternatively, if a 3_{10} helical structure is assumed, they would lie on one side of the helix so that the segment would be strongly amphipathic. On the other hand, if a β-sheet structure is assumed, they would extend alternately toward both sides of the peptide backbone.

It seems reasonable to postulate that the four repeated homology units are oriented in a pseudosymmetric fashion across the membrane. If this is the case, each unit should contain an even number of transmembrane segments because no additional hydrophobic segments are predicted outside the homology units. The transmembrane segments are most likely to be involved directly or indirectly in the formation of the ionic channel.

Noda *et al.* (1986) isolated complementary DNA clones derived from two distinct rat brain mRNAs encoding sodium channel large polypeptides and have determined the complete amino acid sequences of the two polypeptides (designated sodium channels I and II) as deduced from the cDNA sequences. A partial DNA sequence complementary to a third homologous mRNA from a rat brain has also been cloned. The degree of homology is 87%, 62%, and 62% for the rat I/rat II, rat I/*Electrophorus,* and rat II/*Electrophorus* pairs, respectively. Rat I and II sodium channel also contain four homologous repeats (deletions or insertions). The regions corresponding to these repeats are highly conserved among all three sodium channels, whereas the remaining regions, all of which are assigned to the cytoplasmic side of the membrane, are less well conserved except for the region between repeats III and IV. Rat sodium channel I and II have hydropathicity profiles similar to that of the *Electrophorus* sodium channel (Noda *et al.*, 1984). Each internal repeat has five hydrophobic segments (S1, S2, S3, S5, and S6) and one positively charged segment (S4), all of which exhibit predicted secondary structure (Chou and Fasman, 1978b). The distribution of charge is similar to the *Electrophorus* channel. However, based on other considerations, a different arrangement of segments is proposed, namely, that all of the six segments S1–S6 span the membrane, presumably as α helices, forming an ion channel.

Guy and Seetharamulu (1986) have presented alternative models of the sodium channel in *Electrophorus electricus*. The favored model has the same general folding pattern as postulated by Noda *et al.* (1984) except that each domain has four additional transmembrane segments, S3, S4, S6, and S7. The postulated activation gating mechanism involves a screwlike motion

of the S4 helices. The sequence was analyzed with a method that predicts which portions of α helices and β structures are exposed to water, buried inside the protein, or exposed to lipid (Guy, 1984, 1985). Interactions among transmembrane segments were examined by constructing Nicholson molecular models of alternative conformations and by using computer graphics. Primary features of the models are that the sodium channel lining is formed by positively and negatively charged segments and that movement of the positively charged segments underlies voltage-dependent activation.

Table II contains a list of predicted structures of membrane proteins.

Table II. Predicted Structure of Membrane Proteins

Protein	Method used[a]	Reference
Coat protein	C–F	Green and Flanagan (1976)
Filamentous bacteriophages	L	
Pf1; fd; ZJ-2		
Outer envelope of *E. coli*		
Erythrocyte glycophorin		
Bovine liver microsomal	C–F	Fleming *et al.* (1978)
Cytochrome b_5		
Signal sequences	ΔG_{trans}	von Heijne and Blomberg (1979)
Prelysozyme	Chothia	
M-41 κ L chain		
Prealbumin		
β-Lactamase		
Phage fl-coat		
E. coli lipoprotein		
M-104E λ, L chain		
Preconalbumin		
p-450 M_2		
Mitochondrial ADP/ATP translocase	K–D	Saraste and Walker (1982)
	S	
	M	
α and β subunits of ATP synthase myosin; kinases and other ATP-requiring enzymes and a common nucleotide fold	M	Walker *et al.* (1982)
Bacteriorhodopsin	C–F	Senior (1983)
E. coli H^+-ATPase	V–H	
Cytochrome oxidase		
subunit III		
UNC E protein		
Lac permease of *E. coli*	K–D	Foster *et al.* (1983)
	C–F	Menick *et al.* (1986)
UNC operon	K–D	Walker *et al.* (1984)
ATP synthetase	W	
	S–M	
	M	
	S	
Photosynthetic reaction center, B870 antenna and flanking polypeptides from *R. capsulta*	K–D	Youvan *et al.* (1984)

Table II. (Continued)

Protein	Method used[a]	Reference
ATP binding site of oncogene product v-*src*	T–T	Sternberg and Thorton (1984)
Epidermal growth factor receptor, cAMP-dependent protein kinase		
Bovine brain myelin proteolipid	K–D C–F R	Lauresen *et al.* (1984)
Protein disulfide isomerase of the lumen of the endoplasmic reticulum	F–M–S	Edman *et al.* (1985)
GTPase of bovine rod outer segments	C–F	Yatsunami and Khorana (1985)
3-Hydroxymethylglutaryl–coenzyme A reductase (glycoprotein of the endoplasmic reticulum)	F–M–S	Liscum *et al.* (1985)
Phosphocarrier protein factor III^{lac} of lactose-specific phosphotransferase system of *Staphylococcus aureus*	C–F K–D St.	Stuber *et al.* (1985)
Colicin A	G–O–R C–F F–M–S E(1)	Pattus *et al.* (1985)
Human insulin receptor	K–D	Ullrich *et al.* (1985)
Murine anion exchange protein	K–D S–E	Kopito and Lodish (1985)
Wheat chloroplast gene for CF_0 subunit of ATP synthase	K–D C–F G–O–R	Bird *et al.* (1985)
Human glucose transporter	K–D E(2) C–F	Mueckler *et al.* (1985)
Glucose permease of bacterial phosphotransferase	K–D E(3) S–E	Erni and Zanolari (1986)
32-kd Qb-binding chloroplast thylakoid membrane protein	R–H–A	Sayre *et al.* (1986) Rao *et al.* (1983)
Potential membrane proteins of Epstein–Barr virus	C–F G–O–R H–W	Modrow and Wolf (1986)
Nucleotide-binding domain in the β subunit of *E. coli* F_1-ATPase	W–S–G	Duncan *et al.* (1986)
Mannitol permease of *E. coli*	K–D	Stephan and Jacobson (1986) Lee and Saier (1983)
Protein 4.1, human erythrocyte membrane skeleton	F–M–S	Conboy *et al.* (1986)
Ca^{2+}-dependent membrane-binding proteins (col-electron) in *Torpedo marmorata* and mammalian cells	C–F	Geisow *et al.* (1986)
Amelogenin from bovine tooth enamel	C–F	Renugopalakrishnan *et al.* (1986)
Coronavirus protein E1	C–F E(2) L H	Rottier *et al.* (1986)
$GABA_A$ benzodiazepine receptor from bovine brain	K–D H–W	Schofield *et al.* (1987)
Dihydropyridine calcium channel blocker receptor from skeletal muscle (DHP receptor)	K–D C–F	Tanabe *et al.* (1987)

(*continued*)

Table II. (Continued)

Protein	Method used[a]	Reference
Family of receptors coupled to guanine nucleotide regulatory proteins	K–D	Dohlman *et al.* (1987)
Human platelet glycoprotein 1b	K–D	Lopez *et al.* (1987)
Willebrand factor-binding domain of platelet membrane glycoprotein 1b	H–W K–D C–F	Titani *et al.* (1987)
Human cation-dependent mannose-G-phosphate-specific receptor	N–A (C–F) R–R	Dohlman *et al.* (1987)
Bovine retinal S antigen	G–O–R N–T	Shinohara *et al.* (1987)
Murine band 3 gene of erythrocytes	K–D C–F	Kopito *et al.* (1987)
Insulinlike growth factor II receptor	K–D	Morgan *et al.* (1987)
Growth hormone receptor and serum binding protein	K–D	Leung *et al.* (1987)
Myelin P_2 protein	C–F B–B	Shin and McFarlane (1987)
Calcium-dependent membrane-binding proteins p35, p36, and p32	G–O–R	Taylor and Geisow (1987)
Scrapie prion protein	E K–D C–F G–O–R F–M–S	Bazan *et al.* (1987)
Bovine substance-K receptor (SKR)	K–D	Masu *et al.* (1987)

[a]Sources of references: B–B, Barkovsky and Bandarin (1979); C–F, Chou-Fasman (1978a); E(1), Eisenberg *et al.* (1982a); E(2), Eisenberg *et al.* (1984a); E(3), Eisenberg (1982b); F–M–S, Finer-Moore and Stroud (1984); G–O–R, Garnier *et al.* (1978); H, von Heijne (1981b); H–W, Hopp and Woods (1981); K–D, Kyte and Doolittle (1982); L, Lim (1974a,b,c); M, McLachlan (1977); N–A, Novotny and Auffray (1984); N–T, Nozaki and Tanford (1971); R, Rose (1978); R–H–A, Mohana-Rao *et al.* (1983); R–R, Rose and Roy (1980); S, Staden (1982); S–E, Schiffer and Edmunson (1967); S–M, Staden (1982); St., Stuber (1982); T–T, Taylor and Thornton (1983, 1984); V–H, von Heijne (1981a,b); W, Walker *et al.* (1984).

ACKNOWLEDGMENTS. This work was supported by a National Science Foundation grant, DMB-8713193. I would like to thank Pamela Gailey for the preparation of the manuscript.

IX. REFERENCES

Abercrombie, D. M., and Khorana, H. G., 1986, Regeneration of native bacteriorhodopsin following acetylation of ϵ-amino groups of Lys-30, -40 and -41, *J. Mol. Biol.* **261**:4875–4880.

Allen, G., Trinnaman, B. J., and Green, N. M., 1980, The primary structure of the calcium ion-transporting adenosine triphosphatase protein of rabbit skeletal sarcoplasmic reticulum, *Biochem. J.* **187**: 591–616.

Allison, J. P., McIntyre, B. W., and Bloch, D., 1982, Tumor-specific antigen of murine T-lymphoma defined with monoclonal antibody, *J. Immunol.* **129**:2293–2300.

Ananthanaryanan, V. S., and Bandekar, J., 1976, Application of one-dimensional Isling model to the secondary structure in globular proteins: Predicted β-regions, *Int. J. Peptide Protein Res.* **8**:615–623.

Ananthanaryanan, V. S., Brahmachari, S. K., and Paltabiraman, N., 1984, Proline-containing β-turns in peptides and proteins: Analysis of structural data on globular proteins, *Arch. Biochem. Biophys.* **232**:482–495.

Anderer, F. A., 1963, Recent studies on the structure of tobacco mosaic virus, *Adv. Protein Chem.* **13**:1–35.

Anderson, W., Burt, S., and Loew, G., 1979, Energy-conformation studies of frequency of β-turns in tetrapeptide sequences, *Int. J. Peptide Protein Res.* **14**:402–408.

Anfinsen, C. B., 1959, *The Molecular Basis of Evolution*, John Wiley & Sons, New York.

Anfinsen, C. B., 1973, Principles that govern the folding of protein chains, *Science* **181**:233–230.

Anfinsen, C. B., Haber, E., Sela, M., and White, F. H., 1961, The kinetics of formation of native ribonuclease during oxidation of the reduced polypeptide chain, *Proc. Natl. Acad. Sci. U.S.A.* **47**:1309–1314.

Arden, B., Klotz, J. L., Siu, G., and Hood, L., 1985, Diversity and structure of genes of the α family of mouse T-cell antigen receptor, *Nature* **316**:783–787.

Argos, P., 1987a, Analysis of sequence-similar pentapeptides in unrelated protein tertiary structures. Strategies for protein folding and a guide for site-directed mutagenesis, *J. Mol. Biol.* **197**:331–348.

Argos, P., 1987b, A sensitive procedure to compare amino acid sequences, *J. Mol. Biol.* **193**:385–396.

Argos, P., and Mohana-Rao, J. K., 1985, Relationships between exons and the predicted structure of membrane-bound proteins, *Biochim. Biophys. Acta* **827**:283–297.

Argos, P., and Palau, J., 1982, Amino acid distribution in protein secondary structures, *Int. J. Peptide Protein Res.* **19**:380–393.

Argos, P., Schwarz, J., and Schwarz, J., 1976, An assessment of protein secondary structure prediction methods based on amino acid sequence, *Biochem. Biophys. Acta* **439**:261–273.

Argos, P., Rossman, M., and Johnson, J. E., 1977, A four-helical super-secondary structure, *Biochem. Biophys. Res. Commun.* **75**:83–86.

Argos, P., Hanei, M., and Garavito, R. M., 1978, The Chou-Fasman secondary structure prediction method with an extended data base, *FEBS Lett.* **93**:19–24.

Argos, P., Mohana-Rao, J. K., and Hargrave, P. A., 1982, Structural prediction of membrane-bound proteins, *Eur. J. Biochem.* **128**:565–575.

Arzamazova, N. M., Arystarkhova, E. A., Shafieva, G. I., Nazimov, I. V., Aldanova, N. A., and Modyanov, N. N., 1985, Primary structure of the α-subunit of $Na^+ + K^+$-ATPase. I. Analysis of hydrophilic fragments of the polypeptide chains, *Bioorg. Khim.* **11**:1598–1601.

Aubert, J.-P., and Loucheaux-Lefebvre, M.-H., 1976, Conformational study of α_1-acid glycoprotein, *Arch. Biochem. Biophys.* **175**:400–409.

Aubert, J.-P., Biserte, G., and Loucheux-Lefebvre, M. H., 1976, Carbohydrate–peptide linkage in glycoproteins, *Arch. Biochem. Biophys.* **175**:410–418.

Aubert, J.-P., Helbecque, N., and Loucheux-Lefebvre, M.-H., 1981, Circular dichroism studies of synthetic Asn-X-Ser/Thr-containing peptides: Structure of glycosylation relationship, *Arch. Biochem. Biophys.* **208**: 20–29.

Bacon, D. J., and Anderson, W. F., 1986, Multiple sequence alignment, *J. Mol. Biol.* **191**:153–161.

Baldwin, R. L., 1980, The mechanism of folding of ribonucleases A and S, in *Protein Folding* (R. Jaenicke, ed.), Elsevier Amsterdam, pp. 369–384.

Barker, W. C., Hunt, L. T., Orcutt, B. C., George, D. G., Yeh, L. S., Chen, H. R., Blomquist, M. C.,

Johnson, G. C., Seibel-Ross, E. I., and Ledky, R. S., 1984, *Report of the National Biomedical Research Foundation,* Georgetown University, Washington.

Barkovsky, E. V., 1982, Prediction of the secondary structure of globular proteins by their amino acid sequence, *Acta Biol. Med. Germ.* **41:**751–758.

Barkovsky, E. V., and Bandarin, V. A., 1979, Secondary structure prediction of globular proteins from their amino acid sequence, *Bioorg. Khim.* **5:**24–34.

Barlow, D. J., and Thornton, J. M., 1983, Ion-pairs in proteins, *J. Mol. Biol.* **168:**867–885.

Barth, R. K., Kim, B. S., Lan, N. C., Hunkapiller, T., Sobieck, N., Winoto, A., Gershenfeld, H., Okada, C., Hansburg, D., Weissman, I. L., and Hood, L., 1985, The murine T-cell receptor uses a limited repertoire of expressed V_β gene segments, *Nature* **316:**517–523.

Barton, G. J., and Sternberg, M. J. E., 1987a, Evaluation and improvements in the automatic alignment of protein sequences, *Protein Eng.* **1:**89–94.

Barton, G. J., and Sternberg, M. J. E., 1987b, A strategy for the rapid multiple alignment of protein sequences. Confidence levels from tertiary structure comparison, *J. Mol. Biol.* **198:**327–337.

Bash, P. A., Singh, V. C., Langridge, R., and Kollman, P. A., 1987, Free energy calculations by computer simulation, *Science* **236:**564–568.

Bashford, D., Chothia, C., and Lesk, A. M., 1987, Determinants of a protein fold. Unique features of the globin amino acid sequences, *J. Mol. Biol.* **196:**199–216.

Bazan, J. F., Fletterick, R. J., McKinley, M. P., and Pruisner, S. B., 1987, Predicted secondary structure and membrane topology of the scrapie prion protein, *Protein Eng.* **1:**125–135.

Beasty, A. M., and Matthews, C. R., 1985, Characterization of an early intermediate in the folding of the α-subunit of tryptophan synthase by a hydrogen exchange measurement, *Biochemistry* **24:**3547–3553.

Beecher, B., and Cassim, J. Y., 1976, Effects of light adaption on the purple membrane of *Halobacterium halobium, Biophys. J.* **16:**1183–1200.

Beeley, J. G., 1976, Location of the carbohydrate groups on ovomucoid, *Biochem. J.* **159:**335–345.

Beeley, J. G., 1977, Peptide chain conformation and the glycosylation of glycoproteins, *Biochem. Biophys. Res. Commun.* **76:**1051–1055.

Bentley, G., Dodson, E., Dodson, G., Hodgkin, D., and Mercola, D., 1976, Structure of insulin in 4-zinc insulin, *Nature* **261:**166–168.

Billeter, M., Havel, T. F., and Kuntz, I. D., 1987a, A new approach to the problem of docking two molecules: The ellipsoid algorithm, *Biopolymers* **26:**777–793.

Billeter, M., Havel, T. F., and Wuthrich, K., 1987b, The ellipsoid algorithm as a method for the determination of polypeptide conformations from experimental distance constraints and energy minimization, *J. Comput. Chem.* **8:**132–141.

Bird, C. R., Koller, B., Auffret, A. D., Huttly, A. K., Howe, C. J., Dyer, T. A., and Gray, J. C., 1985, The wheat chloroplast gene for CF_0 subunit I of ATP synthase contains a large intron, *EMBO J.* **4:**1381–1388.

Black, S. D., and Glorioso, J. C., 1986, MSEQ: A microcomputer-based approach to the analysis, display, and prediction of protein structure, *Biol. Techniques* **4:**448–460.

Blagdon, D. E., and Goodman, G., 1975, Mechanisms of protein and polypeptide helix initiation, *Biopolymers* **14:**241–245.

Blake, C. C. F., 1979, Exons encode protein functional units, *Nature* **277:**598.

Bloomer, A. C., Champness, J. N., Bricogne, G., Staden, R., and Klug, A., 1978, Protein disk of tobacco mosaic virus at 2.8 Å resolution showing the interactions within and between units, *Nature* **276:**362–368.

Blout, E. R., 1962, The dependence of the conformation of polypeptides and proteins upon amino acid composition, in: *Polyamino Acids, Polypeptides, and Proteins* (M. Stahman, ed.), University of Wisconsin Press, Madison, pp. 275–279.

Blow, D. M., Irwin, M. J., and Nyborg, J., 1977, The peptide chain of tyrosyl t-RNA synthetase: No evidence for a super-secondary structure of four-α-helices, *Biochem. Biophys. Res. Commun.* **76:**728–734.

Blundell, T. L., and Johnson, L. N., 1976, *Protein Crystallography,* Academic Press, New York.

Blundell, T. L., and Sternberg, M. J. E., 1985, Computer-aided design in protein engineering, *Trends Biotechnol.* **3:**228–235.

Blundell, T., Singh, J., Thornton, J., Burley, S. K., and Petsko, G. A., 1986a, Aromatic interactions, *Science* **234:**1005.

Blundell, T. L., Barlow, D., Sibanda, B. L., Thornton, T. M., Taylor, W., Tickle, I. J., Sternberg, M. J. E., Pitts, J. E., Haneef, I., and Hemmings, A. M., 1986b, Three-dimensional aspects of the design of new protein molecules, *Phil. Trans. R. Soc. Lond. [A]* **317:**333–344.

Blundell, T. L., Sibanda, B. L., Sternberg, M. J. E., and Thornton, J. M., 1987, Knowledge-based prediction of protein structures and the design of novel molecules, *Nature* **326**:347–352.

Bonner, T. I., Buckley, N. J., Young, A. C., and Brann, M. R., 1987, Identification of a family of muscarinic acetylcholine receptor genes, *Science* **237**:527–532.

Boswell, D. R., and McLachlan, A. D., 1984, Sequence comparison by exponentially-damped alignment, *Nucleic Acids Res.* **12**:457–464.

Bourgeois, S., and Pfahl, M., 1976, Repressors, *Adv. Protein Chem.* **30**:1–99.

Brandhuber, B. J., Boone, T., Kenney, W. C., and McKay, D. B., 1987, Three-dimensional structure of interleukin-2, *Science* **283**:1707–1709.

Brandl, C., Green, N. M., Korczak, B., and MacLennan, D. H., 1986, Two Ca^{2+} ATPase genes: Homologies and mechanistic implications of deduced amino acid sequences, *Cell* **44**:597–607.

Briggs, M. S., and Gierasch, L. M., 1984, Exploring the conformational role of signal sequences: Synthesis and conformational analysis of λ receptor protein wild type and mutant signal peptides, *Biochemistry* **23**:3111–3114.

Brisson, A., and Unwin, P. N. T., 1985, Quaternary structure of the acetylcholine receptor, *Nature* **315**:474–477.

Brooks, B. R., Bruccoleri, R. E., Olafson, B. D., States, D. J., Swaminathan, S., and Karplus, M., 1983, CHARMM: A program for macromolecular energy minimization and dynamics calculations, *J. Comput. Chem.* **4**:187–217.

