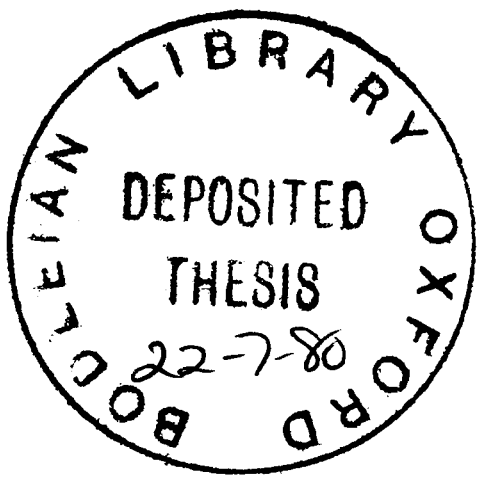
PROTEIN FOLDING

by

FRED E. COHEN



Molecular Biophysics
Department of Zoology

Wolfson College
Trinity Term 1980

ABSTRACT

PROTEIN FOLDING

Fred. E. Cohen, Wolfson College, D.Phil. Thesis, Trinity Term, 1980.

Recent studies of the relationship between protein sequence and protein structure are reviewed. A detailed discussion of past attempts to predict the structure of a protein from its amino acid sequence, the protein folding problem, is presented and the strengths and weaknesses of these methods are examined. The root-mean-square deviation is studied and a benchmark for structural comparisons is established. A combinatorial approach to the protein folding problem is outlined and its advantages over existing methods is discussed. Specific algorithms based on the combinatorial approach are developed and applied to a variety of proteins. The success of this approach in terms of the root-mean-square deviation benchmark as well as the drawbacks of this method are presented.

ACKNOWLEDGEMENTS

There are several colleagues and friends whom I wish to thank for their contributions to the work this thesis describes. I am very grateful to Professor Sir David Phillips and Professor F.M. Richards for teaching me how to think and do science. They are models to be followed. I thank Professor I.D. Kuntz, Professor P.A. Kollman and Dr. T.J. Richmond for many long and stimulating discussions and fruitful collaborations. I am especially grateful to Dr. L.N. Johnson for her comments on the text. W. Taylor, M. Lewis and the rest of F12 helped me keep my wits. Dr. M.J.E. Sternberg deserves my most sincere thanks. I could not count the number of times we discussed protein structure or the number of arguments we have had. I believe that we now choose opposite sides from force of habit. I could have had no better colleague for the past two years, but I will be sorry to leave because I could have had no better friend. None of this would have been possible without the generous financial support of the Rhodes Trust.

To Carolyn

TABLE OF CONTENTS

List of Tables

List of Figures

LIST OF PUBLICATIONS

COHEN, F.E., RICHMOND, T.J. and RICHARDS, F.M. (1979). J. Mol. Biol. 132, 275-288.

RICHARDS, F.M., RICHMOND, T.J., STERNBERG, M.J.E. and COHEN, F.E. (1980). In Protein Folding (ed. R. Jaenicke), Elsevier, Amsterdam, 117-130.

COHEN, F.E., STERNBERG, M.J.E., and TAYLOR, W. (1980). In Protein Folding (ed. R. Jaenicke), Elsevier, Amsterdam, 131-148.

COHEN, F.E. and STERNBERG, M.J.E. (1980). J. Mol. Biol. 137, 9-22.

COHEN, F.E. and STERNBERG, M.J.E. (1980). J. Mol. Biol. 138, 321-333.

COHEN, F.E. and STERNBERG, M.J.E. (1980). In Amino Acids, Peptides and Proteins, Vol. 11, The Biochemical Society, London.

COHEN, F.E., STERNBERG, M.J.E. and TAYLOR, W.R. (1980). Nature, in press.

COHEN, F.E. and STERNBERG, M.J.E. (1981). In Amino Acids, Peptides and Proteins, Vol. 12, The Biochemical Society, London, in press.

CHAPTER I

INTRODUCTION

1. Prelude

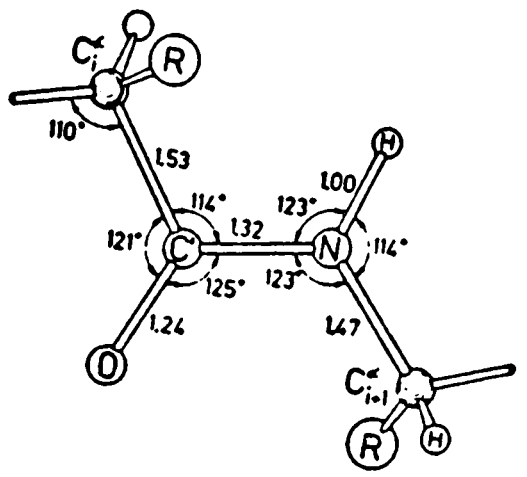
A folded globular protein elegantly satisfies the hydrogen bonding preference of buried nitrogen and oxygen atoms while simultaneously placing the remaining polar groups near the protein-solvent interface and internalising most hydrophobic residues. Since α-helices and β-sheets form networks of main chain hydrogen bonds between amide nitrogens and carbonyl oxygens (Pauling et al., 1951), they are logical candidates for building blocks of protein structure. This thesis investigates the packing of α-helices and β-sheets in an effort to produce rules governing such assemblies. These rules form the basis of a series of algorithms designed ultimately to predict the three dimensional structure of a protein from the amino acid sequence.

2. Amino Acids & Proteins: Some Definitions and Chemistry

Proteins are linear heteropolymers constructed from the head-to-tail condensation of L-α-amino acids and the product of several linkages is a polypeptide chain. The molecular dimensions of the repeating unit are shown in Figure 1.1. Twenty amino acids (see Figure 1.2) are commonly seen in protein structures although some post-translational modifications of specific amino acid residues along the polypeptide chain increase this figure. A wide range of chemical and physical properties are spanned by the twenty essential amino acids. They vary in hydrophobicity, charge, volume, surface area, side chain flexibility and aromaticity (see Table 1.1). Variations with respect to these properties are often invoked to