Bruccoleri, R. E., and Karplus, M., 1987, Prediction of the folding of short polypeptide segments by uniform conformational sampling, *Biopolymers* **26**:137–168.

Bunting, J. R., Athey, T. W., and Cathou, R. E., 1972, Backbone folding of immunoglobulin light and heavy chains: A comparison of predicted β-bend positions, *Biochim. Biophys. Acta* **285**:60–71.

Burgess, A. W., and Scheraga, H. A., 1975, Assessment of some problems associated with the prediction of the three-dimensional structure of a protein from its amino acid sequence, *Proc. Nat. Acad. Sci. U.S.A.* **72**:1221–1225.

Burgess, A. W., Ponnuswamy, P. K., and Scheraga, H. A., 1974, Analysis of conformations of amino acid residues and prediction of backbone topography in proteins, *Israel J. Chem.* **12**:239–286.

Burley, S. K., and Petsko, G. A., 1985, Aromatic–aromatic interaction: A mechanism of protein stabilization, *Science* **229**:23–28.

Burley, S. K., and Petsko, G. A., 1986, Amino–aromatic interactions in proteins, *FEBS Lett.* **203**:139–143.

Busetta, B., and Hospital, M., 1981, Improving the accuracy of secondary structure predictions, *Biochemie* **63**:951–954.

Busetta, B., and Hospital, M., 1982, An analysis of the prediction of secondary structures, *Biochim. Biophys. Acta* **701**:111–118.

Cantor, C. R., and Jukes, T. H., 1966, The repetition of homologous sequences in the polypeptide chains of certain cytochromes and globins, *Proc. Natl. Acad. Sci. U.S.A.* **56**:177–184.

Capaldi, R. A., and Vanderkooi, G., 1972, The low polarity of many membrane proteins, *Proc. Natl. Acad. Sci. U.S.A.* **69**:930–932.

Catterall, W. A., 1984, The molecular basis of neuronal excitability, *Science* **223**:653–661.

Chakravarty, P. K., Mathur, K. B., and Dhar, M. M., 1973, The synthesis of a decapeptide with glycosidase activity, *Experientia* **29**:786–788.

Chang, E. L., Yager, P., Williams, R. W., and Dalziel, A. W., 1983, The secondary structure of reconstituted acetylcholine receptors as determined by Raman spectroscopy, *Biophys. J.* **41**:65a.

Charton, M., and Charton, B. I., 1983, The dependence of the Chou–Fasman parameters on amino acid side-chain structure, *J. Theor. Biol.* **102**:121–134.

Chothia, C., 1973, Conformation of twisted β-pleated sheets in proteins, *J. Mol. Biol.* **75**:295–302.

Chothia, C., 1974, Hydrophobic bonding and accessible surface area in proteins, *Nature* **248**:338–339.

Chothia, C., 1975, Structural invariants in protein folding, *Nature* **254**:303–308.

Chothia, C., 1976, The nature of accessible and buried surfaces in proteins, *J. Mol. Biol.* **105**:1–14.

Chothia, C., 1984, Principles that determine the structure of proteins, *Annu. Rev. Biochem.* **53**:537–572.

Chothia, C., and Janin, J., 1975, Principles of protein-protein recognition, *Nature* **256**:705–708.

Chothia, C., and Janin, J., 1982, Orthogonal packing of β-sheets in proteins, *Biochemistry* **21**:3955–3965.

Chothia, C., and Lesk, A. M., 1982a, Evolution of proteins formed by β-sheets. I. The core of the immunoglobulin domains, *J. Mol. Biol.* **160**:325–342.

Chothia, C., and Lesk, A. M., 1982b, Evolution of proteins formed by β-sheets. II. Plastocyanin and azurin, *J. Mol. Biol.* **160:**303–323.

Chothia, C., and Lesk, A. M., 1986, The relation between the divergence of sequence and structure in proteins, *EMBO J.* **5:**823–826.

Chothia, C., Levitt, M., and Richardson, D., 1977, Structure of proteins: Packing of α-helices and pleated sheets, *Proc. Natl. Acad. Sci. U.S.A.* **74:**4130–4134.

Chothia, C., Levitt, M., and Richardson, D., 1981, Helix to helix packings in proteins, *J. Mol. Biol.* **145:**215–250.

Chothia, C., Novotny, J., Bruccoleri, R., and Karplus, M., 1985, Domain association in immunoglobulin molecules. the packing of variable domains, *J. Mol. Biol.* **186:**651–663.

Chou, P. Y., 1979, New approaches to protein structural analysis and conformational predictions, in: *CECM Protein Folding Workshop,* Université de Paris-Sud, Orsay, France.

Chou, P. Y., 1980, Amino acid compositions of four structural classes of proteins, in: *Abstracts, Second Chemical Congress of the North American Continent,* Las Vegas.

Chou, P. Y., and Fasman, G. D., 1973, Structural and functional role of Leu residues in proteins, *J. Mol. Biol.* **74:**263–281.

Chou, P. Y., and Fasman, G. D., 1974a, Conformational parameters for amino acids in helical, β-sheet, and random coil regions calculated from proteins, *Biochemistry* **13:**211–222.

Chou, P. Y., and Fasman, G. D., 1974b, Prediction of protein conformation, *Biochemistry* **13:**222–245.

Chou, P. Y., and Fasman, G. D., 1975, The conformation of glucagon: Predictions and consequences, *Biochemistry* **14:**2536–2541.

Chou, P. Y., and Fasman, G. D., 1977a, Secondary structural prediction of proteins from their amino acid sequence, *Trends Biochem. Sci.* **2:**128–132.

Chou, P. Y., and Fasman, G. D., 1977b, β-Turns in proteins, *J. Mol. Biol.* **115:**135–175.

Chou, P. Y., and Fasman, G. D., 1977c, Prediction of protein secondary structure, in: *Fifth American Peptide Symposium* (M. Goodman and J. Meienhofer, eds.), John Wiley & Sons, New York, pp. 284–287.

Chou, P. Y., and Fasman, G. D., 1978a, Prediction of the secondary structure of proteins from their amino acid sequence, *Adv. Enzymol.* **47:**45–148.

Chou, P. Y., and Fasman, G. D., 1978b, Empirical predictions of protein conformation, *Annu. Rev. Biochem.* **47:**251–276.

Chou, P. Y., and Fasman, G. D., 1979a, Prediction of β-turns, *Biophys. J.* **26:**367–384.

Chou, P. Y., and Fasman, G. D., 1979b, Conservation of chain reversal regions in proteins, *Biophys. J.* **26:** 385–400.

Chou, P. Y., Adler, A. J., and Fasman, G. D., 1975, Conformational prediction and circular dichroism studies on the *lac* repressor, *J. Mol. Biol.* **96:**29–45.

Chou, K.-C., Pottle, M., Nemethy, G., Veda, Y., and Scheraga, H. A., 1982, Structure of β-sheets, *J. Mol. Biol.* **162:**89–112.

Claudio, T., Ballivet, M. Patrick, J., and Heinemann, S., 1983, Nucleotide and deduced amino acid sequences of *Torpedo californica* acetylcholine receptor γ-subunit, *Proc. Natl. Acad. Sci. U.S.A.* **80:**1111–1115.

Cohen, C., and Parry, D. A. D., 1986, α-Helical coiled-coils: A widespread motif in proteins, *Trends Biochem. Sci.* **11:**245–248.

Cohen, F. E., and Kuntz, I. D., 1987, Prediction of the three-dimensional structure of human growth hormone, *Proteins* **1:**162–166.

Cohen, F. E., and Sternberg, M. J. E., 1980a, The use of chemically derived distant constants in the prediction of protein structure with myoglobin as an example, *J. Mol. Biol.* **137:**9–22.

Cohen, F. E., and Sternberg, M. J. E., 1980b, On the prediction of protein structure: The significance of the root-mean-square deviation, *J. Mol. Biol.* **138:**321–333.

Cohen, F. E., Richmond, J. T., and Richards, F. M. J., 1979, Protein folding: Evaluation of some simple rules for the assembly of helices into tertiary structure with myoglobin as an example, *J. Mol. Biol.* **132:**275–288.

Cohen, F. E., Sternberg, M. J. E., and Taylor, W. R., 1980, Analysis and prediction of protein β-sheet structures by a combinatorial approach, *Nature* **285:**378–382.

Cohen, F. E., Sternberg, M. J. E., and Taylor, W. R., 1981, Analysis of the tertiary structure of protein sandwiches, *J. Mol. Biol.* **148:**253–272.

Cohen, F. E., Sternberg, M. J. E., and Taylor, W. R., 1982, Analysis and prediction of the packing of α-helices against a β-sheet in the tertiary structure of globular proteins, *J. Mol. Biol.* **156:**821–862.

Cohen, F. E., Abarbanel, R. M., Kuntz, I. D., and Fletterick, R. J., 1983, Secondary structure assignment for α/β proteins by a combinatorial approach, *Biochemistry* **22**:4894–4904.

Cohen, F. E., Abarbanel, R. M., Kuntz, I. D., and Fletterick, R. J., 1986a, Turn prediction in proteins using a pattern matching approach, *Biochemistry* **25**:266–275.

Cohen, F. E., Kosen, P. A., Kuntz, I. D., Epstein, C. B., Cardelli, T. C., and Smith, K. A., 1986b, Structure activity studies of interleukin-2, *Science* **234**:349–356.

Conboy, J., Kan, Y. W., Shohet, S. B., and Mohandas, N., 1986, Molecular cloning of protein 4.1, a major structural element of the human erythrocyte membrane skeleton, *Proc. Natl. Acad. Sci. U.S.A.* **83**:9512–9516.

Cook, D. A., 1967, The relation between amino acid sequence and protein conformation, *J. Mol. Biol.* **29**:167–171.

Cornette, J. L., Cease, K. B., Margalit, J. H., Spouge, J. L., Berzofsky, J. A., and DeLisi, C., 1987, Hydrophobicity scales and computational techniques for detecting amphipathic structures in proteins, *J. Mol. Biol.* **195**:659–685.

Corrigan, A. J., and Huang, P. C., 1982, A BASIC microcomputer program for plotting the secondary structure of proteins, *Comput. Prog. Biomed.* **15**:163–168.

Crawford, I. P., Niermann, T., and Kirshner, K., 1987, Predictions of secondary structure by evolutionary comparison: Application to the α-subunit of tryptophan synthase, *Proteins* **1**:118–129.

Crawford, J. L., Lipscomb, W. N., and Schellman, C. G., 1973, The reverse turn as a polypeptide conformation in globular proteins, *Proc. Natl. Acad. Sci. U.S.A.* **70**:538–542.

Creighton, T. E., 1978, Experimental studies of protein folding and unfolding, *Prog. Biophys. Mol. Biol.* **33**: 231–297.

Creighton, T. E., 1979, Electrophoretic analysis of the unfolding of proteins by urea, *J. Mol. Biol.* **129**:235–264.

Criado, M., Hochschwender, S., Sarin, V., Fox, J. L., and Lindstrom, J., 1985, Evidence for unpredicted transmembrane domains in acetylcholine receptor subunits, *Proc. Nat. Acad. Sci. U.S.A.* **82**:2004–2008.

Crick, F. H. C., 1953, The packing of α-helices: Simple coiled coils, *Acta Crystallogr.* **6**:689–697.

Crippen, G. M., 1977a, A statistical approach to the calculation of conformation of proteins. 1. Theory, *Macromolecules* **10**:21–25.

Crippen, G. M., 1977b, A statistical approach to the calculation of conformation of proteins. 2. The reoxidation of reduced trypsin inhibitor, *Macromolecules* **10**:25–28.

Crippen, G., M., 1977c, A novel approach to the calculation of conformation: Distance geometry, *J. Comp. Physiol.* **26**:449–452.

Crippen, G. M., 1978, The tree structural organization of proteins, *J. Mol. Biol.* **126**:315–332.

Crippen, G. M., and Kuntz, I. D., 1978, A survey of atom packing in globular proteins, *Int. J. Peptide Protein Res.* **12**:47–56.

Davies, B. D., and Tai, R. C., 1980, The mechanism of protein secretion across membranes, *Nature* **283**:433–438.

Davies, D. R., 1964, A correlation between amino acid composition and protein structure *J. Mol. Biol.* **9**:605–609.

Dayhoff, M. O., 1972, *Atlas of Protein Sequence and Structure*, National Biomedical Research Foundation, Washington.

Dayhoff, M. O., 1978, *Atlas of Protein Sequence and Structure. Supplement 3*, National Biomedical Research Foundation, Washington.

Dayhoff, M. O., Barker, W. C., and Hunt, L. T., 1983, Establishing homologies in protein sequences, *Methods Enzymol.* **91**:524–545.

DeGrado, W. F., Kezdy, E. J., and Kaiser, E. T., 1981, Design, synthesis and characterization of a cytotoxic peptide with melittin-like activity, *J. Am. Chem. Soc.* **103**:679–681.

de Groot, R. J., Luytjes, W., Horzinek, M. C., van der Zeijst, B. A. M., Spaan, W. J. M., and Lenstra, J. A., 1987, Evidence for a coiled-coil structure in spike proteins of coronaviruses, *J. Mol. Biol.* **196**:963–966.

Deisenhofer, J., Epp, O., Mikki, K., Huber, R., and Michel, H., 1984, X-ray structural analysis of a membrane protein complex electron density map at 3 Å resolution and a model of the chromophores of the photosynthetic reaction centre from *Rhodopseudomonas viridis*, *J. Mol. Biol.* **180**:385–398.

Deisenhofer, J., Epp, O., Mikki, K., Huber, R., and Michel, H., 1985, Structure of the protein subunits in the photoreaction centre of *Rhodopseudomonas viridis* at 3Å resolution, *Nature* **318**:618–624.

Deléage, G., and Roux, B., 1987, An algorithm for protein secondary structure prediction based on class prediction, *Protein Eng.* **1**:289–294.

Deléage, G., Tinland, B., and Roux, B., 1987, A computerized version of the Chou and Fasman method for predicting the secondary structure of proteins, *Anal. Biochem.* **165**:200–207.

Dencher, N. A., and Heyn, M. P., 1978, Formation and properties of bacteriorhodopsin monomers in the non-ionic detergents octyl-β-glucoside and Triton X-100, *FEBS Lett.* **96**:322–396.

DeSantis, P., Giglio, E., Liquori, A. M., and Ripamonti, A., 1965, Van der Waals interaction and the stability of helical polypeptide chains, *Nature* **206**:456–458.

DesJarlais, R. L., Sheridan, R. P., Dixon, J. S., Kuntz, I. D., and Venkataraghavan, R., 1986, Docking flexible ligands to macromolecular receptors by molecular shape, *J. Med. Chem.* **29**:2149–2153.

Devereaux, J., Haeben, P., and Smithies, O., 1984, A comprehensive set of sequence analysis programs for the VAX, *Nucleic Acids Res.* **17**:387–395.

Devillers-Thiery, A., Giraudet, J., Bentaboulet, M., and Changeux, J.-P., 1983, Complete mRNA coding sequence of the acetylcholine binding α-subunit of *Torpedo marmorata* acetylcholine receptor: A model for transmembrane organization of the polypeptide chain, *Proc. Natl. Acad. Sci. U.S.A.* **80**: 2067–2071.

Dickerson, R. E., Takano, T., Eisenberg, D., Kallai, O. B., Samson, L., and Cooper, A., 1971, Ferricytochrome c: General features of the horse and bonito proteins at 2.8 Å resolution, *J. Biol. Chem.* **246**: 1511–1535.

Dill, K. A., 1985, Theory for the folding and stability of globular proteins, *Biochemistry* **24**:1501–1509.

Dixon, R. A., Kobilka, B. K., Strader, D. J., Benovic, J. L., Dohlman, H. G., Frielle, T., Bolanowski, M. A., Bennet, C. D., Rands, E., Diehl, R. E., Mumford, R. A., Slater, E. E., Sigal, I. S., Caron, M. G., Lefkowitz, R. J., and Strader, C. D., 1986, Cloning of the gene and cDNA mammalian β-adrenergic receptor and homology with rhodopsin, *Nature* **321**:75–79.

Dohlman, H. G., Caron, M. G., and Lefkowitz, R. L., 1987, A family of receptors coupled to guanine nucleotide regulatory proteins, *Biochemistry* **26**:2657–2664.

Drexler, K. E., 1980, Molecular engineering: An approach to the development of general capabilities for molecular manipulation, *Proc. Natl. Acad. Sci. U.S.A.* **78**:5275–5278.

Dufton, M. J., and Hider, R. C., 1977, Snake toxin secondary structure predictions. Structure activity relationships, *J. Mol. Biol.* **115**:177–193.

Duncan, T. M., Parsonage, D., and Senior, A. E., 1986, Structure of the nucleotide-binding domain in the β-subunit of *Escherichia coli* F_1-ATPase, *FEBS Lett.* **208**:1–6.

Dunhill, P., 1968, The use of helical net-diagrams to represent protein structures, *Biophys. J.* **8**:865–875.

Ecker, J. G., and Kupferschmid, M., 1982, *Report OR,* Rensselaer Polytechnic, Troy, NY.

Edelman, G. M., Cunningham, B. A., Reeke, G. N., Jr., Becker, J. W., Waxdall, M. J., and Wang, J. L., 1972, The covalent and three-dimensional structure of concanavalin A, *Proc. Natl. Acad. Sci. U.S.A.* **69**: 2580–2584.

Edman, J. C., Ellis, L., Blacher, R. W., Roth, R. A., and Rutter, W. J., 1985, Sequence of protein disulphide isomerase and implications of its relation to thioredoxin, *Nature* **317**:267–270.

Edmonds, D. T., 1985, The α-helix dipole in membranes: A new gating mechanism for ion channels, *Eur. Biophys. J.* **13**:31–35.

Edsall, J. Y., and McKenzie, H. A., 1983, Water and proteins II. The location and dynamics of water in protein systems and its relation to their stability and properties, *Adv. Biophysics* **16**:53–183.

Edwards, M. S., Sternberg, M. J. E., and Thornton, J. M., 1987, Structure and sequence patterns in the loops of βαβ units, *Protein Eng.* **1**:173–181.

Efimov, A. V., 1977, Stereochemistry of the packing of α-helices and the β-structure in a compact globule, *Dokl. Akad. Nauk SSSR* **235**:699–702.

Efimov, A. V., 1979, Packing of α-helices in globular proteins. Layer-structure of globular hydrophobic cores, *J. Mol. Biol.* **134**:23–46.

Efimov, A. V., 1982a, Role of constrictions in formation of protein structures containing four helical regions, *Mol. Biol.* **16**:271–281.

Efimov, A. V., 1982b, Super-secondary structures of β-proteins, *Mol. Biol.* **16**:799–806.

Efimov, A. V., 1984, A novel super-secondary structure of proteins and the relation between the structure and amino acid sequence, *FEBS Lett.* **166**:33–38.

Efimov, A. V., 1985, Standard conformations of polypeptide chains in irregular regions of proteins, *Mol. Biol.* **20**:350–360.

Efimov, A. V., 1986a, Standard structures in protein molecules. I. α–β Hairpins, *Mol. Biol.* **20**:329–339.

Efimov, A. V., 1986b, Standard structures in protein molecules. II. β–α Hairpins, *Mol. Biol.* **20:**340–345.

Eisenberg, D., 1984, Three-dimensional structure of membrane surface proteins, *Annu. Rev. Biochem.* **53:**595–623.

Eisenberg, D., Weiss, R. M., Terwilliger, T. C., and Wilcox, W., 1982a, Hydrophobic moments and protein structure, *Faraday Symp. Chem. Soc.* **17:**109–120.

Eisenberg, D., Weiss, R. M., and Terwilliger, T. C., 1982b, The helical hydrophobic moment: A measure of the amphilicity of a helix, *Nature* **299:**371–374.

Eisenberg, D., Schwartz, E., Komaromy, M., and Wall, R., 1984a, Analysis of membrane and surface protein sequences with the hydrophobic moment plot, *J. Mol. Biol.* **179:**125–142.

Eisenberg, D., Weiss, R. M., and Terwilliger, T. C., 1984b, The hydrophobic moment detects periodicity in protein hydrophobicity, *Proc. Natl. Acad. Sci. U.S.A.* **81:**140–144.

Elleman, T. C., Azad, A. A., and Ward, C. W., 1982, Neuraminidase gene from early Asian strain human influenza virus, A/RI/5^{-}/57(H2N2), *Nucleic Acids Res.* **10:**7005–7015.

Emr, S. D., and Silhavy, T. J., 1983, Importance of secondary structure in the signal sequence for protein secretion, *Proc. Natl. Acad. Sci. U.S.A.* **80:**4599–4603.

Engelman, D. M., and Steitz, T., 1981, The spontaneous insertion of proteins into and across membranes: The helical hairpin hypothesis, *Cell* **23:**411–422.