FIGURE 1.1

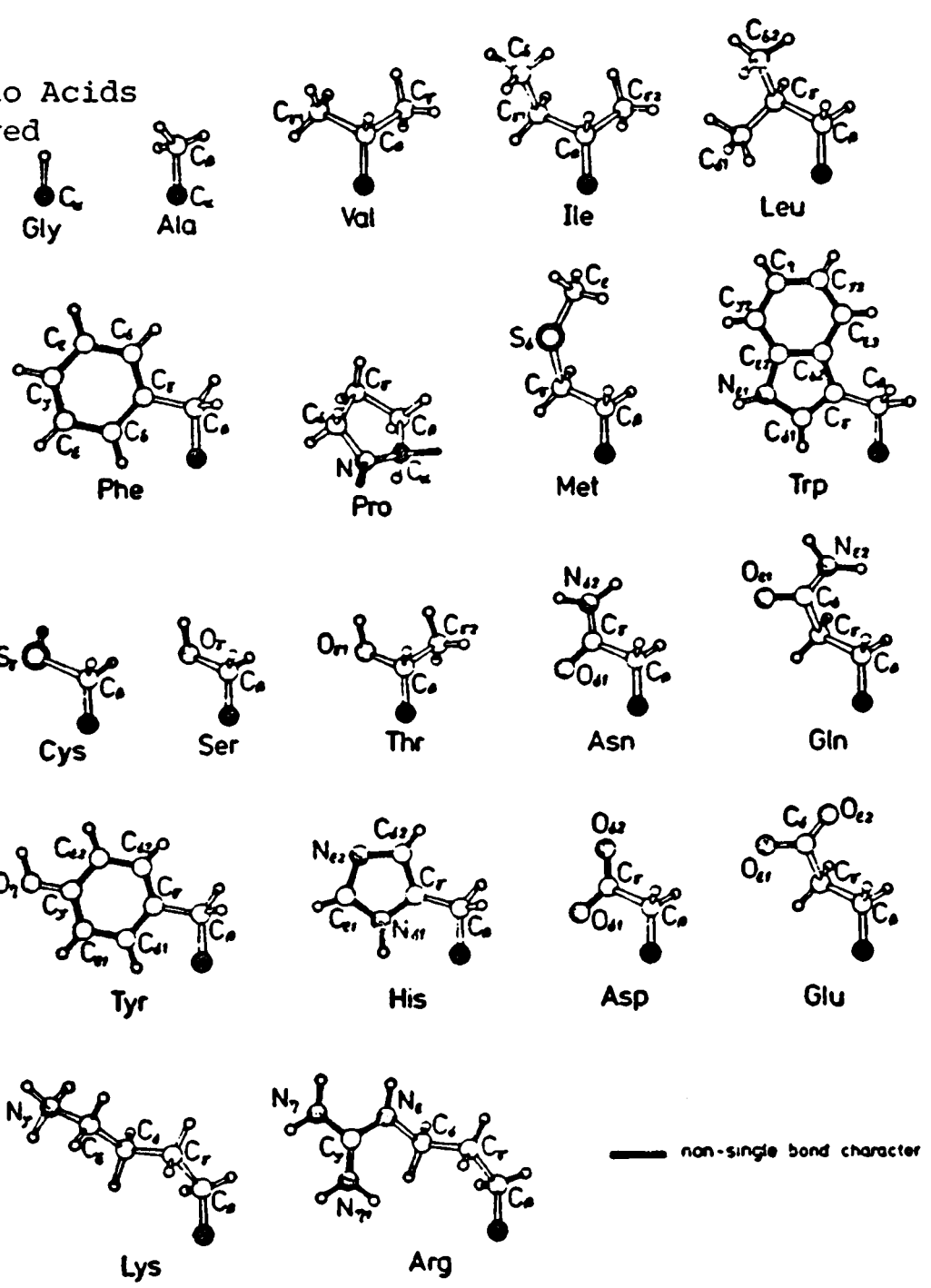
Atomic dimensions of the Peptide Linkage



Bond angles in degrees and bond lengths in Ångstroms are depicted for the atoms of the polypeptide backbone, as described by Pauling et al. (1951). This figure is taken from Schulz and Schirmer (1979).

FIGURE 1.2.

The Twenty Amino Acids Commonly observed in Proteins



The R-group of the twenty amino acids commonly observed in proteins and the entire structure of Proline are shown. The three letter code to identify each residue together with the individual atom labels are included. Bonds with partial double bond character are shaded. This Figure is taken from Schulz and Schirmer (1979).

TABLE 1.1

Properties of the Twenty Essential Amino Acids

	Contact area[a] in Å²		Transfer free energies[b]	Molecular weight	Number of free rotating bonds[c]	$\frac{\Delta G}{\chi_i+\phi+\psi}$
	β	α				
Glycine	24	13	0.0	75	2	0.
Alanine	34	20	0.5	89	3	0.17
Leucine	51	35	1.8	131	6	0.3
Valine	47	33	1.5	117	5	0.3
Isoleucine	54	39	2.97	131	6	0.5
Cysteine	41	25	1.4	121	4	0.35
Methionine	61	43	1.3	149	5	0.26
Phenylalanine	64	46	2.5	165	4	0.62
Tyrosine	64	46	2.3	181	4	0.58
Tryptophan	80	61	3.4	204	4	0.85
Serine	34	20	-0.3	105	3	-0.1
Threonine	41	28	0.4	119	4	0.1
Glutamic acid	48	33	0.55	147	5	0.11
Aspartine acid	40	26	0.54	133	4	0.13
Asparagine	42	28	-0.01	132	4	-0.0003
Glutamine	52	36	-0.1	146	5	0.02
Lysine	63	46	1.5	146	7	0.22
Histidine	53	37	0.5	115	4	0.12
Arginine	72	55	0.73	174	6	0.12
Proline	41	22	2.6	115	1	2.6

[a] From Richards & Richmond (1977)

[b] From Nozaki & Tanford (1971)

[c] $\chi + \phi + \psi$

explain the tendencies of certain residues to favour α-helical, β-strand or β-turn conformations (e.g. Richards & Richmond, 1978) which are observed empirically from known protein structures (Chou & Fasman, 1974). These variations are also responsible for the myriad of specific structural and functional roles of proteins.

Proteins are conveniently divided into two classes: globular and fibrous, on the basis of shape. Fibrous proteins normally play structural roles in cellular organisation. They are frequently found to have a regular repeating unit like the three fold repeat in collagen $(Gly\text{-}X\text{-}Y)_n$ where X is often proline, Y is often hydroxyproline and n can be greater than 300. For collagen, this three fold repeat produces a triple helix where three collagen chains wind on each other. Other fibrous proteins include hair (α-keratin) and silk (β-fibroin).

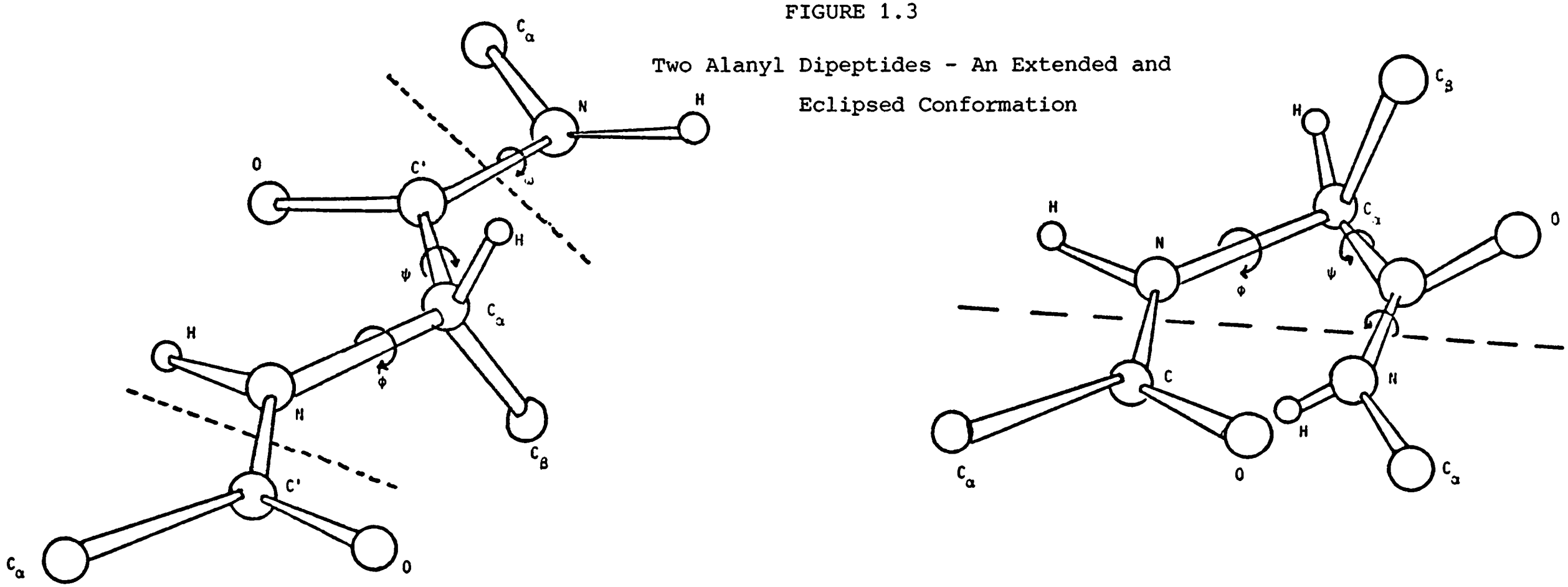
This thesis will focus on the structure of globular proteins. These proteins are central to metabolism and are typically involved in catalysing reactions (enzymes), in the transport of metabolites, or in the control of reactions. The structure of globular proteins is normally described on four levels:

2.1 Primary Structure is the amino acid sequence of the polypeptide chain and disulphide bridges between cystine residues.

2.2 Secondary structure is the local three dimensional arrangement of the polypeptide chain. These structures are often recognised by the regular repeat of specific backbone dihedral angles (ϕ,ψ) and characteristic hydrogen bonds. ϕ and ψ specify rotations about the $N\text{-}C_\alpha$ and $C_\alpha\text{-}C_{carbonyl}$ bonds respectively (see Figure 1.3). A Ramachandran plot (Ramachandran & Sasisekharan, 1968) of the conformation space available to an alanyl

FIGURE 1.3

Two Alanyl Dipeptides - An Extended and Eclipsed Conformation



The dipeptide on the left is in an extended conformation with $\psi = \phi = 180°$. The dipeptide on the right is in an eclipsed conformation with $\psi = \phi = 0°$. The dotted lines define the limits of a single residue. The angle ω, which measures the deviation from planarity of the peptide linkage is zero in both cases.

dipeptide subject to a hard sphere potential suggests that secondary structure units are stable (see Figure 1.4). The fundamental structural units frequently observed in proteins include:

2.2.1 α-helices - a sequence of residues with backbone ϕ,ψ angles near (-57°, -47°) and regular hydrogen bonding between the nitrogen on the i^{th} residue and the carbonyl oxygen on the $i+4^{th}$ residue (see Figure 1.5). This produces a structure with 3.6 residues/turn and a pitch of 1.5Å/residue. Minor distortions from this idealised structure are frequently observed.

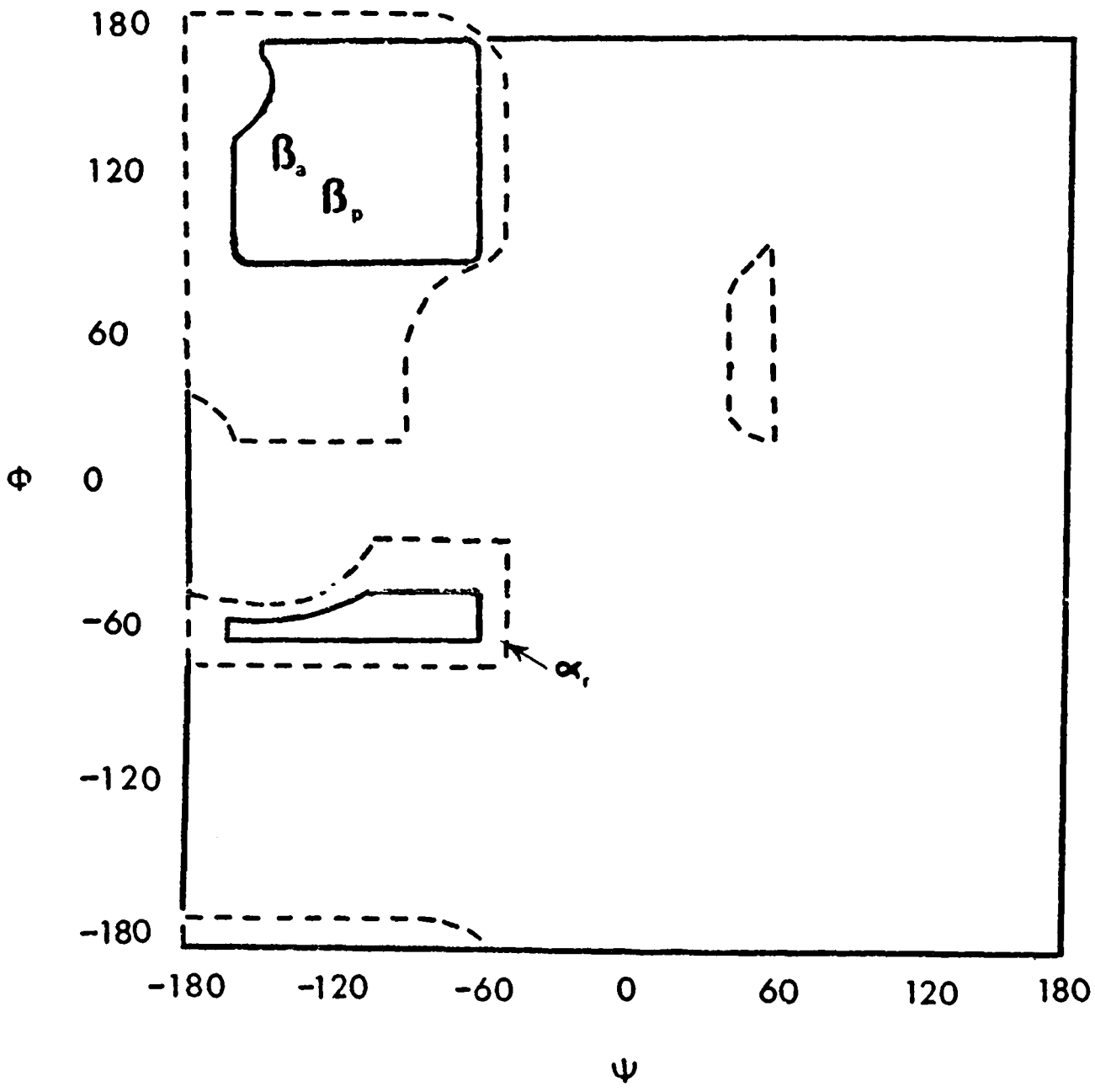
2.2.2 3_{10}-helix - a sequence of residues with backbone ϕ,ψ angles near (-30°, -30°) and regular hydrogen bonding between the nitrogen on the i^{th} residue and the carbonyl oxygen on the $i+3^{th}$ residue (see Figure 1.6). This produces a structure with 3.0 residues/turn and a pitch of 2.0Å/residue. The strain inherent in this structure limits its length to a few residues.