Engelman, D. M., and Steitz, T. A., 1984, On the folding and insertion of globular membrane proteins, in: *The Protein Folding Problem* (D. Wetlaufer, ed.), Westview Press, Boulder, CO, pp. 87–113.

Engelman, D. M., and Zaccai, G., 1980, Bacteriorhopdopsin is an inside-out protein, *Proc. Natl. Acad. Sci. U.S.A.* **77:**5894–5898.

Engelman, D. M., Henderson, R., McLachlan, A. D., and Wallace, B. A., 1980, Path of the polypeptide in bacteriorhodopsin, *Proc. Natl. Acad. Sci. U.S.A.* **77:**2023–2027.

Engelman, D. M., Goldman, A., and Steitz, T., 1986, The identification of helical segments in the polypeptide chain of bacteriorhodopsin, *Methods Enzymol.* **88:**81–88.

Engelman, D. M., Steitz, T. A., and Goldman, A., 1986, Identifying non-polar transbilayer helices in amino acid sequences of membrane proteins, *Annu. Rev. Biophys. Biophys. Chem.* **15:**321–353.

Epand, R. M., 1971, Studies on the conformation of glucagon, *Can. J. Biol. Chem.* **49:**166–169.

Erni, B., and Zanolari, B., 1986, Glucose-permease of the bacterial phosphotransferase system, *J. Biol. Chem.* **261:**16398–16403.

Esipova, N. G., and Tumanyan, V. G., 1972, Factors determining the formation of the tertiary structure of globular protein, *Mol. Biol.* **6:**840–850.

Fasman, G. D., 1980, Prediction of protein conformation from the primary structure, *Ann. N.Y. Acad. Sci.* **348:** 147–159.

Fasman, G. D., 1982, Prediction of the secondary structure of proteins, in: *From Cyclotrons to Cytochromes. Essays in Molecular Biology and Chemistry* (N. O. Kaplan and A. Robinson, eds.), Academic Press, New York, pp. 455–468.

Fasman, G. D., 1985, A critique of the utility of the prediction of protein secondary structure. International Symposium on Biomolecular Structure and Interactions, *J. Biosci.* **8:**15–23.

Fasman, G. D., 1987, The road from poly-α-amino acids to the prediction of protein conformation. Biopolymers and biotechnology symposium in honor of Prof. Ephraim Katzir on his 70th birthday, *Biopolymers* **26:** 559–579.

Fasman, G. D., and Chou, P. Y., 1974, Prediction of protein conformation: Consequences and aspirations, in: *Peptides, Polypeptides, and Proteins* (E. R. Blout, F. A. Bovey, M. Goodman, and N. Lotan, eds.), John Wiley & Sons, New York, pp. 114–125.

Fasman, G. D., Chou, P. Y., and Adler, A. J., 1976, Prediction of the conformation of the histones, *Biophys. J.* **16:**1201–1238.

Fasman, G. D., Chou, P. Y., and Adler, A. J., 1977, Histone conformation: Predictions and experimental studies, in: *The Molecular Biology of the Mammalian Genetic Apparatus*—I, (P. O. P. Ts'o, ed.), Elsevier–Exerpta Medica/North Holland, Amsterdam, pp. 1–52.

Finer-Moore, J., and Stroud, R. M., 1984, Amphipathic analysis and possible formation of the ion channel in acetylcholine receptor, *Proc. Natl. Acad. Sci. U.S.A.* **81:**155–159.

Finkelstein, A. V., Ptitsyn, O. B., and Bendsko, P., (1970), Coiling and topology on the anti-parallel β-structure, *Biofisika* **24:**21–26.

Finney, J. L., Gellatly, B. J., Golton, I. C., and Goodfellow, J., 1980, Solvent effects and polar interactions in the structural stability and dynamics of globular proteins, *Biophys. J.* **32:**17–23.

Fishleigh, R. V., Robson, B., Garnier, J., and Finn, P. W., 1987, Studies on rationales for an expert system

approach to the interpretation of protein sequence data. Preliminary analysis of the human epidermal growth factor receptor, *FEBS Lett.* **214:**219–225.

Fitch, W. M., 1966a, The relation between frequencies of amino acids and ordered trinucleotides, *J. Mol. Biol.* **16:**1–8.

Fitch, W. M., 1966b, An improved method of testing for evolutionary homology, *J. Mol. Biol.* **16:**9–16.

Fitch, W. M., 1966c, Evidence suggesting a partial, internal duplication in the ancestral gene for heme-containing globins, *J. Mol. Biol.* **16:**17–27.

Fitch, W. M., and Smith, T. F., 1983, Optimal sequence alignments, *Proc. Natl. Acad. Sci. U.S.A.* **80:**1382–1386.

Fleming, P. J., Dailey, H. A., Corcoran, D., and Strittmatter, P., 1978, The primary structure of the non-polar segment of bovine cytochrome b_5, *J. Biol. Chem.* **253:**5369–5372.

Flinta, C., von Heijne, G., and Johansson, J., 1983, Helical sidedness and the distribution of polar residues in trans-membrane helices, *J. Mol. Biol.* **168:**193–196.

Foster, D. L., Boublik, M., and Kaback, H. R., 1983, Structure of the *lac* carrier protein of *Escherichia coli*, *J. Biol. Chem.* **258:**31–34.

Froimowitz, M., and Fasman, G. D., 1974, Prediction of secondary structure of proteins using the helix-coil transition theory, *Macromolecules* **7:**583–589.

Fukushima, D. Kupferberg, J. P., Yokoyama, S., Kroon, D. J., Kaiser, E. T., and Kezdy, F. J., 1979, A synthetic amphiphilic helical docosapeptide with the surface properties of plasma apolipoprotein A-I, *J. Am. Chem. Soc.* **101:**3703–3704.

Furois-Corbin, S., and Pullman, A., 1987, Theoretical studies of the packing of α-helices into possible transmembrane bundles: Sequences including alanines, leucines and serines, *Biochim. Biophys. Acta* **902:** 31–45.

Garnier, J., Osguthorpe, D. J., and Robson, B., 1978, Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins, *J. Mol. Biol.* **120:**97–120.

Garratt, R. C., Taylor, W. R., and Thornton, J. M., 1985, The influence of tertiary structure on secondary structure prediction. Accessibility versus predictability for β-structure, *FEBS Lett.* **188:**59–62.

Geisow, M. J., and Roberts, R. D. B., 1980, Amino acid preferences for secondary structure vary with protein class, *Int. J. Biol. Macromol.* **2:**387–389.

Geisow, M. J., Fritsche, U., Hexham, J. M., Dash, B., and Johnson, T., 1986, A consensus amino acid sequence repeat in *Torpedo* and mammalian Ca^{2+}-dependent membrane-binding proteins, *Nature* **320:** 636–638.

Gelin, B., and Karplus, M., 1979, Side-chain torsional potentials: Effect of dipeptide, protein, and solvent environment, *Biochemistry* **18:**1256–1268.

Gerber, G. E., Anderegg, R. J., Herlihy, W. C., Gray, C. P., Bieman, K., and Khorana, H. G., 1979, Partial primary structure of bacteriorhodopsin: Sequencing methods for membrane proteins, *Proc. Natl. Acad. Sci. U.S.A.* **76:**227–231.

Getzoff, E. D., Tainer, J. A., and Olson, A. J., 1986, Recognition and interactions controlling the assemblies of β-barrel domains, *Biophys. J.* **49:**191–206.

Ghelis, C., and Yon, J., 1982, *Protein Folding*, Academic Press, New York.

Gibrat, J.-F., Garnier, J., and Robson, B., 1987, Further developments of protein secondary structure prediction using information theory: New parameters and consideration of residue pairs, *J. Mol. Biol.* **198:**425–443.

Gilbert, W., 1978, Why genes in pieces? *Nature* **271:**501.

Gilson, M. K., and Honig, B. H., 1987, Calculation of electrostatic potentials in an enzyme active site, *Nature* **330:**84–86.

Glaeser, R. M., and Jap, B. K., 1985, Absorption flattening in the circular dichroism spectra of small membrane fragments, *Biochemistry* **24:**6398–6401.

Goad, W. B., and Kanehisa, M. I., 1982, Pattern recognition in nucleic acid sequences I. A general method for finding local homologies and symmetries, *Nucleic Acids Res.* **10:**247–263.

Gordon, D. J., and Holzworth, G., 1971, Artifacts in the measure of optical activity of membrane suspensions, *Arch. Biochem. Biophys.* **142:**481–488.

Grantham, R., 1974, Amino acid difference formula to help explain protein evolution, *Science* **185:** 862–864.

Gratzer, W. B., Bailey, E., and Beaven, G. H., 1967, Conformational states of glucagon, *Biochem. Biophys. Res. Commun.* **28:**914–919.

Gray, T. M., and Matthews, B. W., 1984, Intrahelical hydrogen bonding of serine, threonine, and cysteine residues within α-helices and its relevance to membrane-bound proteins, *J. Mol. Biol.* **175:**75–81.

Green, N. M., and Flanagan, M. T., 1976, The prediction of the conformation of membrane proteins from the sequence of amino acids, *Biochem. J.* **153:**729–732.

Grenningloh, G., Rienitz, A., Schmitt, B., Methfessel, C., Zensen, M., Beyreuther, K., Grundelfinger, E. D., and Betz, H., 1987, The strychnine-binding subunit of the glycine receptor shows homology with nicotinic acetylcholine receptors, *Nature* **328:**215–220.

Gribskov, M., Burgess, R. R., and Devereaux, J., 1986, PEPPLOT, a protein secondary analysis program for the UWGCG sequence analysis software package, *Nucleic Acids Res.* **14:**327–334.

Gribskov, M., McLachlan, A. D., and Eisenberg, D., 1987, Profile analysis: Detection of distantly related proteins, *Proc. Natl. Acad. Sci. U.S.A.* **84:**4355–4358.

Gutte, B., Daumigen, M., and Wittschieber, E., 1979, Design, synthesis and characteristics of a 34-residue polypeptide that interacts with nucleic acids, *Nature* **281:**650–655.

Guy, H. R., 1981, Structural models of the nicotinic acetylcholine receptor and its toxin-binding sites, *Cell. Mol. Neurobiol.* **1:**231–258.

Guy, H. R., 1984, A structural model of the acetylcholine receptor channel based on partition energy and helix packing calculations, *Biophys. J.* **45:**249–261.

Guy, H. R., 1985, Amino acid side-chain partition energies and distribution of residues in soluble proteins, *Biophys. J.* **47:**61–70.

Guy, H. R., and Seetharamulu, P., 1986, Molecular model of the action potential sodium channel, *Proc. Natl. Acad. Sci. U.S.A.* **83:**508–512.

Guzzo, A. V., 1965, The influence of amino acid sequence on protein structure, *Biophys. J.* **5:**809–822.

Haber, J. E., and Koshland, Jr., D. E., 1970, An evaluation of the relatedness of proteins based on comparison of amino acid sequences, *J. Mol. Biol.* **50:**617–639.

Hager, K. M., Mandala, S. M., Davenport, J. W., Speicher, D. W., Benz, Jr., E. J., and Slayman, C. W., 1986, Amino acid sequences of the plasma membrane ATPase of *Neurospora crassa:* Deduction from genomic and c-DNA sequences, *Proc. Natl. Acad. Sci. U.S.A.* **83:**7693–7697.

Hagler, A. T., and Honig, B., 1978, On the formation of protein tertiary structure on a computer. *Proc. Natl. Acad. Sci. U.S.A.* **75:**554–558.

Hagler, A. T., Huler, E., and Lifson, S., 1974, Energy functions for peptides and proteins. I. Derivation of a consistent force field including the hydrogen bond from amide crystals, *J. Am. Chem. Soc.* **96:**5319–5327.

Hammett, L. P., 1970, *Physical Organic Chemistry,* 2nd ed., McGraw-Hill, New York.

Hardman, K. D., and Ainsworth, C. F., 1972, Structure of concanavalin A at 2.4 Å resolution, *Biochemistry* **11:**4910–4919.

Harrison, S. C., 1985, Two for the price of one, *Nature* **313:**736–737.

Havel, T. F., Kuntz, I. D., and Crippen, G. M., 1983, The theory and practice of distance geometry, *Bull. Math. Biol.* **45:**665–720.

Hayes, T. G., 1980, Chou–Fasman analysis of the secondary structure of F and LE interferons, *Biochem. Biophys. Res. Commun.* **95:**872–879.

Hayward, S. B., and Stroud, R. M., 1981, Projected structure of purple membrane determined to 3.7 Å resolution by low temperature electron microscopy, *J. Mol. Biol.* **151:**491–517.

Heber-Katz, E., Hollosi, M., Dietzschold, B., Hudecz, F., and Fasman, G. D., 1985, The T cell response to the glycoprotein D of the herpes simplex virus: The significance of antigen conformation, *J. Immunol.* **135:** 1385–1390.

Hedrick, S. M., Cohen, D. I., Nielsen, E. A., and Davis, M. M., 1984a, Isolation of cDNA clones encoding T-cell-specific membrane associated proteins, *Nature* **308:**149–153.

Hedrick, S. M., Nielsen, E. A., Kavaler, J., Cohen, D. I., and Davis, M. M., 1984b, Sequence relationships between putative T-cell receptor polypeptides and immunoglobulins, *Nature* **308:**153–158.

Hellberg, S., Sjostrom, M., and Wold, S., 1986, The prediction of bradykinin potentiating potency of pentapeptides. An example of a peptide quantitative structure–activity relationship, *Acta Chem. Scand.* **B40:**135–140.

Henderson, R., and Unwin, P. N. T., 1975, Three-dimensional model of purple membrane obtained by electron microscopy, *Nature* **257:**28–32.

Hol, W. G. J., van Duijnen, P. T., and Berendsen, H. J. C., 1978, The α-helix dipole and the properties of proteins, *Nature* **273:**443–446.

Hol, W. G. J., Halie, L. M., and Sander, C., 1981, Dipoles of the α-helix and β-sheet: Their role in protein folding, *Nature* **294:**532–536.

Holmes, M. A., and Matthews, B. W., 1982, Structure of thermolysin refined at 1.6 Å resolution, *J. Mol. Biol.* **160:**623–639.

Hones, J., Jany, K.-D., Pfleider, G., and Wagner, A. F. V., 1987, An integrated prediction of secondary, tertiary, and quaternary structure of glucose dehydrogenase, *FEBS Lett.* **212:**193–198.

Honig, B. H., and Hubbell, W. L., 1984, Stability of "salt bridges" in membrane proteins, *Proc. Natl. Acad. Sci. U.S.A.* **81:**5412–5416.

Honig, B. H., Ray, A., and Levinthal, C., 1976, Conformational flexibility and protein folding: Rigid structural fragments connected by flexible joints in subtilisn BPN, *Proc. Natl. Acad. Sci. U.S.A.* **73:**1974–1978.

Honig, B. H., Hubbell, W. L., and Flewelling, R. F., 1986, Electrostatic interactions in membranes and proteins, *Annu. Rev. Biophys. Biophys. Chem.* **15:**163–193.

Hopp, T. P., and Woods, K. R., 1981, Prediction of protein antigenic determinants from amino acid sequence, *Proc. Natl. Acad. Sci. U.S.A.* **78:**3824–3828.

Hruby, W., Krstenansky, J., Gysin, B., Pelton, J. T., Trivedi, D., and McKee, R. L., 1986, Conformational considerations in the design of glucagon agonists and antagonists: Examination using synthetic analogs, *Biopolymers* **25:**S135–S155.

Huang, K.-S., Bayley, H., Liao, M.-J., London, E., and Khorana, H. G., 1981, Refolding of an integral membrane protein. Denaturation, renaturation, and reconstitution of intact bacteriorhodopsin and two proteolytic fragments, *J. Biol. Chem.* **256:**3802–3809.

Huang, K.-S., Radhakrishnan, R., Bayley, H., and Khorana, H. G., 1982, Orientation of retinal in bacteriorhodopsin as studied by cross-linking using a photosensitive analog of retinal, *J. Biol. Chem.* **257:** 13616–13623.

Huber, R., Kulka, D., Ruhlman, A., and Steigman, W., 1971, Pancreatic trypsin inhibitor (Kunitz) Part 1. Structure and function, *Cold Spring Harbor Symp. Quant. Biol.* **36:**141–150.

Hucho, F., 1986, The nicotinic acetylcholine receptor and its ion channel, *Eur. J. Biochem.* **158:**211–256.

Hurle, M. R., Matthews, C. R., Cohen, F. E., Kuntz, I. D., Toumadje, A., and Johnson, Jr., W. C., 1987, Prediction of the tertiary structure of the α-subunit of tryptophan synthetase, *Proteins* **2:**210–224.

IntelliGenetics, Inc., 1981–1985. *PER References Manual,* IntelliGenetics, Mountain View, CA.

Isogai, Y., Nemethy, G., Rackovsky, S., Leach, S. J., and Scheraga, H. A., 1980, Characterization of multiple bends in proteins, *Biopolymers* **19:**1183–1210.

Jaenicke, R., 1984, Protein folding and protein association, *Angew. Chem. [Engl.]* **23:**395–413.

Jaenicke, R., 1987, Folding and association of proteins, *Prog. Biophys. Mol. Biol.* **49:**117–237.

Janin, J., 1979, Surface and inside volumes in globular proteins, *Nature* **277:**491–492.

Janin, J., and Chothia, C., 1980, Packing of α-helices onto β-pleated sheets and anatomy of α/β proteins, *J. Mol. Biol.* **143:**95–128.

Jap, B. K., and Kong, S. H., 1986, Secondary structure of halorhodopsin, *Biochemistry* **25:**502–505.

Jap, B. K., Maestre, M. F., Hayward, S. B., and Glaeser, R. M., 1983, Peptide-chain secondary structure of bacteriorhodopsin, *Biophys. J.* **43:**81–89.

Jones, D. D., 1975, Amino acid properties and side-chain orientation in proteins: A cross correlation approach, *J. Theor. Biol.* **50:**167–183.

Kabat, E. A., and Wu, T. T., 1973a, The influence of nearest-neighbor amino acids on the conformation of the middle amino acid in proteins: Comparison of predicted and experimental determination of β-sheets in conconavalin A, *Proc. Natl. Acad. Sci. U.S.A.* **70:**1473–1477.

Kabat, E. A., and Wu, T. T., 1973b, The influence of nearest-neighboring amino acid residues on aspects of secondary structure of proteins. Attempts to locate α-helices and β-sheets, *Biopolymers* **12:**751–774.

Kabat, E. A., Wu, T. T., Bilofsky, H., Reid-Miller, M., and Perry, H., 1983, *Sequences of Immunological Interest,* U.S. Dept. of Health and Human Services, Washington.

Kabsch, W., and Sander, C., 1983a, How good are predictions of protein secondary structure? *FEBS Lett.* **155:** 179–182.

Kabsch, W., and Sander, C., 1983b, Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometric features, *Biopolymers* **22:**2577–2637.

Kabsch, W., and Sander, C., 1984, On the use of sequence homologies to predict protein structure: Identical pentapeptides can have completely different conformations, *Proc. Natl. Acad. Sci. U.S.A.* **81:**1075–1078.

Kabsch, W., and Sander, C., 1985, Identical pentapeptides with different backbones, *Nature* **317:**207.

Kanazawa, H., Hama, H., Rosen, B. P., and Futai, M., 1985, Deletion of seven amino acids from the γ subunit

of *Escherichia coli* H^+-ATPases causes total loss of F_1 assembly on membrane, *Arch. Biochem. Biophys.* **241:**364–370.

Karplus, M., and Weaver, D. L., 1979, Diffusion–collision model for protein folding, *Biopolymers* **18:**1421–1437.

Karplus, P. A., and Schulz, G. E., 1985, Prediction of chain flexibility in proteins, *Naturwissenschaften* **72:** 212–213.

Kauzmann, W., 1959, Some factors in the interpretation of protein denaturation, *Adv. Protein Chem.* **14:**1–63.

Kawakami, K., Noguchi, S., Noda, M., Takahashi, H., Ohta, T., Kawamura, M., Nojima, H., Nagano, K., Hirose, T., Inayama, S., Hayashida, H., Miyata, T., and Numa, S., 1985, Primary structure of the α-subunit of *Torpedo californica* ($Na^+ + K^+$)ATPase deduced from cDNA sequence, *Nature* **316:** 733–736.

Kelly, L., and Holladay, L. A., 1987, Comparison of scales of amino acid side chain properties by conservation during evolution of four proteins, *Protein Eng.* **1:**137–140.

Kim, P. S., and Baldwin, R. L., 1982, Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding, *Annu. Rev. Biochem.* **51:**459–489.

Klapper, I., Hagstrom, R., Fine, R., Sharp, K., and Honig, B., 1986, Focussing of electric fields in the active site of Cu–Zn superoxide dismutase: Effects of ionic strength and amino acid modification, *Proteins* **1:**47–59.

Kleffel, B., Garavito, R. M., Baumeister, W., and Rosenbusch, J. P., 1985, Secondary structure of a channel-forming protein: Porin from *E. coli* outer membrane, *EMBO J.* **4:**1589–1592.