2.2.3 β-strand - a sequence of residues with backbone ϕ,ψ angles near (-120°, +120°). This produces a linear structure with axial length of 3.3Å/residue.

2.2.4 β-sheet - the parallel or antiparallel juxtaposition of two or more sequentially distinct β-strands so as to hydrogen bond the nitrogen and carbonyl oxygens on one strand to the carbonyl oxygens and nitrogens of another (see Figure 1.7). More formally, a β-sheet should be a unit of tertiary structure but for historical reasons, it is called secondary structure. The twist between adjacent strands is typically -20° (Chothia,

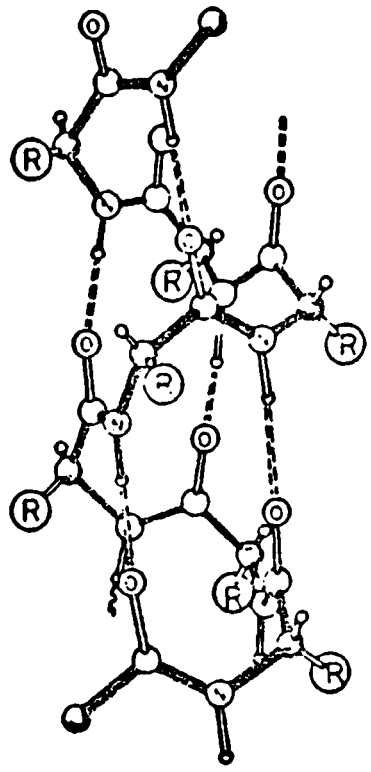
FIGURE 1.4

Ramachandran Plot of Allowed Conformations for an Alanyl Dipeptide.



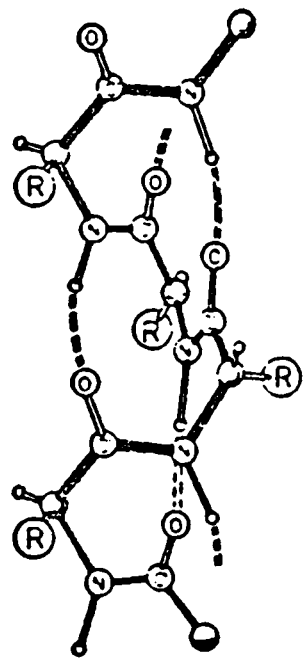
Regions of (ϕ,ψ) space available to a hard sphere model of an alanyl dipeptide. The regions which are completely allowed are enclosed by solid lines and partially allowed regions are enclosed by dotted lines. The location of right-handed α-helical (α_r) and parallel (β_p) and antiparallel (β_a) are marked.

FIGURE 1.5 α-helix



The regular hydrogen bonding pattern of the α-helix was first described by Pauling et al. (1951). It is commonly seen in both globular and fibrous proteins.

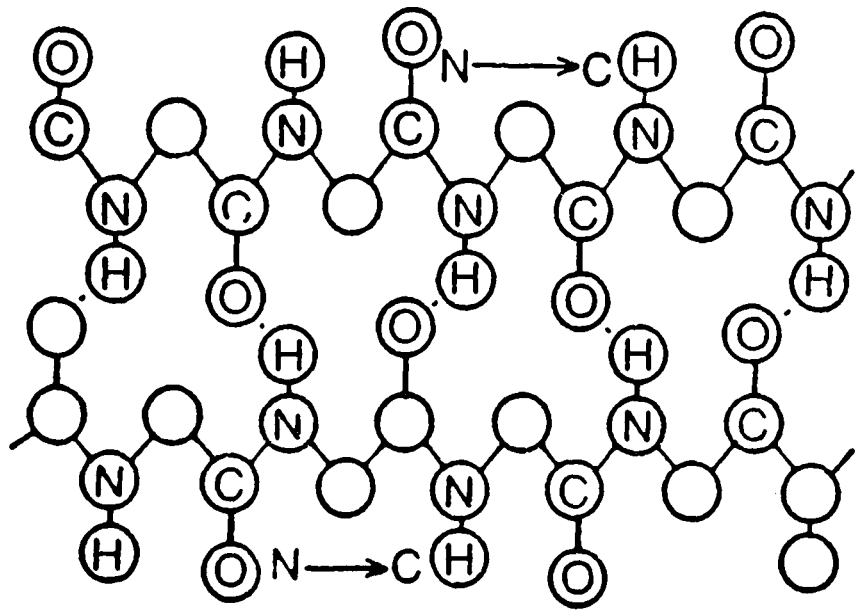
FIGURE 1.6 3_{10}-helix



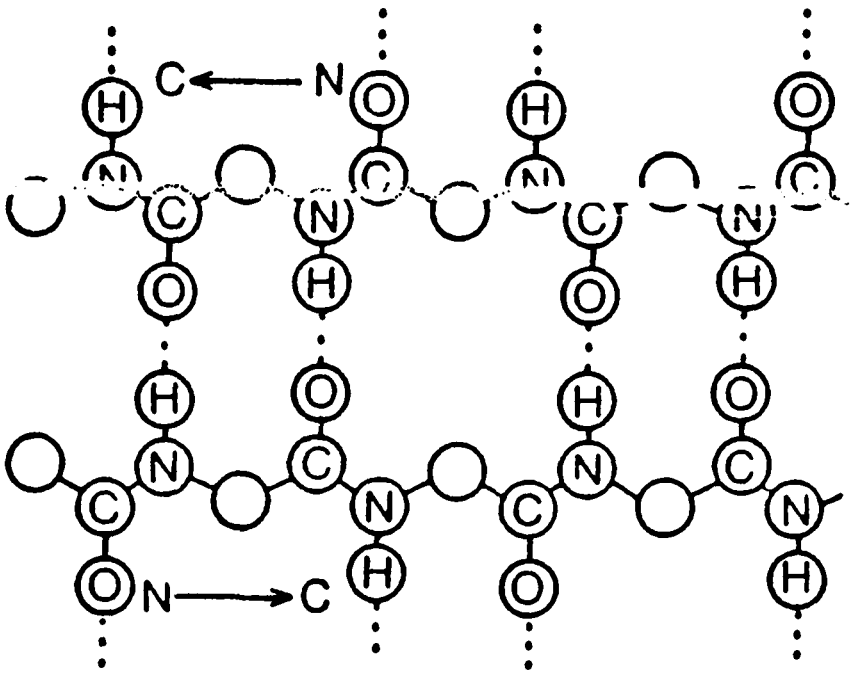
The 3_{10}-helix requires slight distortions in hydrogen bond geometry. Only small stretches of 3_{10}-helix are seen in proteins.

FIGURE 1.7 β-sheets

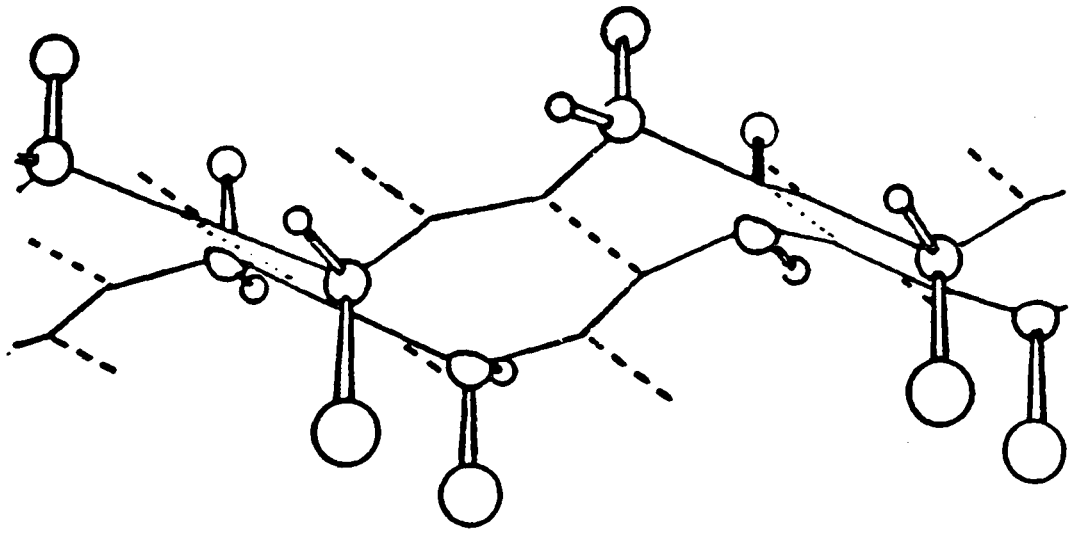
(a)



(b)



(c)

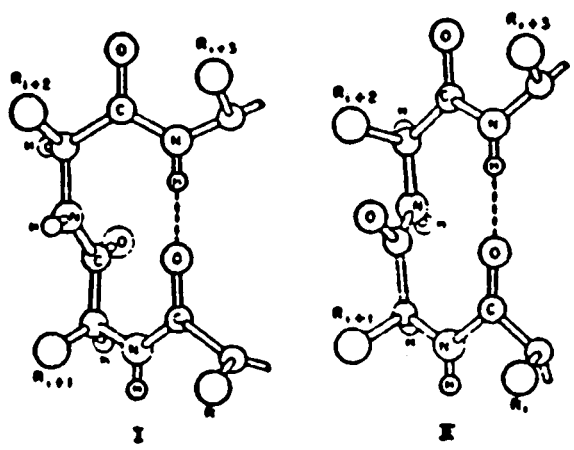


(a) A schematic representation of the backbone hydrogen bonding in a parallel β-sheet.

(b) A schematic representation of the backbone hydrogen bonding in an antiparallel β-sheet. Notice how parallel (antiparallel) hydrogen bonding on one side of a β-strand places the amide nitrogens and carbonyl oxygens on the other side of the strand in position for another strand to be aligned parallel (or antiparallel) to it.

(c) A three dimensional view of the lack of planarity in a β-sheet.

FIGURE 1.8 Two Types of β-bends



Venkatachalam (1968) has described two types of β-bends. They differ in the relative position of the carbonyl oxygen of residue i+1 and the amide nitrogen of residue i+2.

1973). Thus a 5-strand sheet is a noticeably distorted from the planar model of a β-keratin-like silk. Weathersford & Salemme (1979) have attributed this right handed twist to tetrahedral distortion of the peptide nitrogen.

2.2.5 β-bulge - a non-repetitive unit of secondary structure where the register of a β-strand is disrupted. A bulge in the structure results when the i and i+2 residues on one strand are hydrogen bonded to the j and j±3 residues on an adjacent strand (Richardson et al., 1978).

2.2.6 β-turn - a sequence of four residues where a tight hairpin is formed when the i and i+3 residue hydrogen bond (see Figure 1.8). Venkatachalam (1968) has classified these bends into three classes with characteristic composition and ϕ,ψ angles.

2.3 Tertiary Structure is the three dimensional arrangement of an entire connected polypeptide chain. In certain proteins (e.g. insulin or chymotrypsin), the original polypeptide chain has been cleaved and some residues may have been excised by proteolytic enzymes. Although no longer one polypeptide chain, this is still considered tertiary structure.

2.4 Quaternary Structure is the relative arrangement of protein monomers into polymers. These include the dimer of triose phosphate isomerase, the tetramer of haemoglobin, the hexamer of insulin, and the oligomers of tomato bushy stunt virus and tobacco mosaic virus coat proteins.

2.5 Proteins without tertiary structure. Recently, the existence of proteins without well-defined tertiary structure has been suggested

by nuclear magnetic resonance (NMR) studies. These proteins, which include chromogranin A (Daniels, 1977), phospholipase A_2 and κ-casien, cannot be classified as either globular or fibrous and have been excluded from the studies which follow.