Klein, P., 1986, Prediction of protein structural class by discriminant analysis, *Biochim. Biophys. Acta* **874:** 205–215.

Klein, P., and DeLisi, C., 1986, Prediction of protein structural class from amino acid sequence, *Biopolymers* **25:**1659–1672.

Klein, P., Kanehisa, M., and DeLisi, C., 1984, Prediction of protein function from sequence properties. Discriminant analysis of a data base, *Biochim. Biophys. Acta* **787:**221–226.

Klein, P., Kanehisa, M., and DeLisi, C., 1985, The detection of membrane-spanning proteins, *Biochim. Biophys. Acta* **815:**468–476.

Klein, P., Jacquez, J. A., and DeLisi, C., 1986, Prediction of protein function by discriminant analysis, *Math. Biosci.* **81:**177–189.

Kneale, G. G., and Bishop, M. J., 1985, Nucleic acid and protein sequence databases, *Cabios Rev.* **1:**11–17.

Kolaskar, A. S., Ramabrahmam, V., and Soman, K. V., 1980, Reversal of polypeptide chain in globular proteins, *Int. J. Peptide Protein Res.* **16:**1–11.

Kolb, E., Hudson, P. J., and Harris, J. I., 1980, Phosphofructokinase: Complete amino acid sequence of the enzyme from *Bacillus stearothermophilus, Eur. J. Biochem.* **108:**587–597.

Kopito, R. R., and Lodish, H. F., 1985, Primary structure and transmembrane orientation of the murine anion exchange protein, *Nature* **316:**234–238.

Kopito, R. R., Andersson, M., and Lodish, H. F., 1987, Structure and organization of the murine band 3 gene, *J. Biol. Chem.* **262:**8035–8040.

Kosower, E. M., 1982, in: *International Symposium on Structure and Dynamics of Nucleic Acids and Protein,* pp. 52–53.

Kosower, E. M., 1983a, Partial tertiary structure assignment for the acetylcholine receptor on the basis of the hydrophobicity of amino acid sequences and channel location using single group rotation theory, *Biochem. Biophys. Res. Commun.* **111:**1022–1024.

Kosower, E. M., 1983b, Partial tertiary structure assignments of the β_1-, γ-, and δ subunits of the acetylcholine receptor on the basis of the hydrophobicity of amino acid sequences and channel location using single group theory, *FEBS Lett.* **155:**245–247.

Kosower, E. M., 1987, A structural and dynamic model for the nicotinic acetylcholine receptor, *Eur. J. Biochem.* **168:**431–449.

Kotelchuck, D., and Scheraga, H. A., 1968, The influence of short-range interactions on protein conformation. I. Side-chain–backbone interactions with a single peptide unit, *Proc. Natl. Acad. Sci. U.S.A.* **61:**1163–1170.

Kotelchuck, D., and Scheraga, H. A., 1969, The influence of short-range interactions on protein conformation II. A model for predicting the α-helical regions of proteins, *Proc. Natl. Acad. Sci. U.S.A.* **62:**14–21.

Krchnak, V., Mach, O., and Maly, A., 1987, Computer prediction of potential immunogenic determinants from protein amino acid sequence, *Anal. Biochem.* **165:**200–207.

Krebs, K. E., and Phillips, M. C., 1984, The contribution of α-helices to the surface activities of proteins, *FEBS Lett.* **175**:263–266.

Kubo, T., Fukuda, K., Mikami, A., Maeda, A., Takahashi, H., Mishina, M., Hoga, T., Haga, K., Ichiyama, A., Kangawa, K., Kojima, M., Matsuo, H., Hirose, T., and Numa, S., 1986, Cloning, sequencing and expression of complementary DNA encoding the muscarinic acetylcholine receptor, *Nature* **323**:411–416.

Kubota, Y., Takahashi, S., Nishikawa, K., and Ooi, T., 1981, Homology in protein sequences expressed by correlation coefficients, *J. Theor. Biol.* **91**:347–361.

Kuhn, L. A., and Leigh, J. S., Jr., 1985, A statistical technique for predicting membrane protein structure, *Biochim. Biophys. Acta* **828**:351–361.

Kuntz, I. D., 1972, Protein folding, *J. Am. Chem. Soc.* **94**:4009–4012.

Kuntz, I. D., 1975, An approach to the tertiary structure of globular proteins, *J. Am. Chem. Soc.* **97**:4362–4366.

Kuntz, I. D., and Crippen, G. M., 1979, Protein densities, *Int. J. Peptide Protein Res.* **13**:223–228.

Kuntz, I. D., Crippen, G. M., Kollman, P. A., and Kimelman, D., 1976, Calculation of protein tertiary structure, *J. Mol. Biol.* **106**:983–994.

Kuntz, I. D., Blaney, J. M., Oatley, S. J., Langridge, R., and Ferrin, T. E., 1982, A geometric approach to macromolecule–ligand interactions, *J. Mol. Biol.* **161**:269–288.

Kyte, J., and Doolittle, R. F., 1982, A simple method for displaying the hydropathic character of a protein, *J. Mol. Biol.* **157**:105–132.

Lathrop, R. H., Webster, T. A., and Smith, T. F., 1987, ARIADNE: Pattern-directed inference and hierarchical abstraction in protein structure recognition, *Commun. ACM* **30**:909–921.

Laursen, R. A., Samiullah, M., and Lees, M. B., 1984, The structure of bovine brain myelin proteolipid and its organization in myelin, *Proc. Natl. Acad. Sci. U.S.A.* **81**:2912–2916.

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

Lee, C. A., and Saier, M. H., Jr., 1983, Mannitol-specific enzyme II of the bacterial phosphotransferase system, *J. Biol. Chem.* **258**:10761–10767.

Lenstra, J. A., 1977, Evolution of secondary structure prediction of proteins, *Biochim. Biophys. Acta* **491**:333–398.

Lenstra, J. A., Hofsteenge, J., and Beintema, J. J., 1977, Invariant features of the structure of pancreatic ribonuclease. A test of different predictive models, *J. Mol. Biol.* **109**:185–193.

Lesk, A. M., and Chothia, C., 1980, How different amino acid sequences determine similar protein structures: The structure and evolutionary dynamics of the globins, *J. Mol. Biol.* **136**:225–270.

Lesk, A. M., and Rose, G. D., 1981, Folding units in globular proteins. *Proc. Natl. Acad. Sci. U.S.A.* **78**: 4304–4308.

Lesk, A., Levitt, M., and Chothia, C., 1986, Alignment of the amino acids sequences of distantly related proteins using variable gap penalties, *Protein Eng.* **1**:77–78.

Leszczynski, J., and Rose, G. D., 1986, Loops in globular proteins: A novel category of secondary structure, *Science* **234**:849–855.

Leung, D. W., Spenser, S. A., Cachianes, G., Hammonds, R. G., Collins, C., Henzel, W. J., Barnard, R., Waters, M. J., and Wood, W. I., 1987, Growth hormone receptor and serum binding protein: Purification, cloning and expression, *Nature* **330**:537–543.

Levin, J. M., Robson, B., and Garnier, J., 1986, An algorithm for secondary structure determination in proteins based on sequence similarity, *FEBS Lett.* **205**:303–308.

Levinthal, C., 1968, Are there pathways for protein folding? *J. Chem. Phys.* **65**:44–45.

Levitt, M., 1974, On the nature of the binding of hexa-N-acetylglucosamine substrate to lysozyme, in: *Peptides, Polypeptides, and Proteins* (E. R. Blout, F. A. Bovey, M. Goodman, and N. Lotan, eds.), John Wiley & Sons, New York, pp. 99–113.

Levitt, M., 1976, A simplified representation of protein conformations for rapid simulation of protein folding, *J. Mol. Biol.* **104**:59–107.

Levitt, M., 1978, Conformational preferences of amino acids in globular proteins, *Biochemistry,* **17**:4277–4285.

Levitt, M., 1983, Protein folding by restrained energy minimization and molecular dynamics, *J. Mol. Biol.* **170**: 723–764.

Levitt, M., and Chothia, C., 1976, Structural patterns in globular proteins, *Nature* **261**:552–558.

Levitt, M., and Greer, J., 1977, Automatic identification of secondary structure in globular proteins, *J. Mol. Biol.* **114:**181–293.

Levitt M., and Lifson, S., 1969, Refinement of protein conformations using a macromolecular energy minimization procedure, *J. Mol. Biol.* **46:**269–279.

Levitt, M., and Warshel, A., 1975, Computer simulation of protein folding, *Nature* **253:**694–698.

Lewis, P. N., and Bradbury, E. M., 1974, Effect of electrostatic interactions on the prediction of helices in proteins, *Biochim. Biophys. Acta* **336:**153–164.

Lewis, P. N., and Scheraga, H. A., 1971, Predictions of structural homologies in cytochrome c proteins, *Arch. Biochem. Biophys.* **144:**576–583.

Lewis, P. N., Go, N., Go, M., Kotelchuck, D., and Scheraga, H. A., 1970, Helix probability profiles of denatured proteins and their correlation with native structures, *Proc. Natl. Acad. Sci. U.S.A.* **65:** 810–815.

Lewis, P. N., Momany, F. A., and Scheraga, H. A., 1971, Folding of polypeptide chains in proteins: A proposed mechanism for folding, *Proc. Natl. Acad. Sci. U.S.A.* **68:**2293–2297.

Lewis, P. N., Momany, F. A., and Scheraga, H. A., 1973a, Chain reversals in proteins, *Biochim. Biophys. Acta* **303:**211–229.

Lewis, P. N., Momany, F. A., and Scheraga, H. A., 1973b, Energy parameters in polypeptides. VI. Conformational energy analysis of the N-acetyl-N′-methyl amides of the twenty naturally occurring amino acids, *Israel J. Chem.* **11:**121–152.

Lifson, S., and Roig, A., 1961, On the theory of helix–coil transitions in polypeptides, *J. Chem. Phys.* **34:** 1963–1974.

Lifson, S., and Sander, C., 1979, Antiparallel and parallel β-strands differ in amino acid residue preferences, *Nature* **282:**109–111.

Lifson, S., and Sander, C., 1980a, Specific recognition in the tertiary structure of β-sheets in proteins, *J. Mol. Biol.* **139:**627–639.

Lifson, S., and Sander, C., 1980b, Composition, cooperativity and recognition in proteins, in: *Protein Folding* (R. Jaenicke, ed.), Elsevier/North-Holland Biomedical Press, Amsterdam, pp. 289–316.

Lifson, S., and Warshel, A., 1968, Consistent force field calculations, vibrational spectra, and enthaipies of cycloalkane and *n*-alkane molecules, *J. Chem. Phys.* **49:**5116–5129.

Liljas, A., and Rossman, M. G., 1974, X-ray studies of protein interactions, *Annu. Rev. Biochem.* **43:**475–507.

Lim, V. I., 1974a, Structural principles of the globular organization of protein chains: A stereochemical theory of globular protein secondary structure, *J. Mol. Biol.* **88:**857–872.

Lim, V. I., 1974b, Algorithms for prediction of α-helices and β-structural regions in globular proteins, *J. Mol. Biol.* **88:**873–894.

Lipman, D. J., and Pearson, W. R., 1985, Rapid and sensitive protein similarity searches, *Science* **227:**1435–1441.

Lisium, L., Finer-Moore, J., Stroud, R. M., Luskey, K. L., Brown, M. S., and Goldstein, J. L., 1985, Domain structure of 3-hydroxy-3-methylglutaryl coenzyme A reductase, a glycoprotein of the endoplasmic reticulum, *J. Biol. Chem.* **260:**522–530.

London, E., and Khorana, H. G., 1982, Denaturation and renaturation of bacteriorhodopsin in detergents and lipid–detergent mixtures, *J. Biol. Chem.* **257:**7003–7011.

Long, M. M., Urry, D. W., and Stoeckenius, W., 1977, Circular dichroism of biological membranes: Purple membrane of *Halobacterium halobium, Biochem. Biophys. Res. Commun.* **75:**725–731.

Lopez, J. A., Chung, D. W., Fujikawa, K., Hagen, F. S., Papayannopoulou, T., and Roth, G. J., 1987, Cloning of the α chain of human platlet glycoprotein 1b: A transmembrane protein with homology to leucine-rich α_2-glycoprotein, *Proc. Natl. Acad. Sci. U.S.A.* **84:**5615–5619.

Loucheux-Lefebvre, M.-H., 1978, Predicted β-turns in peptide and glycopeptide antifreezes, *Biochem. Biophys. Res. Commun.* **81:**1352–1356.

Loucheux-Lefebvre, M.-H., Aubert, J.-P., and Jolles, P., 1978, Prediction of the conformation of the cow and sheep k-caseins, *Biophys. J.* **23:**323–336.

Low, B. W., Lovell, F. M., and Rudko, A. D., 1968, Prediction of α-helical regions in proteins of known sequence, *Proc. Natl. Acad. Sci. U.S.A.* **60:**1519–1526.

MacLennan, D. H., Brandl, C. J., Korczak, B., and Green, N. M., 1985, Amino acid sequences of a Ca^+ + Mg^+-dependent ATPase from rabbit muscle sarcoplasmic reticulum, deduced from its complementary DNA sequence, *Nature* **316:**696–700.

Maizel, J. V., Jr., and Lenk, R. P., 1981, Enhanced graphic matrix analysis of nucleic acid and protein sequence, *Proc. Natl. Acad. Sci. U.S.A.* **78:**7665–7669.

Manavalan, P., and Ponnuswamy, P. K., 1977, A study of the preferred environment of amino acid residues in globular proteins, *Arch. Biochem. Biophys.* **184:**476–487.

Manavalan, P., and Ponnuswamy, P. K., 1978, Hydrophobic character of amino acid residues in globular proteins, *Nature* **275:**673–674.

Mao, D., and Wallace, B. A., 1984, Differential light scattering and absorption flattening optical effects are minimal in circular dichroism spectra of small unilamellar vesicles, *Biochemistry* **23:**2667–2673.

Masu, Y., Nakayama, K., Tamaki, H., Harada, Y., Kuno, M., and Nakanishi, S., 1987, cDNA cloning of bovine substance-K receptor through oocyte expression system, *Nature* **329:**836–838.

Matthew, J. B., 1985, Electrostatic effects in proteins, *Annu. Rev. Biophys. Biophys. Chem.* **14:**387–417.

Matthew, J. B., and Gurd, F. R. N., 1986a, Calculation of electrostatic interactions in proteins, *Methods Enzymol.* **130:**413–436.

Matthew, J. B., and Gurd, F. R. N., 1986b, Stabilization and destabilization of protein structure by charge interactions, *Methods Enzymol.* **130:**437–453.

Matthews, B. W., 1975, Comparison of the predicted and observed secondary structure of T4 phage lysozyme, *Biochim. Biophys. Acta* **405:**442–451.

Matthews, F. S., Argos, P., and Levine, M., 1971, The structure of cytochrome b_5 at 2.0 Å resolution, *Cold Spring Harbor Symp. Quant. Biol.* **36:**387.

Maxfield, F. R., and Scheraga, H. A., 1975, The effect of neighboring charges on the helix forming ability of charged amino acids in proteins, *Macromolecules* **8:**491–493.

Maxfield, F. R., and Scheraga, H. A., 1976, Status of empirical methods for the prediction of protein backbone topography, *Biochemistry* **15:**5138–5153.

Maxfield, F. R., and Scheraga, H. A., 1979, Improvements in the prediction of protein backbone topography by reduction of statistical errors, *Biochemistry* **18:**697–704.

McCammon, J. A., Gelin, B. R., and Karplus, M., 1977, Dynamics of folded proteins. *Nature* **267:**585–590.

McCarthy, M. P., Earnest, J. P., Young, E. F., Choe, S., and Stroud, R. M., 1986, The molecular neurobiology of the acetylcholine receptor, 1986, *Annu. Rev. Neurosci.* **9:**383–413.

McCubbin, W. D., Oikawa, K., and Kay, C. M., 1971, Circular dichroism studies on concanavalin A, *Biochem. Biophys. Res. Commun.* **43:**666–674.

McGregor, M. J., Islam, S. A., and Sternberg, M. J. E., 1987, Analysis of the relationship between side-chain conformation and secondary structure in globulary proteins, *J. Mol. Biol.* **198:**295–310.

McLachlan, A. D., 1971, Tests for comparing related amino-acid sequences. Cytochrome c and cytochrome c_{551}, *J. Mol. Biol.* **61:**409–424.

McLachlan, A. D., 1977, Quantum chemistry and protein folding: The art of the possible, *Int. J. Quant. Chem.* **13**(Suppl. 1):371–385.

McLachlan, A. D., and Karn, J., 1983, Periodic features in the amino acid sequence of nemotode myosin rod, *J. Mol. Biol.* **164:**605–626.

McLachlan, A. D., and Stewart, M., 1976, The 14-fold periodicity in α-tropomyosin and the interaction with actin, *J. Mol. Biol.* **103:**271–298.

McLachlan, A. D., Bloomer, A. C., and Butler, P. J. G., 1980, Structural repeats and evolution of tobacco mosaic virus coat protein and RNA, *J. Mol. Biol.* **136:**203–224.

Meirovitch, S., Rackovsky, S., and Scheraga, H. A., 1980, Empirical studies of hydrophobicity. 1. Effect of protein size on the hydrophobic behavior of amino acids, *Macromolecules* **13:**1398–1405.

Menick, D. R., Carrasco, N., Antes, L., Patel, L., and Kaback, H. R., 1986, *Lac* permease of *Escherichia coli:* Arginine-302 as a component of the postulated proton relay, *Biochemistry* **26:**6638–6644.

Mercier, J.-C., and Chobert, J. M., 1976, Comparative study of the amino acid sequences of the caseinomacropeptides from seven species, *FEBS Lett.* **72:**208–214.

Mercier, J.-C., Uro, J., Ribadeau-Daumas, B., and Grosclaude, F., 1972, Structure primaire du caséinomacropeptide de la caséine k_{β_1} bovine, *Eur. J. Biochem.* **27:**535–547.

Miles, E. W., Yutani, K., and Ogarsahara, K., 1982, Guanidine hydrochloride-induced unfolding of the α-subunit of tryptophan synthetase and of the two α-proteolytic fragments. Evidence for stepwise unfolding of the two α domains, *Biochemistry* **21:**2586–2592.

Milner-White, E. J., 1988, Recurring loop motif in proteins that occurs in right-handed and left-handed forms. Its relationship with alpha-helices and beta-bulge loops, *J. Mol. Biol.* **199:**503–511.

Milner-White, E. J., and Poet, R., 1986, Four classes of β-hairpins in proteins, *Biochem. J.* **240:**289–292.

Milner-White, E. J., and Poet, R., 1987, Loops, bulges, turns and hairpins in proteins, *Trends Biochem. Sci.* **12:**189–192.

Modrow, S., and Wolf, H., 1986, Characterization of two related Epstein–Barr virus-encoded membrane proteins that are differentially expressed in Burkitt lymphoma and *in vitro*-transformed cell lines, *Proc. Natl. Acad. Sci. U.S.A.* **83:**5703–5707.

Mohana-Rao, J. K., and Argos, P., 1986, A conformational preference parameter to predict helices in integral membrane proteins, *Biochim. Biophys. Acta* **869:**197–214.

Mohana-Rao, J. K., Hargrave, P. A., and Argos, P., 1983, Will the seven-helix bundle be a common structure for integral membrane proteins? *FEBS Lett.* **156:**165–169.

Momany, F. A., MacGuire, R. F., Burgess, A. W., and Scheraga, H. A., 1975, Energy parameters in polypeptides. VII. Geometric parameters, partial atomic charges, nonbonded interactions, hydrogen bond interaction, and intrinsic torsion potentials for the naturally occurring amino acids, *J. Phys. Chem.* **79:** 2361–2381.

Mononen, I., and Karjalainen, E., 1984, Structural comparisons of protein sequences around potential N-glycosylation sites, *Biochim. Biophys. Acta* **788:**364–367.

Moore, W. M., Holladay, L. A., Puett, D., and Brody, R. N., 1974, On the conformation of the acetylcholine receptor protein from *Torpedo nobiliana, FEBS Lett.* **45:**145–149.

Moran, E. C., Chou, P. Y., and Fasman, G. D., 1977, Conformational transitions of glucagon in solution: The $\alpha \rightleftarrows \beta$ transition, *Biochem. Biophys. Res. Commun.* **77:**1300–1306.

Morgan, D. O., Edman, J. C., Standring, D. N., Fried, V. A., Smith, M. C., Roth, R. A., and Rutter, W. J., 1987, Insulin-like growth factor II receptor as a multifunctional binding protein, *Nature* **329:** 301–307.

Morgan, R. S., and McAdon, J. H., 1980, Predictor for sulfur–aromatic interactions in globular proteins, *Int. J. Peptide Protein Res.* **15:**177–180.

Moser, R., Thomas, R. M., and Gutte, B., 1983, An artificial crystalline DDT-binding polypeptide, *FEBS Lett.* **157:**247–251.