3. The Determination of Protein Structure

In the sixty years since Bragg first realised that X-rays could be used to probe the spatial arrangement of atoms within molecules (Bragg, 1921), the detailed structures of larger and larger molecules have been determined. Twenty years ago, sperm whale myoglobin became the first protein structure solved to atomic resolution (Kendrew et al., 1960). Currently, the structures of 161 proteins are known (Dickerson, 1979). For a review of protein crystallography, see Phillips (1967).

In spite of the large number of structures available, two major problems complicate crystallographic analysis:

(1) The growth of crystals suitable for analysis

(2) Isolating isomorphous heavy metal derivatives of the native protein to solve the phase problem.

Although many protein structures have been solved to a resolution which permits the path of the polypeptide chain and the orientation of most side chains to be described (∿3Å), a small subset of these are studied at atomic resolution (1.5Å). Moreover, atomic positions (e.g. human lysozyme, Artymiuk, 1979) and occasionally topologies (e.g. hexokinase, Steitz, 1976) are revised as the resolution of the data increase.

Given these difficulties, a theoretical procedure for predicting the three dimensional arrangement of atoms in globular proteins would be desirable. The fact that many biologically interesting systems will never be suitable for crystallisation increases the value of a solution to the protein folding problem.

4. Principles of Organisation of Globular Proteins

4.1 Globular Proteins are compact with only small packing defects.

Although the polypeptide chain of myoglobin with 153 residues has a length of 550 Å (3.6Å/residue), the diameter of the native structure is only 36Å. Richards (1974) showed that the packing density, the ratio between the minimum volume of an object and the actual volume occupied, was 0.75 for proteins. This ratio is 0.74 for close packed spheres. However, for lysozyme and ribonuclease S, this ratio varies between 0.6 and 0.85 in various parts of the structure. It has been suggested that packing defects could be associated with dynamic fluctuations (Richards, 1979)

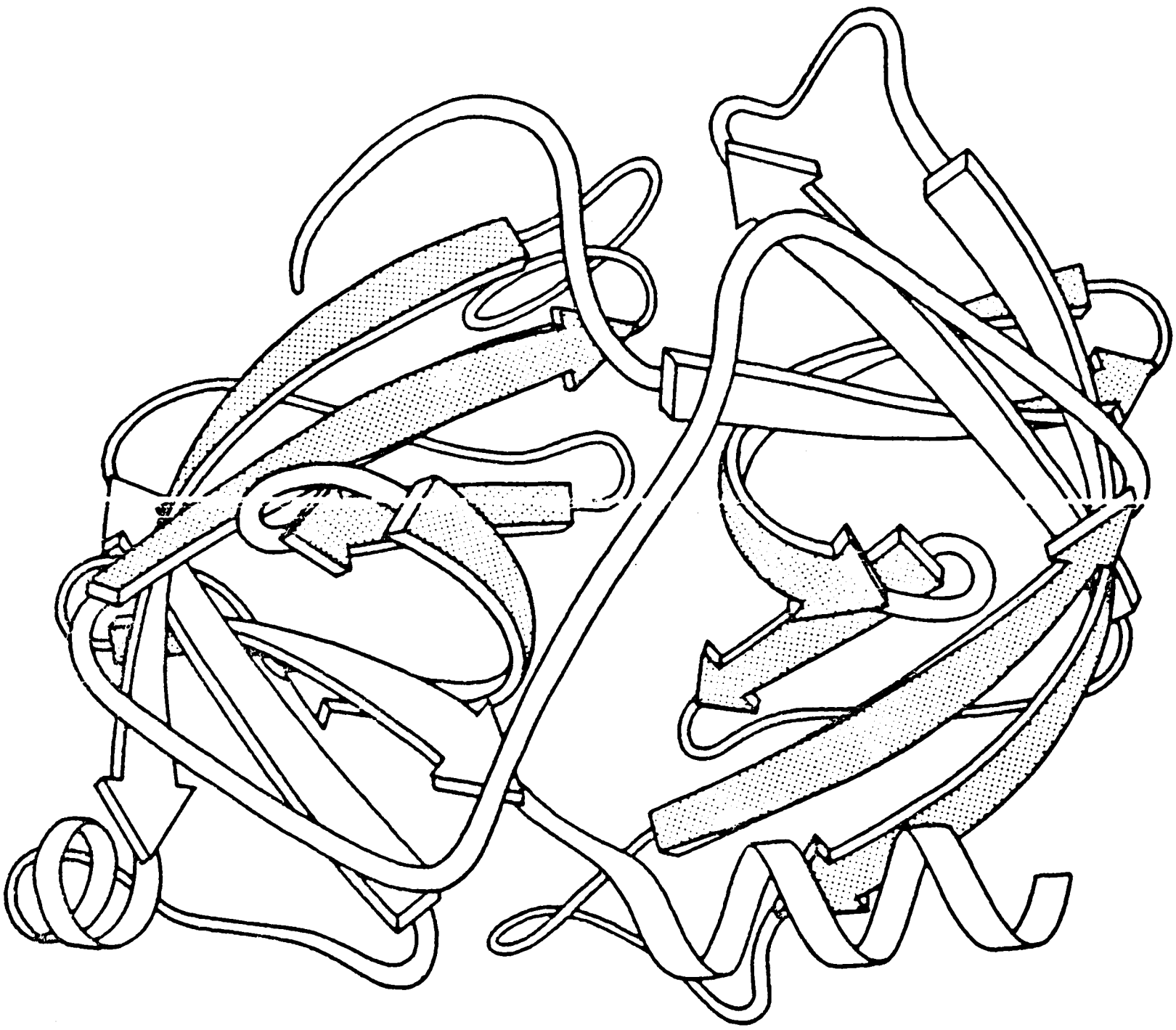
4.2 Large Proteins often have folding domains

Wetlaufer (1973) observed that a cleavage plane to segment a protein into two sequentially distinct structural units can often be found for proteins with molecular weights greater than 20000 daltons. Crippen (1978) has created a tree search algorithm for objectively locating folding domains by examining the properties of the matrix of interatomic distances. Rose (1979) has developed a similar procedure to find domains but relies on constructing the optimal cutting plane to segment the protein. The existence of semi-independent structural units within a single peptide has supported a hierarchic condensation model for protein folding: a protein folds through the sequential coalescing of increasingly complex polypeptidc segments.

Examples of domains in proteins include the nucleotide binding domain in the dehydrogenases (e.g. Rossmann et al., 1975), the helical and β-sheet domains in thermolysin (Colman et al., 1972), the two antiparallel 6-stranded β-barrels in elastase (Sawyer et al., 1973) (see Figure 1.9) and the bilobal structure of phosphoglycerate kinase (Banks et al., 1979).

FIGURE 1.9

The two Domains of Elastase



Elastase domains 1 & 2, β barrel exteriors stippled

This Figure is taken from Richardson (1979).