Moser, R., Frey, S., Münger, K., Hehlgans, T., Klausen, S., Langen, H., Winnacker, E.-L., Mertz, R., and Gutte, B., 1987, Expression of the synthetic gene of an artificial DDT-binding polypeptide of *E. coli, Protein Eng.* **1:**339–343.

Moult, J., and James, M. N. G., 1987, An algorithm for determining the conformation of polypeptide segments in proteins by systematic search, *Proteins* **1:**146–163.

Mueckler, M., Caruso, C., Baldwin, S. A., Panico, M., Blench, I., Morris, H. R., Ailard, W. J., Lienhard, G. E., and Lodish, H. F., 1985, Sequence and structure of a human glucose transporter, *Science* **299:**941–945.

Murakami, M., 1985, Mutation affecting the 12th and 61st amino acids of p21 protein result in decreased probability of β-turn occurrence around the mutation positions: A prediction, *J. Theor. Biol.* **114:**193–198.

Murakami, M., 1987, Critical amino acids of p21 protein are located within β-turns: Further evaluation, *J. Theor. Biol.* **128:**339–347.

Murata, M., Richardson, J. S., and Sussman, J., 1985, Simultaneous comparison of three protein sequences, *Proc. Natl. Acad. Sci. U.S.A.* **82:**3073–3077.

Murphy, J., Zhang, W.-J., Macaulay, W., Fasman, G., and Merrifield, R. B., 1988, The relation of predicted structure to observed conformation and activity of glucagon analogs containing replacements at positions 19, 22 and 23, *J. Biol. Chem.* **262:**17304–17312.

Murzin, A. G., and Finkelstein, A. V., 1983, Polyhedra describing the packing of helices in a protein globule, *Biofisika* **28:**905–911.

Mutter, M., 1985, The construction of new proteins and enzymes. A prospect for the future, *Angew. Chem. [Engl.]* **24:**639–653.

Nabedryk, E., Bardin, A. M., and Breton, J., 1985, Further characterization of protein secondary structures in purple membrane by circular dichroism and polarized infrared spectroscopies, *Biophys. J.* **48:**873–876.

Nagano, K., 1973, Logical analysis of the mechanism of protein folding. I. Prediction of helices, loops and β-structures from primary structure, *J. Mol. Biol.* **75:**401–420.

Nagano, K., 1974, Logical analysis of the mechanism of protein folding. II. The nucleation process, *J. Mol. Biol.* **84:**337–372.

Nagano, K., 1977, Triplet information in helix prediction applied to the analysis of super-secondary structures, *J. Mol. Biol.* **109:**251–274.

Nagarajan, M., and Rao, V. S. R., 1977, Conformational analysis of glycoproteins. Part I. Conformation of the protein segment at the site of peptide-sugar linkage, *Current Sci.* **46**:395–400.

Nakashima, H., Nishikawa, K., and Ooi, T., 1986, The folding type of a protein is relevant to the amino acid composition, *J. Biochem. (Tokyo)* **99**:153–162.

Narayana, S. V. L., and Argos, P., 1984, Residue contacts in protein structures and implications for protein folding, *Int. J. Peptide Protein Res.* **24**:25–39.

Needleman, S. B., and Blair, T. T., 1969, Homology of *Pseudomonas* cytochrome c-551 with eukaryotic c-cytochromes, *Proc. Natl. Acad. Sci. U.S.A.* **63**:1227–1233.

Needleman, S. B., and Wunsch, C. D., 1970, A general method applicable to the search for similarities in the amino acid sequences of two proteins, *J. Mol. Biol.* **48**:443–453.

Nemethy, G., 1974, Conformational energy calculations and the folding of proteins, *PAABS Rev.* **3**:51–61.

Nemethy, G., and Scheraga, H. A., 1977, Protein folding, *Q. Rev. Biophys.* **3**:239–352.

Nemethy, G., and Scheraga, H. A., 1980, Stereochemical requirements for the existence of hydrogen bonds in β-bends, *Biochem. Biophys. Res. Commun.* **95**:320–327.

Neuberger, A., and Marshall, R. D., 1968, Aspects of the structure of glycoproteins, in: *Carbohydrates and Their Roles* (H. W. Schultze, R. F. Cain, and R. W. Molstad, eds.), AVI, Westport, CT, p. 115.

Nishikawa, K., 1983, Assessment of secondary structure prediction of proteins: Comparisons of computerized Chou–Fasman method with others, *Biochim. Biophys. Acta* **748**:285–299.

Nishikawa, K., and Ooi, T., 1986, Amino acid sequence homology applied to the prediction of protein secondary structure, and joint prediction with existing methods, *Biochim. Biophys. Acta* **871**:45–54.

Nishikawa, K., Momany, F. A., and Scheraga, H. A., 1974, Low energy structure of two dipeptides and their relationship to bend conformation, *Macromolecules* **7**:797–806.

Noda, M., Takahashi, H., Tanabe, T., Toyosato, M., Furutani, Y., Hirose, T., Asai, M., Inayama, S., Miyata, T., and Numa, S., 1982, Primary structure of α-subunit precursor of *Torpedo californica* acetylcholine receptor deduced from c-DNA sequence, *Nature* **299**:793–797.

Noda, M., Takahashi, H., Tanabe, T., Toyosato, M., Kikyotani, S., Furutani, Y., Hirose, T., Takashima, G., Inayama, S., Miyata, T., and Numa, S., 1983, Structural homology of *Torpedo californica* receptor subunits, *Nature* **302**:528–532.

Noda, M., Shimizu, S., Tanabe, T., Takai, T., Kayano, T., Ikeda, T., Takahashi, H., Nakayama, H., Kanaoka, Y., Minamino, N., Kangawa, K., Matsuo, H., Raftery, M. A., Hirose, T., Inayama, S., Hayashida, H., Miyata, T., and Numa, S., 1984, Primary structure of *Electrophorus electricus* sodium channel deduced from c-DNA sequence, *Nature* **312**:121–127.

Noda, M., Ikeda, T., Kayano, T., Suzuki, H., Takeshima, J.-H., Kurasaki, M., Takahashi, T., and Numa, S., 1986, Existence of distinct sodium channel messenger RNAs in rat brain, *Nature* **320**:188–192.

Novotny, J., and Auffray, C., 1984, A program for prediction of protein secondary structure from nucleotide sequence data: Application to histocompatability antigens, *Nucleic Acids Res.* **12**:243–255.

Novotny, J., and Haber, E., 1985, Structural invariants of antigen binding: Comparison of immunoglobulin V_2–V_H and V_2–V_2 domain dimers, *Proc. Natl. Acad. Sci. U.S.A.* **82**:4592–4596.

Novotny, J., Bruccoleri, R. E., and Newell, J., 1984, Twisted hyperboloid (strophoid) as a model of β-barrels in proteins, *J. Mol. Biol.* **177**:567–573.

Nozaki, Y., and Tanford, C., 1971, The solubility of amino acids and two glycine peptides in aqueous ethanol and dioxane solutions, *J. Biol. Chem.* **246**:2211–2217.

Nussbaum, S., Beaudette, N. V., Fasman, G. D., Potts, J. T., Jr., and Rosenblatt, M., 1985, Design of analogues of parathyroid hormone: A conformational approach, *J. Protein Chem.* **4**:391–406.

Otvös, L., Jr., Hollosi, M., Perczel, A., Dietzschold, B., and Fasman, G. D., 1988, Phosphorylation loops in synthetic peptides of the human neurofilament protein middle-sized subunit, *J. Protein Chem.* **7**:365–376.

Ovchinnikov, Yu. A., 1982, Rhodopsin and bacteriorhodopsin: Structure-function relationships, *FEBS Lett.* **148**:179–191.

Ovchinnikov, Yu., A., Abdulaev, N. G., Feigina, M. Y., Kiselev, A. V., and Lobanov, N. A., 1979, The structural basis of the functioning of bacteriorhodopsin: An overview, *FEBS Lett.* **100**:219–224.

Ovchinnikov, Yu., A., Abdulaev, N. G., Vasilov, R. G., Vturina, I. Yu., Kuryatov, A. B., and Kiselev, A. V., 1985, The antigenic structure and topography of bacteriorhodopsin in purple membranes as determined by interaction with monoclonal antibodies, *FEBS Lett.* **179**:343–350.

Ovchinnikov, Yu., A., Modyanov, N. N., Brovde, N. E., Petrukhin, K. E., Grishin, A. V., Arzamazova, N. M., Aldanova, N. A., Monastyrskaya, G. S., and Sverdlov, E. D., 1986, Pig kidney, $Na^+ + K^+$ATPase, *FEBS Lett.* **201**:237–245.

Padlan, E. A., 1977, Structural implications of sequence variability in immunoglobulins, *Proc. Natl. Acad. Sci. U.S.A.* **74**:2551–2555.

Palau, J., Argos, P., and Puigdomenech, P., 1982, Protein secondary structure studies on the limits of prediction accuracy, *Int. J. Peptide Protein Res.* **19**:394–401.

Pallai, P. V., Mabilla, M., Goodman, M., Vale, W., and Rivier, J., 1983, Structural homology of corticotropin-releasing factor, sauvagine and urotensin I: Circular dichroism and prediction studies, *Proc. Natl. Acad. Sci. U.S.A.* **80**:6770–6774.

Parrilla, A., Domenech, A., and Querol, E., 1986, A PASCAL microcomputer program for prediction of protein secondary structure and hydropathic segments, *Cabios* **2**:211–215.

Pashley, R. M., McGuiggan, P. M., Ninham, B. W, and Evans, D. F., 1985, Attractive forces between uncharged hydrophobic surfaces: Direct measurements in aqueous solution, *Science* **229**:1088–1089.

Patten, P., Yokota, T., Rothbard, J., Chien, Y.-H., Arai, K.-I., and Davis, M. M., 1984, Structure, expression and divergence of T-cell receptor β-chain variable regions, *Nature* **312**:40–46.

Pattus, F., Heitz, F., Martinez, C., Provencher, S. W., and Lazdunski, C. L., 1985, Secondary structure of the pore-forming colicin A and its C-terminal fragment. Experimental fact and structure prediction, *Eur. J. Biochem.* **152**:681–689.

Paul, C., and Rosenbusch, J. P., 1985, Folding patterns of porin and bacteriorhodopsin, *EMBO J.* **4**:1593–1597.

Paul, C. H., 1982, Building models of globular proteins. Molecules from their amino acid sequences. I. Theory, *J. Mol. Biol.* **155**:53–62.

Pauling, L., and Corey, R. B., 1951, Configurations of polypeptide chains with favored orientations around single bonds: Two new pleated sheets, *Proc. Natl. Acad. Sci. U.S.A.* **37**:729–740.

Peralta, E. G., Winslow, J. W., Peterson, G. L., Smith, D. H., Ashkenazi, A., Ramachandran, J., Schimerlik, M. I., and Capon, D. J., 1987, Primary structure and biochemical properties of an M_2 muscarinic receptor, *Science* **236**:600–605.

Periti, P. F., Quagliarotti, G., and Liquori, A. M., 1967, Recognition of α-helical segments in proteins of known primary structure, *J. Mol. Biol.* **24**:313–322.

Perutz, M. F., 1980, Electrostatic effects in proteins, *Science* **201**:1187–1191.

Perutz, M. F., Gronenborn, A. M., Clore, G. M., Fogg, J. H., and Shih, D. T.-B, 1985, The pK_a values of two histidine residues in human haemoglobin, the Bohr effect, and the dipole moments of α-helices, *J. Mol. Biol.* **183**:491–498.

Phillips, D. C., 1970, in: *British Biochemistry, Past and Present* (T. W. Goodwin, ed.), Academic Press, London, pp. 11–28.

Pincus, M. R., and Klausner, R. D., 1982, Predictions of the three-dimensional structure of the leader sequence of pre-κ-light chain, a hexadecapeptide, *Proc. Natl. Acad. Sci. U.S.A.* **79**:3413–3417.

Pohlman, R., Nagel, G., Schmidt, B., Stein, M., Lorkowski, G., Krentler, C., Cully, J., Meyer, H. E., Grzeschik, K.-H., Mersmann, G., Hasilik, A., and von Figura, K., 1987, Cloning of a c-DNA encoding the human cation-dependent mannose-6-phosphate receptor, *Proc. Nat. Acad. Sci. U.S.A.* **84**:5575–5579.

Ponder, J. W., and Richards, F. M., 1987, Tertiary templates for proteins. Use of packing criteria in the enumeration of allowed sequences for different structural classes, *J. Mol. Biol.* **193**:775–791.

Pongor, S., and Szaley, A. A., 1985, Prediction of homology and divergence in the secondary structure of polypeptides, *Proc. Natl. Acad. Sci. U.S.A.* **82**:366–370.

Ponnuswamy, P. K., Worme, P. K., and Scheraga, H. A., 1973, Role of medium-range interactions in proteins, *Proc. Natl. Acad. Sci. U.S.A.* **70**:830–833.

Ponnuswamy, P. K., Prabhakran, M., and Manavalan, P., 1981, Hydrophobic packing and spatial arrangement of amino acid residues in globular proteins, *Biochim. Biophys. Acta* **623**:301–316.

Popot, J.-L., Gerchman, S.-E., and Engelman, D. M., 1987, Refolding of bacteriorhodopsin in lipid bilayers. A thermodynamically controlled two-stage process, *J. Mol. Biol.* **198**:655–676.

Post, C. B., Brooks, B. R., Karplus, M., Dobson, C. M., Artymiuk, P. C., Cheetham, J. C., and Phillips, D. C., 1986, Molecular dynamics. Simulations of native and substrate bound lysozyme. A study of the average structures and atomic fluctuations, *J. Mol. Biol.* **190**:455–479.

Potts, J. T., Jr., Kronenberg, H. M., and Rosenblatt, M., 1982, Parathyroid hormone: Chemistry, biosynthesis, and mode of action, *Adv. Protein Chem.* **35**:322–396.

Prothero, J. W., 1966, Correlation between the distribution of amino acids and alpha helices, *Biophys. J.* **6**: 367–370.

Prothero, J. W., 1968, A model of alpha-helical distribution in proteins, *Biophys. J.* **8**:1236–1255.

Ptitsyn, O. B., 1969, Statistical analyses of the distribution of amino acid residues among helical and nonhelical regions in globular proteins, *J. Mol. Biol.* **42:**501–510.

Ptitsyn, O. B., 1981, Protein folding: General physical model, *FEBS Lett.* **131:**197–202.

Ptitsyn, O. B., 1985, Physical principles of protein structure and protein folding, *J. Biosci.* **8:**1–13.

Ptitsyn, O. B., and Finkelstein, A. V., 1970a, Connection between the secondary and primary structures of globular proteins, *Biofisika* **15:**757–768.

Ptitsyn, O. B., and Finkelstein, A. V., 1970b, Prediction of helical portions of globular proteins according to their primary structure, *Dokl. Akad. Nauk. SSSR* **195:**221–224.

Ptitsyn, O. B., and Finkelstein, A. V., 1979, Coiling and topology of the parallel β-structure, *Biofisika* **24:**27–30.

Ptitsyn, O. B., and Finkelstein, A. V., 1983, Theory of protein secondary structure and algorithm of its prediction, *Biopolymers* **22:**15–25.

Ptitsyn, O. B., and Rashin, A. A., 1975, A model of myoglobin self-organization, *Biophys. Chem.* **3:**1–20.

Ptitsyn, O. B., Finkelstein, A. V., and Falk, P., 1979, Principal folding pathway and topology of all β-proteins, *FEBS Lett.* **101:**1–5.

Pullman, B., and Pullman, A., 1974, Molecular orbital calculations on the conformation of amino acid residues of proteins, *Adv. Protein Chem.* **28:**347–526.

Pumphrey, R. S. H., 1986a, Computer models of the human immunoglobulins. I. Shape and segmental flexibility, *Immunol. Today* **7:**174–178.

Pumphrey, R. S. H., 1986b, Computer models of the human immunoglobulins. II. Binding sites and molecular interactions, *Immunol. Today* **7:**206–211.

Quiocho, F. A., Sack, J. S., and Vyas, N. K., 1987, Stabilization of charges on isolated ionic groups sequestered in proteins by polarized peptide units, *Nature* **329:**561–564.

Rackovsky, S., and Goldstein, D. A., 1987, Differential geometry and protein conformation. V. Medium-range conformational influence of the individual amino acids, *Biopolymers* **26:**1163–1187.

Ralph, W. W., Webster, T., and Smith, T. F., 1987, A modified Chou and Fasman protein structure algorithm, *Cabios* **3:**211–216.

Ramachandran, G. N., Ramakrishnan, C., and Sasisekharan, V., 1963, Stereochemistry of polypeptide chain configurations, *J. Mol. Biol.* **7:**95–99.

Rao, S. T., and Rossman, M. G., 1973, Comparison of super-secondary structures in proteins, *J. Mol. Biol.* **76:** 241–250.

Rao, J. K. M., Hargrave, P. A., and Argos, P., 1983, Will the seven-helix bundle be a common structure for integral membrane proteins? *FEBS Lett.* **156:**165–169.

Rashin, A. A., 1981, Location of domains in globular proteins, *Nature* **291:**85–86.

Rashin, A. A., and Honig, B., 1984, On the environment of ionizable groups in globular proteins, *J. Mol. Biol.* **173:**515–521.

Rawlings, N., Ashman, K., and Wittman-Leibold, B., 1983, Computerized version of the Chou and Fasman protein secondary structure predictive method, *Int. J. Peptide Protein Res.* **22:**515–524.

Remington, S. J., and Matthews, B. W., 1978, A general method to assess similarity of protein structures, with applications to T4 bacteriophage lysozyme, *Proc. Natl. Acad. Sci. U.S.A.* **75:**2180–2184.

Remington, S. J., Anderson, W. F., Owen, J., Ten Eyck, L. F., Grainger, C. T., and Matthews, B. W., 1978, Structure of the lysozyme from bacteriophage T4: An electron density map at 2.4 Å resolution, *J. Mol. Biol.* **118:**81–98.

Renugopalakrishnan, V., Strawich, E. S., Horowitz, P. M., and Glimcher, M. J., 1986, Studies on the secondary structures of amelogenin from bovine tooth enamel, *Biochemistry* **25:**4879–4887.

Ricard, J. M., Perez, J. J., Pons, M., and Giralt, E., 1983, Conformational basis of N-glycosylation of proteins: Conformational analysis of Ac-Asn-Ala-Thr-NH_2, *Int. J. Biol. Macromol.* **5:**279–282.

Richards, F. M., 1974, The interpretation of protein structures: Total volume, group volume distributions and packing density, *J. Mol. Biol.* **82:**1–14.

Richards, F. M., 1977, Areas, volumes, packing, and protein structure, *Annu. Rev. Biophys. Bioeng.* **6:**151–176.

Richardson, J. S., 1976, Handedness of crossover connections in β sheets. *Proc. Natl. Acad. Sci. U.S.A.* **73:** 2619–2623.

Richardson, J. S., 1977, β-Sheet topology and the relatedness of proteins, *Nature* **268:**495–500.

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

Richmond, T. J., 1984, Solvent accessible surface area and excluded volume in proteins, *J. Mol. Biol.* **178:**63–89.

Richmond, T. J., and Richards, F. M., 1978, Packing of α-helices: Geometrical constraints and contact areas, *J. Mol. Biol.* **119:**775–791.

Robson, B., 1974, Analysis of the code relating sequence to conformation in globular proteins: Theory and application of expected information, *Biochem. J.* **141:**853–867.

Robson, B., and Garnier, J., 1986, *Introduction to Proteins and Protein Engineering,* Elsevier, Amsterdam.

Robson, B., and Osguthorpe, D. J., 1979, Refined models for computer simulations of protein folding. Applications to the study of conserved secondary structure and flexible hinge points during the folding of pancreatic trypsin inhibitor, *J. Mol. Biol.* **132:**19–51.

Robson, B., and Pain, R. H., 1971, Analyses of the code relating sequence to conformation in proteins: Possible implications for the mechanism of formation of helical regions, *J. Mol. Biol.* **58:**237–259.

Robson, B., and Pain, R. H., 1972, Directional information transfer in protein helices, *Nature (New Biol.)* **238:** 107–108.

Robson, B., and Pain, R. H., 1974a, Analysis of the code relating sequence to conformation in globular proteins: Development of a stereochemical alphabet on the basis of intra-residue information, *Biochem. J.* **141:**869–882.

Robson, B., and Pain, R. H., 1974b, Analysis of the code relating sequence to conformation in globular proteins: An informational analysis of the role of the residue in determining the conformation of its neighbours in the primary sequence, *Biochem. J.* **141:**869–882.

Robson, B., and Pain, R. H., 1974c, Analysis of the code relating sequence to conformation in globular proteins: The distribution of residue pairs in turns and kinks in the backbone chain, *Biochem. J.* **141:**899–904.

Robson, B., and Platt, E., 1986, Refined models for computer calculations in protein engineering. Calibration and testing of atomic potential functions compatible with more efficient calculations, *J. Mol. Biol.* **188:** 259–281.

Robson, B., and Suzuki, E., 1976, Conformational properties of amino acid residues in globular proteins, *J. Mol. Biol.* **107:**327–356.

Robson, B., Platt, E., Fishleigh, R. V., Marsden, A., and Millard, P., 1987, Expert system for protein engineering: Its application in the study or chloroamphenicol acetyltransferase and avian pancreatic polypeptide, *J. Mol. Graphics* **5:**8–17.

Rogers, N. K., 1986, The modelling of electrostatic interactions in the function of globular proteins, *Prog. Biophys. Mol. Biol.* **48:**37–66.

Rogers, N. K., and Sternberg, M. J. E., 1984, Electrostatic interactions in globular proteins. Different dielectric models applied to the packing of α-helices, *J. Mol. Biol.* **174:**527–542.

Rose, G. D., 1978, Prediction of chain turns in globular proteins on a hydrophobic basis, *Nature* **272:**586–590.

Rose, G. D., 1979, Hierarchic organization of domains in globular proteins, *J. Mol. Biol.* **134:**447–470.

Rose, G. D., and Roy, S., 1980, Hydrophobic basis of packing in globular proteins, *Proc. Natl. Acad. Sci. U.S.A.* **77:**4643–4647.

Rose, G. D., and Seltzer, J. P., 1977, A new algorithm for finding the peptide chain turns in a globular proteins, *J. Mol. Biol.* **113:**153–164.

Rose, G. D., and Wetlaufer, D. B., 1977, The number of turns in globular proteins, *Nature* **268:**769–770.

Rose, G. D., Young, W. B., and Gierasch, L. M., 1983, Interior turns in globular proteins, *Nature* **304:**655–657.

Rose, G. D., Gierasch, L. M., and Smith, J. A., 1985, Turns in peptides and proteins, *Adv. Protein Chem.* **37:** 1–109.

Rosenblatt, M., Habener, J. F., Tyler, F. A., Shepard, G. L., and Potts, J. T., Jr., 1979, Chemical synthesis of the precursor-specific region of preproparathyroid hormone, *J. Biol. Chem.* **254:**1414–1421.

Rosenblatt, M., Beaudette, N. V., and Fasman, G. D., 1980, Conformational studies of the synthetic precursor-specific regions of pre-parathyroid hormone, *Proc. Natl. Acad. Sci. U.S.A.* **77:**3983–3987.

Rosenblatt, M., Majzoub, J. A., Beaudette, N. V., Kronenberg, H. M., Potts, J. T., Fasman, G. D., and Habener, J. F., 1981, Chemically synthesized precursor-specific fragment of preproparathyroid hormone: Conformational and biological properties, in: *Peptides 1980. Proceedings of the Sixteenth European Peptide Symposium* (K. Brunfeldt, ed.), Scriptor, Copenhagen, pp. 572–577.

Rossman, M. G., and Argos, P., 1981, Protein folding, *Annu. Rev. Biochem.* **50:**497–533.

Rossman, M. G., and Liljas, A., 1974, Recognition of structural domains in globular proteins, *J. Mol. Biol.* **85:** 177–181.

Rottier, P. J. M., Welling, G. W., Welling-Wester, S., Niesters, G. M., Lenstra, J. A., and van der Zeijst, B. A. M., 1986, Predicted membrane topology of the coronavirus E1, *Biochemistry* **25:**1335–1339.

Sack, G. H., Jr., 1983, Molecular cloning of human genes for serum amyloid A, *Gene 22:* 19–29.

Salemme, F. R., 1981, Conformational and geometrical properties of β-sheets in proteins: III. Isotropically stressed configurations, *J. Mol. Biol.* **146:**143–156.

Salemme, F. R., 1983, Structural properties of protein β-sheets, *Prog. Biophys. Mol. Biol.* **42:**95–133.

Salemme, F. R., and Weatherford, D. W., 1981a, Conformational and geometrical properties of β-sheets in proteins: I. Parallel β-sheets, *J. Mol. Biol.* **146:**101–117.

Salemme, F. R., and Weatherford, D. W., 1981b, Conformation and geometrical properties of β-sheets in proteins: II. Antiparallel and mixed β-sheets, *J. Mol. Biol.* **146:**119–141.

Sander, C., and Schulz, G. E., 1979, Degeneracy of the information contained in amino acid sequences: Evidence for overlaid genes, *J. Mol. Evol.* **13:**245–252.

Saraste, M., and Walker, J. E., 1982, Internal sequence repeats and the path of polypeptide in mitochondrial ADP/ATP translocase, *FEBS Lett.* **144:**250–254.

Sawyer, L., and James, M. N. G., 1982, Carboxyl–carboxylate interactions in proteins, *Nature* **295:**79–80.

Sayre, R., Anderson, B., and Bogorad, L., 1986, The topology of a membrane protein: The orientation of the 32 kd Qb-binding chloroplast thylakoid membrane protein, *Cell* **47:**601–608.

Scheraga, H. A., 1960, Structural studies of ribonuclease III. A model for the secondary and tertiary structure, *J. Am. Chem. Soc.* **82:**3847–3852.

Scheraga, H. A., 1968, Calculations of conformations of polypeptides, *Adv. Phys. Org. Chem.* **6:**103–184.

Scheraga, H. A., 1971, Theoretical and experimental studies of conformations of polypeptides, *Chem. Rev.* **71:** 195–217.

Scheraga, H. A., 1985, Calculations of the three-dimensional structures of proteins, *Ann. N.Y. Acad. Sci.* **439:** 170–194.

Schiffer, M., and Edmundson, A. B., 1967, Use of helical wheels to represent the structures of proteins and to identify segments with helical potential, *Biophys. J.* **7:**121.

Schiffer, M., Wu, T. T., and Kabat, E. A., 1986, Subgroups of variable regions genes of β-chains of T-cell receptors for antigen, *Proc. Natl. Acad. Sci. U.S.A.* **83:**4461–4463.

Schofield, P. R., Darlison, M. G., Fujita, N. Burt, D. R., Stephenson, F. A., Rodriguez, H., Rhee, L. M., Ramachandran, J., Reale, V., Glencorse, T. A., Seeburg, P. H., and Barnard, E. A., 1987, Sequence and functional expression of the $GABA_A$ receptor shows a ligand-gated receptor super-family, *Nature* **328:** 221–227.

Schulz, G. E., 1977, Structural rules for globular proteins, *Angew. Chem. [Engl.]* **16:**23–32.

Schulz, G. E., 1980, Gene duplication in glutathione reductase, *J. Mol. Biol.* **138:**335–347.

Schulz, G. E., and Schirmer, R. H., 1974, Topological comparison of adenyl kinase with other proteins, *Nature* **250:**142–144.

Schulz, G. E., and Schirmer, R. H., 1979, *Principles of Protein Structure,* Springer-Verlag, New York.

Schulz, G. E., Barry, C. D., Friedman, J., Chou, P. Y., Fasman, G. D., Finkelstein, A. V., Lim, V. I., Ptitsyn, O. B., Kabat, E. A., Wu., T. T., Levitt, M., Robson, B., and Nagano, K., 1974a, Comparison of predicted and experimentally determined secondary structure of adenylate kinase, *Nature* **250:** 140–142.

Schulz, G. E., Elzinga, M., Marx, F., and Schirmer, R. H., 1974b, Three-dimensional structure of adenyl kinase, *Nature* **250:**120–123.

Segrest, J. P., and Feldman, R. J., 1977, Amphipathic helices and plasma lipoproteins. A computer study, *Biopolymers* **16:**2053–2065.

Sellers, P., 1974, On the theory and computation of evolutionary distances, *J. Appl. Math.* **26:**787–793.

Sellers, P., 1979, Pattern recognition in genetic sequences, *Proc. Natl. Acad. Sci. U.S.A.* **76:**3041.

Senior, A. E., 1983, Secondary and tertiary structure of membrane proteins involved in proton translocation, *Biochim. Biophys. Acta* **726:**81–95.

Serrano, R., Kiedland-Brandt, M. C., and Fink, G. R., 1986, Yeast plasma membrane ATPase is essential for growth and has homology with $(Na^+ + K^+)$, K^+- and Ca^{2+}-ATPases, *Nature* **319:**689–693.

Sheridan, R. P., Dixon, J. S., Venkataraghavan, R., Kuntz, I. D., and Scott, K. P., 1985, Amino acid composition and hydrophobicity patterns of protein domains correlate with their structure, *Biopolymers* **24:** 1995–2023.

Sheridan, R. P., and Allen, L. C., 1980, The electrostatic potential of the alpha helix (electrostatic potential/ α-helix/secondary structure/helix dipole), *Biophys. Chem.* **11:**133–136.

Sheridan, R. P., Levy, R. M., and Salemme, F. R., 1982, α-Helix dipole model and electrostatic stabilization of 4-α-helical proteins, *Proc. Natl. Acad. Sci. U.S.A.* **79:**4545–4549.

Shin, H.-C., and McFarlane, E. F., 1987, The secondary structure of myelin P_2 protein derived by secondary structure prediction methods, circular dichroism, and 400-MHz ^{1}H-NMR spectroscopy: Implications for tertiary structure, *Biochim. Biophys. Acta* **913:**155–162.

Shinohara, T., Dietzschold, B., Craft, C. M., Wistow, G., Early, J. J., Donoso, L. A., Horowitz, J., and Tao, R., 1987, Primary and secondary structure of bovine retinal S antigen (48 kDa protein), *Proc. Natl. Acad. Sci. U.S.A.* **84:**6975–6979.

Shipman, L. L., and Christoffersen, R. E., 1973, *Ab initio* calculations on large molecules using molecular fragments. Model peptide studies, *J. Am. Chem. Soc.* **95:**1408–1416.

Shoemaker, K. R., Kim, P. S., York, E. J., Stewart, J. M., and Baldwin, R. L., 1987, Tests of the helix dipole model for stabilization of α-helices, *Nature* **326:**563–567.

Shor, N. Z., 1977, Cut-off method with space extension in convex programming problems, *Cybernetics* **12:**94–96.

Shotton, D. M., and Watson, H. C., 1970, Three-dimensional structure of tosyl-elastase, *Nature* **235:**811–816.

Shull, G., Schwartz, A., and Lingrel, J. B., 1985, Amino-acid sequence of the catalytic subunit of the (Na^+ + K^+) ATPase deduced from a complementary DNA, *Nature* **316:**691–695.

Sibanda, B. L., and Thornton, J. M., 1985, β-Hairpin families in globular proteins, *Nature* **316:**170–174.

Simon, I., Nemethy, G., and Scheraga, H. A., 1978, Conformational energy calculations of the effects of sequence variation on the conformations of two tetrapeptides, *Macromolecules* **11:**797–804.

Singh, J., and Thornton, J. M., 1985, The interaction between phenylalanine rings in proteins, *FEBS Lett.* **191:** 1–6.

Sippl, M. J., 1982, On the problem of comparing protein structures. Development and application of a new method for the assessment of structural similarities of polypeptide conformations, *J. Mol. Biol.* **156:**359–388.

Small, D., Chou, P. Y., and Fasman, G. D., 1977, Occurrence of phosphorylated residues in predicted β-turns: Implications for β-turns participation in control mechanisms, *Biochem. Biophys. Res. Commun.* **79:**341–346.

Smith, T. F., and Waterman, M. S., 1981, Identification of common molecular subsequences, *J. Mol. Biol.* **147:**195–197.

Smythies, J. R., 1980, An hypothesis concerning the molecular structure of the nicotinic acetylcholine receptor, *Med. Hypothesis* **6:**943–950.

Sneath, P. H. A., 1966, Relations between chemical structure and biological activity in peptides, *J. Theor. Biol.* **12:**157–195.

Snell, C. R., and Smyth, D. G., 1975, Proinsulin: A proposed three-dimensional structure, *J. Biol. Chem.* **250:** 6291–6295.

Srere, P. A., and Brooks, G. C., 1969, The circular dichroism of glucagon solutions, *Arch. Biochem. Biophys.* **129:**708–710.

Staden, R., 1982, An interactive graphics program for comparing and aligning nucleic acid and amino acid sequences, *Nucleic Acids Res.* **10:**2951–2961.

Steitz, T. A., Goldman, A., and Engelman, D. M., 1982, Quantitative application of the helical hairpin hypothesis to membrane proteins, *Biophys. J.* **37:**124–125.

Stephan, M. M., and Jacobson, G. R., 1986, Membrane disposition of the *Escherichia coli* mannitol permease: Identification of membrane-bound and cytoplasmic domains, *Biochemistry* **25:**8230–8234.

Sternberg, M. J. E., 1983, The analysis and prediction of protein structure, in: *Computing in Biological Sciences* (M. Geisow and A. Barret, eds.), Elsevier, Amsterdam, pp. 143–177.

Sternberg, M. J. E., and Cohen, F. E., 1982, Prediction of the secondary and tertiary structures of interferon from four homologous amino acid sequences, *Int. J. Biol. Macromol.* **4:**137–144.

Sternberg, M. J. E., and Taylor, W. R., 1984, Modelling the ATP-binding site of oncogene products, the epidermal growth factor receptor and related proteins, *FEBS Lett.* **175:**387–392.

Sternberg, M. J. E., and Thornton, J. M., 1976, On the conformation of proteins: The handedness of the β-strand–α-helix–β-strand unit, *J. Mol. Biol.* **105:**367–382.

Sternberg, M. J. E., and Thornton, J. M., 1977a, On the conformation of proteins: The handedness of the connection between parallel β-strands, *J. Mol. Biol.* **110:**269–283.

Sternberg, M. J. E., and Thornton, J. M., 1977b, On the conformation of proteins: An analysis of β-pleated sheets, *J. Mol. Biol.* **110:**285–296.

Sternberg, M. J. E., and Thornton, J. M., 1977c, On the conformation of proteins: Hydrophobic ordering of strands in β-pleated sheets, *J. Mol. Biol.* **115:**1–17.

Sternberg, M. J. E., and Thornton, J. M., 1978, Prediction of protein structure from amino acid sequence, *Nature* **271:**15–20.

Sternberg, M. J. E., Cohen, F. E., Taylor, W. R., and Feldman, R. J., 1981, Analysis and prediction of structural motifs in the glycolytic enzymes, *Phil. Trans. R. Soc. Lond. [Biol.]* **293:**177–189.

Sternberg, M. J. E., Cohen, F. E., and Taylor, W. R., 1982, A combinatorial approach to the prediction of the tertiary fold of globular proteins, *Biochem. J.* **10:**299–301.

Sternberg, M. J. E., Hayes, F. R. F., Russell, A. J., Thomas, P. G., and Ferscht, A. R., 1987, Prediction of electrostatic effects of engineering of protein charges, *Nature* **330:**86–88.

Stroud, R. M., and Finer-Moore, J., 1985, Acetylcholine receptor structure, function and evolution, *Annu. Rev. Cell. Biol.* **1:**317–351.

Stuber, K., Deutscher, J., Sobek, H. M., Hengstenberg, W., and Beyreuther, K., 1985, Amino acid sequence of the amphiphilic phosphocarrier protein factor III^{lac} of the lactose-specific phosphotransferase system of *Staphylococcus aureus, Biochemistry* **24:**1164–1168.

Stuber, M., 1982, Doctoral Thesis, University of Cologne, Cologne, Germany.

Sundaralingam, M., Sekharudu, Y. C., Yathindra, N., and Ravichandran, V., 1987, Ion pairs in alpha-helices, *Proteins: Structure, Function and Genetics* **2:**64–71.

Sweet, R. M., 1986, Evolutionary similarity among peptide segments is a basis for prediction of protein folding, *Biopolymers* **25:**1565–1577.

Sweet, R. M., and Eisenberg, D., 1983, Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure, *J. Mol. Biol.* **171:**479–488.

Tanabe, T., Takeshima, H., Mikami, A., Flockerzi, V., Takahashi, H., Kangawa, K., Kojima, M., Matsuo, H., Hirose, T., and Numa, S., 1987, Primary structure of the receptor for calcium channel blockers from skeletal muscle, *Nature* **328:**313–318.

Tanaka, S., and Scheraga, H. A., 1976a, Statistical mechanical treatment of protein conformation. I. Conformational properties of amino acids in proteins, *Macromolecules* **9:**142–159.

Tanaka, S., and Scheraga, H. A., 1976b, Statistical mechanical treatment of protein conformation. II. A three-state model for specific-sequence copolymers of amino acids, *Macromolecules* **9:**159–167.

Tanaka, S., and Scheraga, H. A., 1976c, Statistical mechanical treatment of protein conformation. III. Prediction of protein conformation based on a three-state model, *Macromolecules* **9:**168–182.

Tanaka, S., and Scheraga, H. A., 1976d, Statistical mechanical treatment of protein conformation. IV. A four-state model for specific-sequence copolymers of amino acids, *Macromolecules* **9:**812–833.

Tanford, C., 1962, Contributions of hydrophobic interactions to the stability of the globular conformation of proteins, *J. Am. Chem. Soc.* **84:**4240–4247.

Tanford, C., 1980, *The Hydrophobic Effect,* 2nd ed., John Wiley & Sons, New York.

Taylor, W. R., 1984, An algorithm to compare secondary structure predictions, *J. Mol. Biol.* **173:**512–514.

Taylor, W. R., 1986a, Identification of protein sequence homology by consensus template alignment, *J. Mol. Biol.* **188:**233–258.

Taylor, W. R., 1986b, The classification of amino acid conservation, *J. Theor. Biol.* **119:**205–218.

Taylor, W. R., and Geisow, M. J., 1987, Predicted structure for the calcium-dependent membrane-binding proteins, p35, p36, p32, *Protein Eng.* **1:**183–187.

Taylor, W. R., and Thornton, J. M., 1983, Prediction of super-secondary structure in proteins, *Nature* **301:** 540-542.

Taylor, W. R., and Thornton, J. M., 1984, Recognition of super-secondary structures in proteins, *J. Mol. Biol.* **173:**487–514.

Thornton, J. M., and Chakauya, B. L., 1982, Conformation of the terminal regions in proteins, *Nature* **298:** 296–297.

Thornton, J. M., and Sibanda, B. L., 1983, Amino and carboxyl-terminal regions in globular proteins, *J. Mol. Biol.* **167:**443–460.

Tillinghast, J. P., Behlke, M. A., and Loh, D. Y., 1986, Structure and diversity of the human T-cell receptor β-chain variable region genes, *Science* **233:**879–883.

Titani, K., Hermodson, M. A., Ericsson, C. H., Walsh, K. A., and Neurath, H., 1982, Amino acid sequence of thermolysin, *Nature [New Biol.]* **238:**35–37.

Titani, K., Takio, K., Handa, M., and Ruggeri, Z. M., 1987, Amino acid sequences of the von Willebrand factor-binding domain of platelet membrane glycoprotein Ib, *Proc. Natl. Acad. Sci. U.S.A.* **84:** 5610–5614.

Toda, M., Takahashi, H., Tanabe, T., Toyosato, M., Furutani, Y., Hirose, T., Asai, M., Inayama, S., Miyata, T., and Numa, S., 1982, Primary structure of α-subunit precursor of *Torpedo californica* acetylcholine receptor deduced from cDNA sequence, *Nature* **299**:793–797.

Toh, H., Hayashida, H., and Miyata, T., 1983, Sequence homology between retroviral reverse transcriptase and putative polymerases of hepatitis B virus and cauliflower mosaic virus, *Nature* **305**:827–829.

Toitskii, G. V., and Zav'yalov, V. P., 1972, Calculation of the conformations of proteins with the aid of a modified nonagram. Establishment of the interrelationship between the primary and secondary structures of the polypeptide chain, *J. Mol. Biol.* **6**:645–647.

Trewhella, H., Anderson, S., Fox, R., Gogol, R., Khan, S., Engelman, D., and Zaccai, G., 1983, Assignment of segments of the bacteriorhodopsin sequence to positions in the structural map, *Biophys. J.* **42**:233–241.

Trewhella, J., Gogol, E., Zaccai, G., and Engelman, D. M., 1984, Neutron diffraction studies of bacteriorhodopsin structure, in: *Neutrons in Biology* (B. P. Schoenborn, ed.), Plenum Press, New York, pp. 227–246.

Ullrich, A., Bell, J. R., Chen, E. Y., Herrera, R., Petruzzelli, L. M., Dull, T. J., Gray, A., Coussens, L., Liao, Y.-C., Tsubokawa, M., Mason, A., Seeburg, P. H., Grunfeld, C., Rosen, O. M., and Ramachandran, J., 1985, Human insulin receptor and its relationship to the tyrosine kinase family of oncogenes, *Nature* **313**:756–761.

Van Belle, D., Couplet, I., Prevost, M., and Wodak, S., 1987, Calculations of electrostatic properties in proteins. Analysis of contributions from induced protein dipoles, *J. Mol. Biol.* **198**:721–735.

van Duijnen, P. T., Thole, B. T., and Hol, W. G. J., 1979, On the role of the active site helix in papain. An *ab initio* molecular orbital study, *Biophys. Chem.* **9**:273–280.