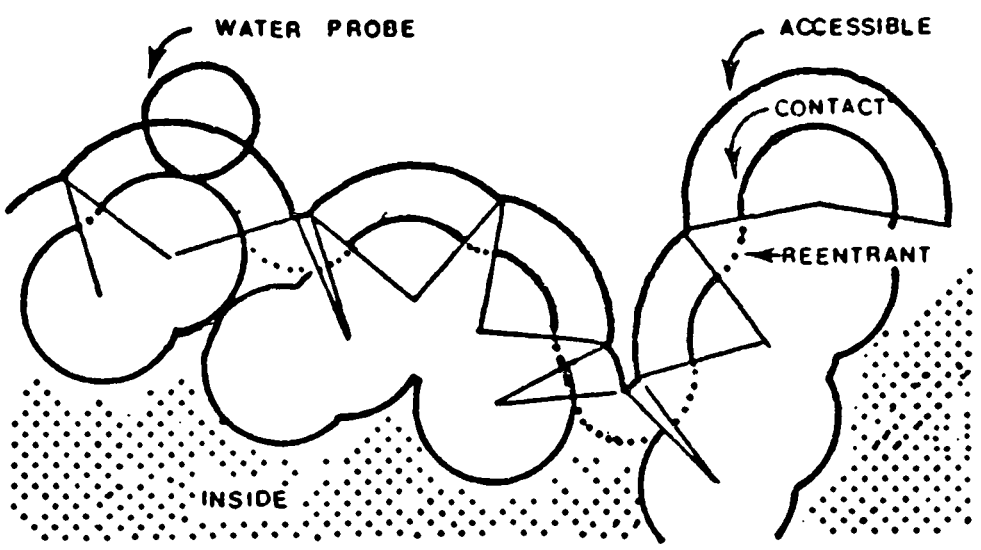
4.3 Non-polar side-chains prefer to be in the interior of the protein, while charged and polar residues tend to lie near the surface.

Kauzmann (1959) hypothesised that the driving force stabilising the folded state must be the reduction of interactions between non-polar side chains and water. The presence of non-polar molecules in water lowers the number of degrees of freedom of the water molecules. They can only interact with non-polar molecules through dipole-induced dipole interactions which are weak. Moreover, the presence of non-polar molecules limits the way in which water molecules might interact with each other. This is the basis of the hydrophobic effect. Lee & Richards (1971) introduced the concept of accessible surface area to quantify the tendencies of residues to be buried or exposed (see Figure 1.10). A general conclusion of this work and subsequent studies (Shrake & Rupley, 1973; Richards, 1974, 1977; Chothia, 1976) demonstrated that the average polar side chain is nearly 3.5 times as accessible as the average non-polar side chain. Of course, polar residues have some hydrophobic character and salt bridges or hydrogen bonds between polar side chains decrease the hydrophilicity of the polar pair, but the hydrophobic effect remains a powerful component in the energetics of protein folding (see section 6.1.6).

Chothia (1974) was able to derive an approximate relationship between accessible surface area and free energy. The correlation between data on the transfer free energy of amino acids from an organic solvent to water (Nozaki & Tanford, 1971) and the accessible surface area of the residue in an extended conformation of Gly-X-Gly lead to the equation $1\text{Å}^2 \equiv 23\text{cal}$ (see Figure 1.11). Strictly speaking, this relationship is reasonable only for the burial of non-polar atoms. Richmond & Richards (1978) have done detailed accessibility calculations on myoglobin and shown that major changes in surface area on folding an extended chain

FIGURE 1.10

Distinction between Accessible, Contact and Reentrant Surface Area



The concept of accessible surface area was introduced by Lee & Richards (1971). In the planar section of the van der Waals surface of a hypothetical molecule, accessible surface area is calculated by computing the length of the arc traced by the centre of a water molecule probe with radius 1.4Å as it rolls along the molecular surface. These arc lengths are numerically integrated over equi-spaced set of parallel sections. Contact area is that part of the molecular surface which can be touched by the water probe as it rolls along the surface. Contact area is roughly proportional to, but always smaller than, accessible surface area. Reentrant area is the difference between the molecular surface as viewed by a water probe and the contact surface area. This Figure is re-drawn from Richards (1977).