Varghese, J. N., Laver, W. G., and Colman, P. M., 1983, Structure of the influenza virus glycoprotein antigen neuraminidase at 2.9Å resolution, *Nature* **303**:35–40.

Venkatachalan, C. M., 1968, Stereochemical criteria for polypeptides and proteins. V. Conformation of a system of three linked peptide units, *Biopolymers* **6**:1425–1436.

Vickery, L. E., 1987, Interactive analysis of protein structure using a microcomputer spread sheet, *Trends Biochem. Sci.* **12**:37–39.

Visser, L., and Blout, E. R., 1971, Elastase. II. Optical properties and the effects of sodium dodecyl sulfate, *Biochemistry* **10**:743–752.

Vogel, H., and Jähnig, F., 1986, Models for the structure of outer-membrane proteins of *Escherichia coli* derived from Raman spectroscopy and prediction methods, *J. Mol. Biol.* **190**:191–199.

Vogel, H., Wright, J. K., and Jähnig, F., 1985, The structure of the lactase permease derived from Raman spectroscopy and prediction methods, *EMBO J.* **4**:3625–3631.

Vogel, S., Freist, W., and Hoppe, J., 1986, Assignment of conserved amino acid residues to the ATP site in the protein kinase domain of the receptor for epidermal growth factor, *Eur. J. Biochem.* **154**:529–532.

Vondervisztt, F., and Simon, I., 1986, A possible way for prediction of domain boundaries in globular proteins from amino acid sequence, *Biochem. Biophys. Res. Commun.* **139**:11–17.

Vondervisztt, F., Matrai, G., and Simon, I., 1986, Characteristic sequential residue environment of amino acids in proteins, *Int. J. Peptide Protein Res.* **27**:483–492.

von Heijne, G., 1981a, On the hydrophobic nature of signal sequences, *Eur. J. Biochem.* **116**:419–422.

von Heijne, G., 1981b, Membrane proteins. The amino acid composition of membrane-penetrating segments, *Eur. J. Biochem.* **120**:275–278.

von Heijne, G., and Blomberg, C., 1977, The β-structure: Inter-strand correlations, *J. Mol. Biol.* **117**:821–824.

von Heijne, G., and Blomberg, C., 1978, Some global β-sheet characteristics, *Biopolymers* **7**:2033–2037.

von Heijne, G., and Blomberg, C., 1979, Trans-membrane translocation of proteins. The direct transfer model, *Eur. J. Biochem.* **97**:175–181.

Wada, A., 1976, The α-helix as an electric macro-dipole, *Adv. Biophysics* **9**:1–63.

Wada, A., and Nakamura, H., 1981, Nature of the charge distribution in proteins, *Nature* **293**:757–758.

Walker, J. E., Crane, A. F., and Schmitt, H., 1979, The topology of the purple membrane, *Nature* **278**:653–654.

Walker, J. E., Saraste, M., Runswick, M. J., and Gay, N. J., 1982, Distantly related sequences in the α- and β-subunits of ATP synthase, myosin, kinases, and other ATP-requiring enzymes and a common nucleotide binding fold, *EMBO J.* **1**:945–951.

Walker, J. E., Saraste, M., and Gay, N. J., 1984, The *unc* operon. Nucleotide sequence, regulation and structure of ATP-synthase, *Biochim. Biophys. Acta* **768**:164–200.

Wallace, B. A., Cascio, M., and Mielke, D. L., 1986, Evaluation of methods for the prediction of membrane protein secondary structure, *Proc. Natl. Acad. Sci. U.S.A.* **83**:9423–9427.

Warshel, A., and Russell, S. T., 1984, Calculations of electrostatic interactions in biological systems and in solution, *Quart. Rev. Biophys.* **17**:283–422.

Warwicker, J., and Watson, H. C., 1982, Calculation of the electric potential in the active site cleft due to α-helix dipoles, *J. Mol. Biol.* **157**:671–679.

Waterman, M. S., Smith, T. F., and Beyer, W. A., 1976, Some biological sequence metrics, *Adv. Math.* **20**: 367–387.

Weatherford, D. W., and Salemme, F. R., 1979, Conformations of twisted parallel β-sheets and the origin of chirality in protein structures, *Proc. Nat. Acad. Sci. U.S.A.* **76**:19–23.

Weber, P. C., and Salemme, F. R., 1980, Structural and functional diversity in 4-α-helical proteins, *Nature* **287**:82–84.

Webster, T. A., Lathrop, R. H., and Smith, T. F., 1987, Prediction of a common structural domain in amino acid-tRNA synthetases through use of a new pattern-directed inference system, *Biochemistry* **26**:6950–6957.

Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S., and Weiner, P., 1984, A new force field for molecular mechanical simulation of nucleic acids and proteins, *J. Am. Chem. Soc.* **106**:765–784.

Wertz, D. H., and Scheraga, H. A., 1978, Influence of water on protein structure. An analysis of the preferences of amino acids residues for the inside or outside and for specific conformations in a protein molecule, *Macromolecules* **11**:9–15.

Wetlaufer, D. B., 1973, Nucleation, rapid folding, and globular intrachain regions in proteins, *Proc. Natl. Acad. Sci. U.S.A.* **70**:697–701.

Wetlaufer, D. B., 1981, Folding of protein fragments, *Adv. Protein Chem.* **34**:61–92.

Wetlaufer, D. B., and Ristow, S., 1973, Acquisition of three-dimensional structure of proteins, *Annu. Rev. Biochem.* **42**:135–158.

Wickner, W., 1979, The assembly of proteins into biological membranes: The membrane trigger hypothesis, *Annu. Rev. Biochem.* **48**:23–45.

Wierenga, R. K., Terpstra, P., and Hol, W. G. J., 1986, Prediction of the occurrence of the ADP-binding βαβ-fold in proteins, using an amino acid sequence fingerprint, *J. Mol. Biol.* **187**:101–107.

Wilbur, W. J., and Lipman, D. J., 1983, Rapid similarity searches of nucleic acids and protein data banks, *Proc. Natl. Acad. Sci. U.S.A.* **80**:726–730.

Williams, R. W., Chang, A., Juretic, D., and Loughran, S., 1987, Secondary structure predictions and medium-range interactions, *Biochim. Biophys. Acta* **916**:200–204.

Wilmot, C. M., and Thornton, J. M., 1988, Analysis and prediction of the different types of β-turn in proteins, *J. Mol. Biol.* **203**:221–232.

Wilson, I. A., Haft, D. H., Getzoff, E. D., Tainer, J. A., Lerner, R. A., and Brenner, S., 1985, Identical short peptide sequences in unrelated proteins can have different conformations: A testing ground for theories of immune recognition, *Proc. Natl. Acad. Sci. U.S.A.* **82**:5255–5259.

Wodak, S. J., and Janin, J., 1980, Analytical approximation to the accessible surface area of proteins, *Proc. Natl. Acad. Sci. U.S.A.* **77**:1736–1740.

Wodak, S. J., and Janin, J., 1981, Location of structural domains in proteins, *Biochemistry* **20**:6544–6552.

Wolfenden, R., 1983, Waterlogged molecules, *Science* **222**:1087–1093.

Wolfenden, R., Anderson, L., Cullis, P. M., and Southgate, C. C. B., 1981, Affinities of amino acid side chains for solvent water, *Biochemistry* **20**:849–855.

Wolfenden, R. V., Cullis, P. M., and Southgate, C. C. F., 1979, Water, protein folding, and the genetic code, *Science* **206**:575–577.

Wu, C.-S. C., Hachimori, A., and Yang, J. T., 1982, Lipid induced ordered conformation of some peptide hormones and bioactive oligopeptides: Predominance of helix over β-form, *Biochemistry* **21**:4556–4562.

Wu, T. T., and Kabat, E. A., 1970, An analysis of the sequences of the variable regions of Bence Jones proteins and Myeloma light chains and their implications for antibody complementarity, *J. Expt. Med.* **132**:211–250.

Yatsunami, K., and Khorana, H. G., 1985, GTPase of bovine rod outer segments: The amino acid sequence of the α-subunit as derived from the c-DNA sequence, *Proc. Natl. Acad. Sci. U.S.A.* **82**:4316–4320.

Yockey, H. P., 1977, A prescription which predicts functionally equivalent residues at given sites in protein sequences, *J. Theor. Biol.* **67**:337–343.

Youvan, D. C., Bylina, E. J., Albert, M., Begusch, H., and Hearest, J., 1984, Nucleotide and deduced polypeptide sequence of the photosynthetic reaction center, B870 antenna, and flanking polypeptides from *R. capsulata, Cell* **37**:949–957.

Yuschok, T. J., and Rose, G. D., 1983, Hierarchic organization of globular proteins. A control study, *Int. J. Peptide Prot. Res.* **21**:479–484.

Zehfus, M. H., Seltzer, J. P., and Rose, G. D., 1985, Fast approximation for accessible surface area and molecular volume of protein segments, *Biopolymers* **24**:2511–2519.

Zimm, B. H., and Bragg, J. K., 1959, Theory of the phase transition between the helix and random chain in polypeptide chains, *J. Chem. Phys.* **31**:526–535.

Zimmerman, J. M., Eliezer, N., and Simha, R., 1968, The characterization of amino acid sequences in proteins by statistical methods, *J. Theor. Biol.* **21**:170–201.

Zimmerman, S. S., and Scheraga, H. A., 1977, Local interactions in bends of protein, *Proc. Natl. Acad. Sci. U.S.A.* **74**:4126–4129.

Zimmerman, S. S., Pottle, M. S., Nemethy, G., and Scheraga, H. A., 1977, Conformational analysis of the 20 naturaliy occurring amino acid residues using ECEPP, *Macromolecules* **10**:1–9.

Zvelebil, M. J., Barton, G. J., Taylor, W. R., and Sternberg, M. J. E., 1987, Prediction of protein secondary structure and active sites using alignment of homologous sequence, *J. Mol. Biol.* **195**:957–961.

X. APPENDIXES

Appendix 1: List of Reviews on Protein Folding and Prediction of Secondary and Tertiary Structure

Argos, P., and Mohana Rao, J. K., 1986, Prediction of protein structure, *Methods Enzymol.* **130:**185–207.

Bajaj, M., and Blundell, T., 1984, Evolution and the tertiary structure of proteins, *Annu. Rev. Biophys. Biophys. Chem.* **13:**453–492.

Blake, C. C. F., and Johnson, L. N., 1984, Protein structure, *Trends Biochem. Sci.* **9:**147–151.

Blundell, T., and Sternberg, M. J. E., 1985, Computer-aided design in protein engineering, *Trends Biotechnol.* **3:**228–235.

Blundell, T. L., Sibanda, B. L., Sternberg, M. J. E., and Thornton, J. M., 1987, Knowledge-based prediction of protein structures and the design of novel molecules, *Nature* **326:**347–352.

Cantor, C. R., and Schimmel, P. R., 1980, *Biophysical Chemistry,* Volume I, W. H. Freeman, San Francisco.

Chothia, C., 1984, Principles that determine the structure of proteins, *Annu. Rev. Biochem.* **53:**537–572.

Chou, P. Y., and Fasman, G. D., 1977, Secondary structural prediction of proteins from their amino acid sequence, *Trends Biochem. Sci.* **2:**128–132.

Chou, P. Y., and Fasman, G. D., 1978a, Empirical predictions of protein conformation, *Annu. Rev. Biochem.* **47:**251–276.

Chou, P. Y., and Fasman, G. D., 1978b, Prediction of the secondary structure of proteins from their amino acid sequence, *Adv. Enzymol.* **47:**45–108.

Creighton, T. E., 1983, *Proteins,* W. H. Freeman, New York.

Doolittle, R. F., 1986, *Of URFS and ORFS. A Primer on How to Analyse Derived Amino Acid Sequences,* University Science Books, Mill Hill, CA.

Edsall, J. T., and McKenzie, H. A., 1983, Water and proteins II. The location and dynamics of water in protein systems and its relation to their stability and properties, *Adv. Biophys.* **16:**53–183.

Eisenberg, D., 1984, Three-dimensional structure of membrane and surface proteins, *Annu. Rev. Biochem.* **53:** 595–623.

Engelman, D. M., Steitz, T. A., and Goldman, A., 1986, Identifying nonpolar transbilayer helices in amino acid sequences of membrane proteins, *Annu. Rev. Biophys. Biophys. Chem.* **15:**321–353.

Fasman, G. D., 1980, Prediction of protein conformation from the primary structure, *Ann. N.Y. Acad. Sci.* **348:** 147–159.

Fasman, G. D., 1985, A critique of the utility of the prediction of protein secondary structure, *J. Biosci.* **8:**15–23.

Fasman, G. D., 1987, The road from poly-α-amino acids to the prediction of protein conformation, *Biopolymers* **26:**S59–S79.

Fletterick, R., and Zoller, M., eds., 1986, Computer graphics and molecular modeling, in: *Current Communications in Molecular Biology,* Cold Spring Harbor Laboratory, Cold Spring Harbor, NY.

Ghelis, C., and Yon, J., 1982, *Protein Folding,* Academic Press, New York.

Go, N., 1983, Theoretical studies of protein folding, *Annu. Rev. Biophys. Biophys. Chem.* **12:**183–210.

Hohne, E., and Kretschmer, R. G., 1985, Description of secondary structure in proteins, *Stud. Biophys.* **108:** 165–186.

Honig, B. H., Hubbell, W. L., and Flewling, R. F., 1986, Electrostatic interactions in membranes and proteins, *Annu. Rev. Biophys. Biophys. Chem.* **15:**163–193.

Jaenicke, R., ed., 1984, *Protein Folding,* Elsevier/North-Holland Biomedical Press, Amsterdam.

Jaenicke, R., 1987, Folding and association of proteins, *Prog. Biophys. Mol. Biol.* **49:**117–237.

Janin, J., and Wodak, S. J., 1983, Structural domains in proteins and their role in the dynamics of protein function, *Prog. Biophys. Mol. Biol.* **42:**21–78.

Jungck, J. R., Friedman, R. M., 1984, Mathematical tools for molecular genetics data: An annotated bibliography, *Bull. Math. Biol.* **46:**699–744.

Kauzmann, W., 1959, Some factors in the interpretation of protein denaturation, *Adv. Protein Chem.* **14:**1–63.

Kneale, G. G., and Bishop, M. J., 1985, Nucleic acid and protein sequence databases, *Cabios* **1:**11–17.

Kollman, P., 1987, Molecular modeling, *Annu. Rev. Phys. Chem.* **38:**303–316.

Lesk, A. M., and Hardman, K. D., 1985, Computer-generated pictures of proteins, *Methods Enzymol.* **115:** 381–390.

Levitt, M., 1982, Protein conformation, dynamics, and folding by computer simulation. *Annu. Rev. Biophys. Bioeng.* **11:**251–271.

Matthew, J. B., and Gurd, F. R. N., 1986a, Stabilization and destabilization of protein structure by charge interactions, *Methods Enzymol.* **130:**437–453.

Matthew, J. B., and Gurd, F. R. N., 1986b, Calculation of electrostatic interactions in proteins, *Methods Enzymol.* **130:**413–436.

Matthews, J. B., 1985, Electrostatic effects in proteins, *Annu. Rev. Biophys. Biophys. Chem.* **14:**387–417.

Nagano, K. and Ponnuswamy, P. K., 1984, Prediction of packing of secondary structure, *Adv. Biophys.* **18:** 115–148.

Nemethy, G., and Scheraga, H. A., 1977, Protein folding, *Q. Rev. Biophys.* **10:**239–352.

Ptitsyn, O. B., and Finkelstein, A. V., 1980, Similarities of protein topologies: Evolutionary divergence, functional convergence or principles of folding, *Q. Rev. Biophys.* **13:**339–386.

Richards, F. M., 1977, Areas, volumes, packing, and protein structure, *Adv. Biophys. Bioeng.* **6:**151–176.

Richards, J. S., 1985, Schematic drawings of protein structures, *Methods Enzymol.* **115:**359–380.

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

Richardson, J. S., 1985, Describing patterns of protein tertiary structure, *Methods Enzymol.* **115:**341–358.

Robson, B., 1982, The prediction of molecular conformation, *Biochem. J.* **10:**297–298.

Robson, R., and Garnier, J., 1986, *Introduction to Proteins and Protein Engineering,* Elsevier, Amsterdam.

Rose, G. D., Gierasch, L. M., and Smith, J. A., 1985, Turns in peptides and proteins, *Adv. Protein Chem.* **37:** 1–109.

Rossman, M. G., and Argos, P., 1981, Protein folding, *Annu. Rev. Biochem.* **53:**497–533.

Salemme, F. R., 1983, Structural properties of protein β-sheets, *Prog. Biophys. Mol. Biol.* **42:**95–133.

Scheraga, H. A., 1985, Calculations of the three-dimensional structure of proteins, *Ann. N.Y. Acad. Sci.* **439:** 170–194.

Schulz, G. E., 1977, Structural rules for globular proteins, *Angew. Chem. [Engl.]* **16:**23–32.

Schulz, G. E., and Schirmer, R. H., 1979, *Principles of Protein Structure,* Springer-Verlag, Berlin.

Sternberg, M. J. E., 1983, The analysis and prediction of protein structure, in: *Computing in Biological Sciences* (M. S. Geisow and A. N. Barrett, eds.), Elsevier Biomedical Press, Amsterdam, pp. 143–177.

Sternberg, M. J. E., 1986, Prediction of protein structure from amino acid sequence, *Anticancer Drug Design* **1:** 169–178.

Sternberg, M. J. E., and Thornton, J. M., 1978, Prediction of protein structure from amino acid sequence, *Nature* **271:**15–20.

Taylor, W. R., 1987, Protein structure prediction in: *Nucleic Acid and Protein Sequence Analysis* (M. J. Bishop and G. J. Rawlings, eds.), IRL Press: Oxford.

von Heijne, G., 1987, *Sequence Analysis in Molecular Biology,* Academic Press, New York.

Warshel, A., and Russell, S. T., 1984, Calculations of electrostatic interactions in biological systems and in solutions, *Q. Rev. Biophys.* **17:**283–422.

Wetlaufer, D. B., 1981, Folding of protein fragments, *Adv. Protein Chem.* **34:**335–347.

Wetlaufer, D. B., ed., 1984, *The Protein Folding Problem,* AAAS, Washington.

Wetlaufer, D. B., and Ristow, S., 1973, Acquisition of three-dimensional structure of proteins, *Annu. Rev. Biochem.* **42:**135–158.

Appendix 2: Programs Available through This Publication for Protein Secondary Structure Prediction

Deposited at:
1. The Protein Identification Resource (PIR), National Biomedical Research Foundation, Georgetown University Medical Center, 3900 Reservoir Road, NW, Washington, D.C. 20007
2. Molecular Biology Computer Research Resource (MBCRR), Dana Farber Cancer Institute, 44 Binney Street, Boston, Massachusetts 02115
3. Whitehead Institute for Biochemical Research, 9 Cambridge Center, Cambridge, Massachusetts 02142, c/o W. Gilbert

1. Chou–Fasman–Prevelige derived from the original Chou–Fasman algorithm [C-F-P]:
   a. Chou, P. Y., and Fasman, G. D., 1974, Prediction of protein conformation, *Biochemistry* **13**:222–245.
   b. Chou, P. Y., and Fasman, G. D., 1978, Prediction of the secondary structure of proteins from their amino acid sequence, *Adv. Enzymol.* **47**:45–148.
   c. Chou, P. Y., and Fasman, G. D., 1979, Prediction of β-turns, *Biophys. J.* **26**:367–384.
   d. Chou, P. Y., Fasman, G. D., and Prevelige, P., Chapters 9 and 12, this volume.

   Written in C for IBM-PC-XT.

2. Deléage, F., Tinland, B., and Roux, B., 1987, A computerized version of the Chou and Fasman method for predicting the secondary structure of proteins, *Anal. Biochem.* **163**:292–297. [D-T-R]

   Some of the qualitative rules in the original rules have been converted to numeric scales to obtain unambiguous predictions.
   Written for an Apple IIe (128k) microcomputer.

3. Eisenberg, D., Wesson, M., and Wilcox, W., Chapter 16, this volume. [E]

   Written in FORTRAN to be used on a Vax computer.

4. Finer-Moore, J., and Stroud, R. M., 1984, Amphipathic analysis and possible formation of the ion channel in an acetylcholine receptor, *Proc. Natl. Acad. Sci. U.S.A.* **81**:155–159. [F-M-S]
   Finer-Moore, J., Bazan, F., Rubin, J., and Stroud, R. M., 1989, Identification of membrane proteins and soluble protein secondary structural elements, domain structure, and packing arrangements by Fourier-transform amphipathic analysis, Chapter 19, this volume.

   Written in FORTRAN for use on a VAX computer on a VMS operating system.

5. Garnier, J., Osguthorpe, D. G., and Robson, B., 1978, Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins, *J. Mol. Biol.* **120**:97–120. [G-O-R]
   and updated:
   Gibrat, J.-F., Garnier, J., and Robson, B., 1987, Further developments of protein secondary structure prediction using information theory, new parameters, and consideration of residue pairs, *J. Mol. Biol.* **198**:425–443.
   Garnier, J., and Robson, B., 1989, The G-O-R method for predicting secondary structure in proteins, Chapter 10, this volume.

Written in FORTRAN for use on a Micro VAX II computer. Another program is available to be run on a microcomputer (e.g., IBM PC).

6. Vogel, H., Wright, J. K., and Jähnig, F., 1985, The structure of the lactose permease derived from raman spectroscopy and prediction methods, *EMBO J.* **4:**3625–3631. [J]
   Vogel, H., and Jähnig, F., 1986, Models for the structure of outer-membrane proteins of *Escherichia coli* derived from raman spectroscopy and prediction methods, *J. Mol. Biol.* **190:**191–199.
   Jähnig, F., 1989, Structure prediction for membrane proteins, Chapter 18, this volume.

   Written in FORTRAN for use on an IBM PC/AT computer.

7. Klein, P., 1986, Prediction of protein structural class by discriminant analysis. *Biochim. Biophys. Acta* **874:**205–215. [K]

   Program STRCLS, written for VAX/VMS in FORTRAN.

8. Klein, P., Kanehisa, M., and DeLisi, C., 1985, The detection and classification of membrane-spanning regions, *Biochim. Biophys. Acta* **815:**468–476. [K-K-D]

   Written in FORTRAN.

9. Kyte, J., and Doolittle, R. F., 1982, A simple method for displaying the hydropathic character of a protein, *J. Mol. Biol.* **157:**105–132.* [K-D]

   Program SOAP, written in language C for use in the software system Unix Vax with a C compiler (K-D: Program 8). Will send other programs on 1600 bpi tape.

   Doolittle Programs: Protein sequence alignment and phylogenetic tree construction. D.-Feng and R. F. Doolittle, 1987, Progressive sequence alignment as a prerequisite to correct phyogenetic trees, *J. Mol. Evol.* **23:**351.

   Seven programs: format.c–for DNA or protein
   Score.c–for nearest relationships
   prealign.c
   dfalign.c
   blen.c
   mulpub.c
   dfplot.c

   (The .c indicates that the programs are written in C language. All these programs are in their uncompiled form. Instructions are given to compile the C programs.)

10. Lim, V. I., 1974, Algorithms for prediction of α-helical and β-structural regions in globular proteins, *J. Mol. Biol.* **88:**873–894.

    Programs written by: Johannes A. Lenstra, Vakgroep Infectieziekten en Immunologie, Facultair Recombinant DNA Laboratorium Fakulteit Der Diergeneeskunde, Rijksuniversiteit Te, Utrecht, Yalelaan 1, Postbus 80-165, 3508 TD Utrecht, The Netherlands, and
    Kabsch, W. and Sander, C., Biophysics Department, Max Planck Institute of Medical Research, D-6900 Heidelberg, Federal Republic of Germany.

*Data-sieving program can be used. Based on a running median (between 5 and 19 amino acids) smooths the raw data, rendering the domain more visible. Bangham, J. A., 1988, Data-sieving hydrophobicity plots, *Anal. Biochem.* **174:**142–145.

11. Nagano, K., 1973, Logical analysis of the mechanism of protein folding. I. Prediction of helices, loops and β-structures from primary structure, *J. Mol. Biol.* **75:**401–420. [N]

    Nagano, K., 1974, Logical analysis of the mechanism of protein folding. II. The nucleation process, *J. Mol. Biol.* **84:**337–372.

    Nagano, K., 1977a, Logical analysis of the mechanism of protein folding. IV. Super-secondary structure, *J. Mol. Biol.* **109:**235–250.

    Nagano, K., 1977b, Triplet information in helix prediction applied to the analysis of super-secondary structures, *J. Mol. Biol.* **109:**251–274.

    Nagano, K., 1980, Logical analysis of the mechanism of protein folding. V. Packing game simulation of α/β proteins, *J. Mol. Biol.* **138:**797–832.

    Nagano, K., and Ponnuswamy, P. K., 1984, Prediction of packing of secondary structure, *Adv. Biophys.* **18:**115–148.

    Nagano, K., 1989, Prediction of packing of secondary structure, Chapter 11, this volume.

    Written in FORTRAN for use with an HITAC M-682H/680 computer system; compatible with the IBM 370 series computer.

12. Rose, G. D., Geselowitz, A. R., Lesser, G. J., Lee, R. H., and Zehfus, M. H., 1985, Hydrophobicity of amino acids in globular proteins, *Science* **229:**834–838. [R-D]

    Dworkin, J. E., and Rose, G. D., 1987, Hydrophobicity profiles revisited, in *Methods in Protein Sequence Analysis* (K. A. Walsh, ed.), Humana Press, Clinton, New Jersey, pp. 573–586.

    Rose, G. D., and Dworkin, J. E., 1989, The hydrophobicity profile, Chapter 15, this volume.

    Written in FORTRAN for use on a VAX or MICROVAX computer on a VMS operating system.

Appendix 3. Commercially Available Programs

1. HIBIO-PROSIS

   Secondary structure prediction:

   1. Chou, P. Y., and Fasman, G. D., 1978, *Adv. Enzymol.* **47**:45–148; 1978, *Annu. Rev. Biochem.* **47**:251–276.
   2. Garnier, J., Osguthorpe, D. J., and Robson, B., 1978, *J. Mol. Biol.* **120**:97–120.

   Hydrophobicity:

   1. Kyte, J., and Doolittle, R. F., 1982, *J. Mol. Biol.* **157**:105–132.
   2. Hopp, T. P., and Woods, K. R., 1981, *Proc. Natl. Acad. Sci. U.S.A.* **78**:3824–3828.
   3. Rose, G., 1978, *Nature* **272**:586–590.

   Written in C for use on any IBM-XT,AT microcomputer.

   Available from: Pharmacia LKB Biotechnology, 800 Centennial Avenue, P.O. Box 1327, Piscataway, New Jersey 08855-1327.

2. MSEQ: A Microcomputer-Based Approach to the Analysis, Display and Prediction of Protein Structure. Black, S. D., and Gloriso, J. C., 1986, *Bio Techniques* **4**:448–460.

   Chou, P. Y., and Fasman, G. D., 1978, Secondary structure prediction, *Annu. Rev. Biochem.* **47**:251–276.

   Hydrophobicity — 4 scales:

   1. Argos, P., and Palau, J., 1982, *Int. J. Peptide Prot. Res.* **19**:380–393.
   2. von Heijne, G., 1981, *Eur. J. Biochem.* **116**:419–422.
   3. Hopp, T. P., and Woods, K. R., 1981, *Proc. Natl. Acad. Sci. U.S.A.* **78**:3824–3828.
   4. Kyte, J., and Doolittle, R. F., 1982, *J. Mol. Bio.* **157**:105–132.

   Hydrophobic moments:

   Eisenberg, D., Weiss, R. M., and Terwilliger, T. C., 1982, *Nature* **299**:371–374; *Proc. Natl. Acad. Sci. U.S.A.* **81**:140–144.

   Graphic cartoons.

   Written in Basic for use on an IBM microcomputer family IBM-PC XT.

   Contact: Mr. Fred Reinhardt, University of Michigan Software, Intellectual Properties Office, 225W Engineering, Ann Arbor, Michigan 48109–1092.

3. NEWAT 85

   Protein sequence data base and programs.

   Categorized phylogenetically: *E. coli,* other prokaryotes, DNA viruses, eykaryotes (except vertebrate animals), vertebrates (except human), human and viruses.

   Software to enter new sequences: search for homologies between a sequence and the data base and to align a pair of sequences.

   Programs for displaying the hydropathy of a protein sequence; translating DNA sequences into putative amino acid sequences.

Written for IBM-PC.

Available from: Newat Distribution Co., Inc., P.O. Box 12822, La Jolla, California 92037.

4. PEP: Analyzing protein sequences.

   Reverse translates peptides and indicates ambiguity due to codon preferences or to the degeneracy of the genetic code.

   Identifies the least ambiguous regions of a peptide for making hybridization probes of different lengths.

   Simulates and maps protease digestion fingerprints.

   Determines hydropathicity to predict antigenic sites or membrane-binding regions.

   Predicts and maps protein secondary structure with the Chou-Fasman algorithm.

   Allows variably set amino acid equivalencies for similarity searches and alignments.

   Written in Mainsail for use on the VMS VAX, Microvax II comuters and Sunwork Stations.

   Available from: Intelligenetics, 1975 El Camino Road, Mountain View, California 94040-2216.

5. PEPPLOT

   Gribskov, M., Burgess, R. R., and Devereux, J., 1986, PEPPLOT: A protein secondary structure analysis program for the UWGCG sequence analysis software package.

   Structure Prediction:

   1. Chou, P. Y., and Fasman, G. D., 1978, Secondary structure prediction, *Adv. Enzymol.* **47**:45–147.
   2. Garnier, J., Osguthorpe, D. J., and Robson, B., 1978, Secondary structure prediction, *J. Mol. Biol.* **120**:97–120.
   3. Kyte, J., and Doolilttle, R. F., 1982, Hydropathy profile, *J. Mol. Biol.* **157**:105–132.
   4. Eisenberg, D., Sweet, R. M., and Terwilliger, T. C., 1984, Hydrophobic moment, *Proc. Natl. Acad. Sci. U.S.A.* **81**:140–144.

   Written in FORTRAN 77 for use on a Vax computer running version 3 or 4 of VMS.

   Available from: John Devereux, University of Wisconsin, Biotechnology Center, 1710 University Avenue, Madison, Wisconsin 53705.

Appendix 4: Relevant Programs Described in the Literature

Arnold, J., Eckerrode, U. K., Lemke, J., Phillips, G. J., and Schaeffer, S. W., 1986, A comprehensive package for DNA sequence analysis in FORTRAN IV for the PDP-11, *Nucleic Acids Res.* **14:**239–254.

Klein, P., and DeLisi, C., 1986, Prediction of protein structural class from the amino acid sequence, *Biopolymers* **25:**1659–1672.

Mount, D. W., 1986, Improved programs for DNA and protein sequence analysis on the IBM personal computer and other standard computer systems, *Nucleic Acids Res.* **14:**443–454.

Nakashima, H., Nishikawa, K., and Ooi, T., 1986, The folding type of a protein is relevant to the amino acid composition, *J. Biochem. (Tokyo)* **99:**153–162.

Nishikawa, K., and Ooi, T., 1986, Amino acids sequence homology applied to the prediction of protein secondary structures, and joint prediction with existing methods, *Biochim. Biophys. Acta* **871:**45–54.

Novotny, J., and Auffray, C., 1984, A program for prediction of protein secondary structure from nucleotide sequence data: Application to histocompatibility antigens, *Nucleic Acids Res.* **12:**243–253. (Editor's note: A combination of Chou, P. Y., and Fasman, G. D., 1978, *Adv. Enzymol.* **47:**45–148 and Rose, G. D., and Roy, S., 1980, *Proc. Natl. Acad. Sci. U.S.A.* **77:**4643–4647)

Peltola, H., Soderlund, H., and Ukkonen, E., 1986, Algorithms for the search of amino acid patterns in nucleic acid sequences, *Nucleic Acids Res.* **14:**99–107.

Reisner, A. H., and Bucholtz, C. A., 1986, The MTX package of computer programs for the comparison of sequences of nucleotides and amino acid residues, *Nucleic Acids Res.* **14:**233–238.

Staden, R., 1986, The current status and portability of our sequence handling software, *Nucleic Acids Res.* **14:** 217–231.

Taylor, P., 1986, A computer program for translating DNA sequences in protein, *Nucleic Acids Res.* **14:**437–441.

Trifonov, E. D., and Brendel, V., 1986, *GNOMIC. A Dictionary of Genetic Codes,* Balaban Publishers, Philadelphia.

van der Berg, J. A., and Osinga, M., 1986, A peptide to DNA conversion program, *Nucleic Acids Res.* **14:**137–140.

Review Articles

Moore, J., Engelberg, A., and Bairoch, A., 1988, Using PC/GENE for proteins and nucleic acid analysis, *Biotechniques* **6:**566–572.

Roe, B. A., 1988, Computer programs for molecular biology: An overview of DNA sequencing and protein analysis packages, *Biotechniques* **6:**560–565.

Appendix 5. National Resource Data Bases

1. BIONET™: National Computer Resource for Molecular Biology

Smith, D. H., Brutlag, D., Friedland, P., and Kedes, L. H., 1986, *Nucleic Acids Res.* **14:**17–20.

Kirstofferson, D., 1987, *Nature* **325:**555–556.

Available through: IntelliGenetics, 700 East El Camino Real, Mountain View, CA 94040.

IntelliGenetics Software

CLONER:	Recombinant DNA simulation system
DDMATRIX:	Dot matrix sequence similarity program
GEL:	DNA sequencing project management system
GENALIGN:	Multiple sequence alignment program
GENED:	Genetic sequence editor
IFIND:	Sequence similarity and alignment program
MAP:	Restriction map generator and editor
PEP:	Polypeptide sequence analysis system
QUEST:	Biological searching system
SEQ:	DNA sequence analysis system
SIZER:	Fragment length analysis system

2. The EMBL Data Library

Hamm, G. H., and Cameroa, G. N., 1986, The EMBL Data Library, *Nucleic Acids Res.* **14:**5–9.

The EMBL Data Library was the first internationally supported central resource for nucleic acid sequence data. Working in close collaboration with its American counterpart, GenBank, the library prepares and makes available to the scientific community a comprehensive collection of the published nucleic acid sequences.

Available through: European Molecular Biology Laboratory, Meyerhofstrasse 1, 6900 Heidelberg, Federal Republic of Germany.

3. The GenBank® Genetic Sequence Databank

Bilofsky, H. S., Burks, C., Fickett, J. W., Goad, W. B., Lewitter, F. I., Rindone, W., Swindell, C. D., and Tung, C.-S., 1986, The GenBank genetic sequence databank, *Nucleic Acids Res.* **14:**1–4.

The GenBank® Genetic Sequence Data Bank contains over 5700 entries for DNA and RNA sequences that have been reported since 1967. This paper briefly describes the contents of the data base, the forms in which the data base is distributed, and the services offered to scientists who use the GenBank data base.

Available through: GenBank, 700 East El Camino Real, Mountain View, CA 94040.

4. Molecular Biology Computer Research Resource (MBCRR)

Smith, T. F., Grushin, K., Tolman, S., and Faulkner, D., 1986, *Nucleic Acids Res.* **14**:25–29.

Analytic tools:

MASE: Full-screen multiple aligned sequence editor with regular expression highlighter.

LOCAL: Dynamic programming maximum local subsequence alignment algorithm, 1981, *J. Mol. Biol.* **147**:195–197.

PRSTRC: A modified Chou and Fasman protein structure algorithm, Ralph, W. W., Webster, T., and Smith, T. F., 1987, *Cabios* **3**:211–216.

DASHER: A high-speed hash-linked list sequence similarity search tool which ranks identified similarities by chi-square test on the occurrence distribution of common *n*-mers.

GGREP: Regular expression pattern search tool for GenBank and NBFR analogous to UNIX GREP. Employs the regular expression handler from GNU-EMACS.

ARIADNE: A pattern-directed inference and hierarchical abstraction in protein structure recognition, Lathrop, R. H., Webster, T. A., and Smith, T. F., 1987, *Commun. ACM* **30**:909–921.

RZMAP: A branch and bound algorithm to reconstruct restriction maps from double digest lengths, 1983, *Gene* **22**:19–29.

The MBCRR distributes the source, documentation, and MS-DOS executables of various utilities and programs for genetic sequence analyses to the noncommercial scientific community:

1. "Fristensky Package": Brian Fristensky's Cornell DNA sequence analysis programs.
2. "Mount Package": The Genetics PC-Software Center of the University of Arizona sequence analysis tools (developed by D. W. Mount, B. Conrad, and E. Myers).
3. "Lipman/Pearson Package": David Lipman and William Pearson's rapid biosequence similarity analysis code. *Science* **227**:1435–1441.
4. "Shalloway Package": David Shalloway's restriction/functional site data base management program (IBM-compatible executable code only).
5. "Zucker Package": Michael Zucker's RNA secondary structure software.
6. "Caltech Package": Alan Goldin's series of routines to analyze DNA or protein sequence data.

IntelliGenetics Programs (see above).

Available through: Molecular Biology Computer Research Resource, Dana Farber Cancer Institute, 44 Binney Street, Boston, MA 02115.

5. Protein Identification Resource (PIR). Sponsored by the National Biomedical Research Foundation (NBRF)

George, D. G., Barker, W. C., and Hunt, L. T., 1986, The protein identification resource, *Nucleic Acids Res.* **14**:11–15.

The NBRF–PIR Protein Sequence Database and the NBRF Nucleic Acid Sequence Database

are distributed on magnetic tape in VAX/VMX and ASCII card image formats on a quarterly basis. The PSQ (Protein Sequence Query) and NAQ (Nucleic Acid Query) programs for browsing and information retrieval are distributed with the respective data bases in VAX/VMS format. The PIR sequence analysis software is updated approximately once a year. The data bases and software of the PIR are in the public domain and may be freely copied and redistributed provided the Protein Identification Resource is acknowledged as the source.

The NBRF–PIR Protein Sequence Database includes the following: all substantially sequenced proteins, including sequences translated from nucleic acid sequences; bibliographic citations for amino acid sequences, nucleic acid sequences, x-ray crystallography, active site determination, etc.; annotations identifying posttranslational modifications, active site, signal sequences, activation cleavages, disulfide bonds, intron locations, etc. Sequences translated from nucleotide sequences are checked against the author's translations and against sequences reported from protein sequencing. An auxiliary data base includes sequences in preparation as well as additional fragmentary and predicted sequences.

The NBRF Nucleic Acid Sequence Database contains entries annotated to show the locations of protein coding regions. The VAX/VMS nucleotide database tape also includes programs to reformat GenBank® and EMBL nucleotide databases to be accessible by the NAQ program.

The NBRF-PIR Sequence Analysis Software programs run on a VAX-11/780 operating under VAX/VMS version 4.2. All programs are written in VAX-11 FORTRAN (a superset of ANSI FORTRAN 77), with the exception of the Lipman–Pearson programs, which are written in VAX-11 C.

PSQ: Protein sequence query retrieval program.

NAQ: Nucleic acid query retrieval program.

FASTP and FASTN: Programs written by William Pearson of the University of Virginia that use the Lipman–Pearson algorithm (*Science* **227:**1435–1441, 1985) to search the protein and nucleic acid sequence data bases. The programs produce near-optimal alignments of segments identified in the search; they are extremely rapid.

RDF: Lipman–Pearson program for evaluation of RASTP results.

SEARCH: To compare a protein segment of, for instance, 25 residues with every 25-residue segment in the data base (gaps not permitted) using a scoring matrix.

ISEARCH: An interactive adaptation of our SEARCH program that includes a procedure for searching for ambiguous segments in which two or more amino acids may be found at some positions.

ALIGN: Uses a version of the Needleman–Wunsch algorithm to determine the best alignment of two sequences by computing a maximum match score; this score is compared with the average maximum match score from random permutations of the two sequences to derive an alignment score in standard deviation units. ALIGN uses a scoring matrix to assign a value to each pair of aligned amino acids in the sequences being compared; certain scoring matrices increase sensitivity for detecting similarity between distantly related sequences.

IALIGN: Interactive ALIGN program.

RELATE: To compare all segments of a given length from one sequence with all segments of a second one, using a scoring matrix. Statistics can be generated by comparing the results with those from permuted sequences.

DOTMATRIX: Graphic segment comparison display program similar to program RELATE. Output is specific for Printronix 300-line printer.

PRPLOT: A general-purpose program that plots values that can be used for Chou–Fasman-type predictions, such as hydrophilicity, hydrophobicity, β-turn-forming potential, etc., averaged over several contiguous amino acids.

CHOFAS: A secondary structure prediction program with a convenient and compact display developed by M. Kanehisa as part of the IDEAS package.

HYDRO: Hydrophobicity scoring matrix.

Levitt, M., 1976, *J. Mol. Biol.* **104:**59–109.

Available through: National Biomedical Research Foundation. Georgetown University Medical Center, 3900 Reservoir Road NW, Washington, D.C. 20007.

6. Protein Data Bank: A computer-based archival file for macromolecular structures.

Bernstein, F. C., Koeyzle, T. F., Williams, G. J. B., Meyer, J. E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tatsumi, M., 1977, The Protein Data Bank: A computer-based archival file for macromolecular structures, *J. Mol. Biol.* **112:**525–542.

Available through: Protein Data Bank, Chemistry Department, Brookhaven National Laboratory, Upton, NY 11973; University Chemical Laboratory, Lensfield Road, Cambridge CB2 1EW, United Kingdom; and University of Tokyo, Hongo, Tokyo, Japan.