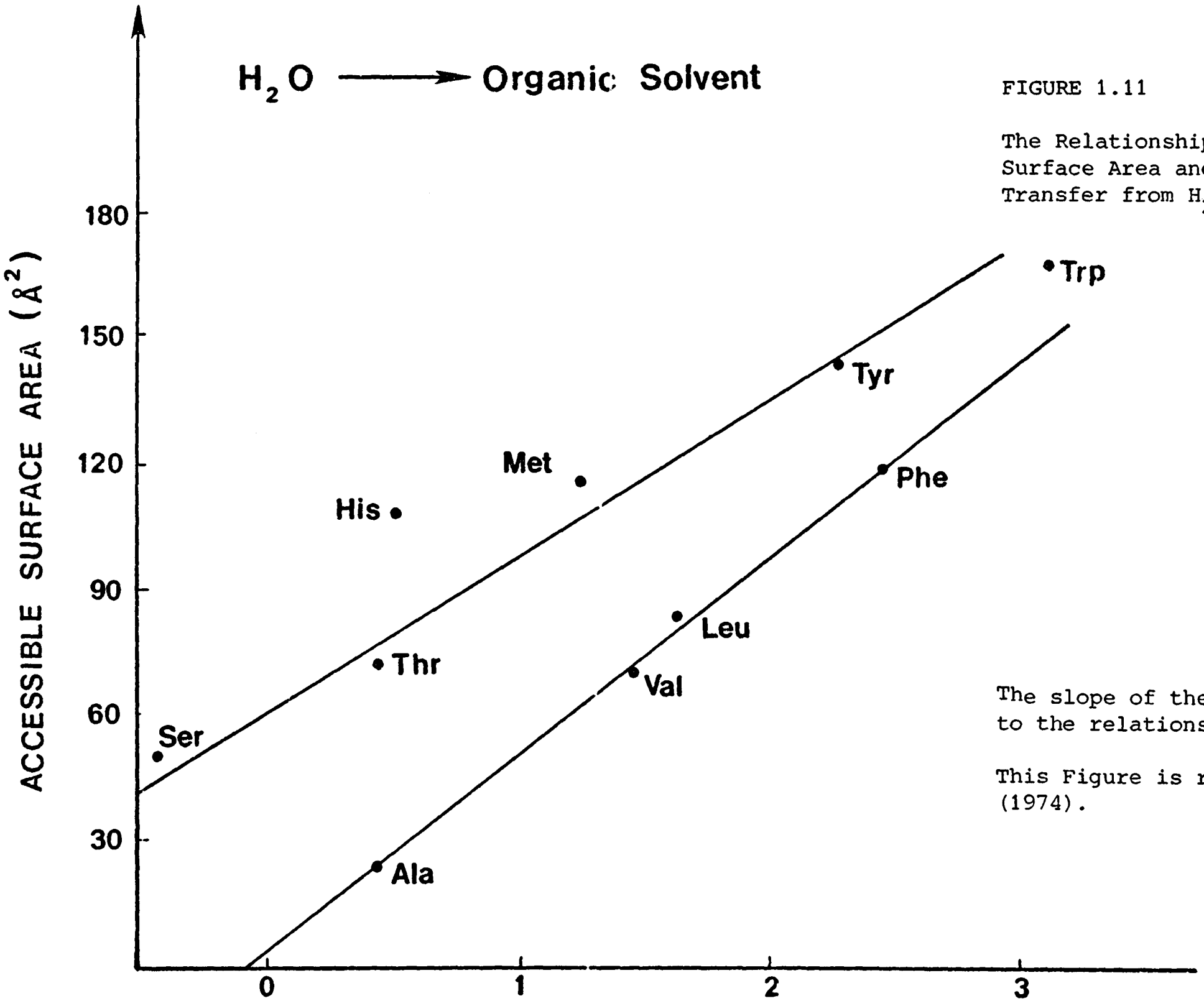


FIGURE 1.11

The Relationship between Accessible Surface Area and Free Energy of Transfer from H_2O to Organic Solvent.

The slope of the lower line gave rise to the relationship 1Å ≡ 23 cal.

This Figure is re-drawn from Chothia (1974).

occur in the formation of secondary structure and the interaction of pairs of secondary structure.

4.4 Internal bends are exceedingly rare

While the polypeptide chain traverses the diameter of a protein several times, chain reversals or "bends" are rarely found in the interior of proteins (Kuntz, 1972). Thus one finds that polar residues tend to constitute bends. Exceptions to this include the chain reversal at Ile 55 in lysozyme and a few examples in phosphorylase b (L. Johnson, personal communication).

4.5 Secondary Structures pack in defined classes

Levitt & Chothia (1976) categorised the packing of secondary structure in proteins into four classes:

αα - composed entirely of α-helices, e.g. myoglobin, cytochrome b_{562} (see Figure 1.12)

ββ - composed entirely of β-sheets, usually a packed pair of β-sheets, e.g. superoxide dismutase, the immunoglobulin domains (see Figure 1.13)

βαβ- α-helices packed against a predominantly parallel β-sheet e.g. flavodoxin, adenyl kinase (see Figure 1.14)

β+α - an αα domain and a ββ domain e.g. thermolyn.

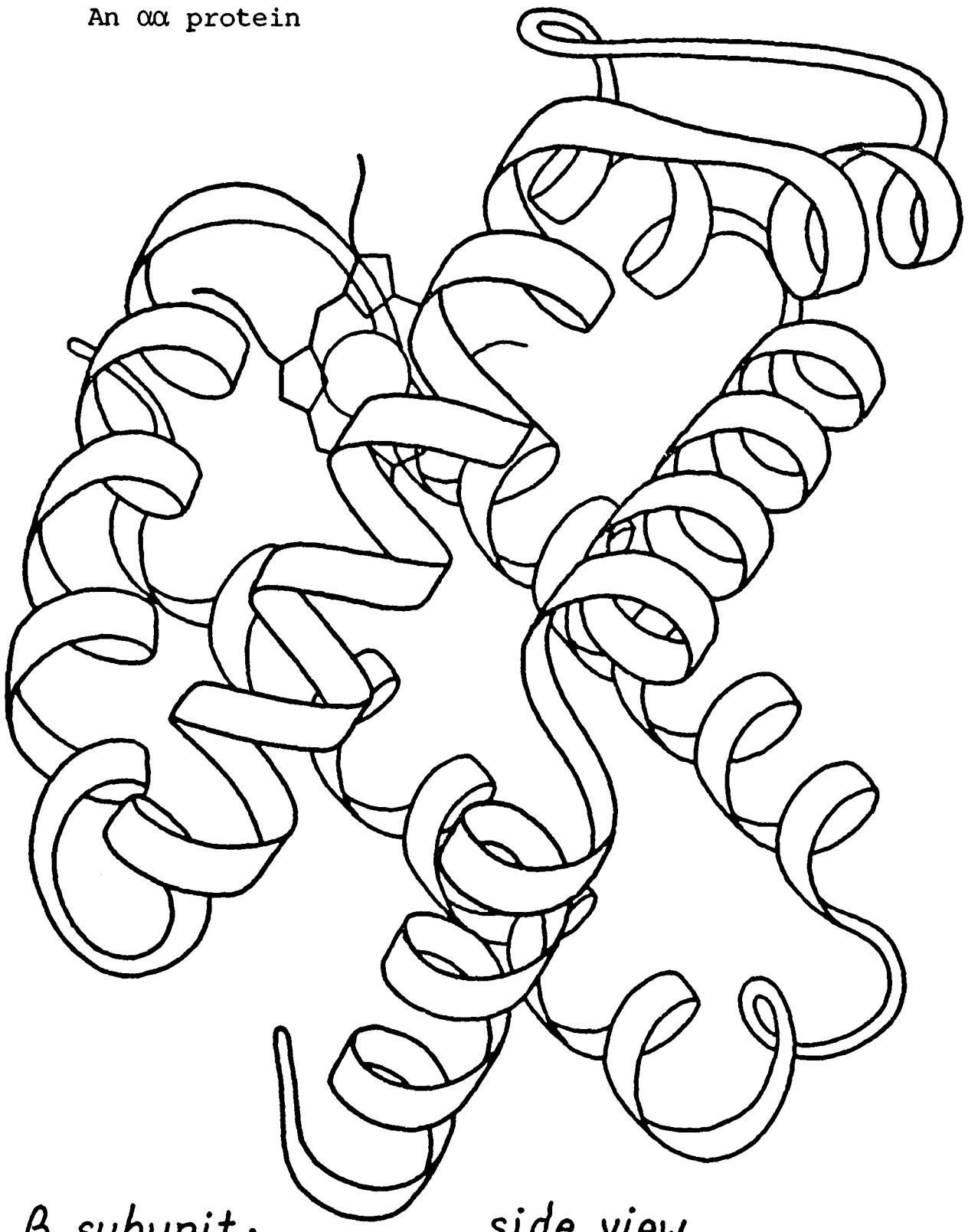
Chothia et al. (1977), Richmond & Richards (1978), Cohen et al. (1979, 1980a) and Sternberg et al. (unpublished data) then went on to describe the specific nature of helix-helix and sheet-sheet interactions in terms of: dihedral interaxial angles, interaxial separation, and the residues central to the interactions.

4.5.1 Right handed connections between parallel β-strands predominate

Sternberg & Thorton (1976) and Richardson (1977) observed that

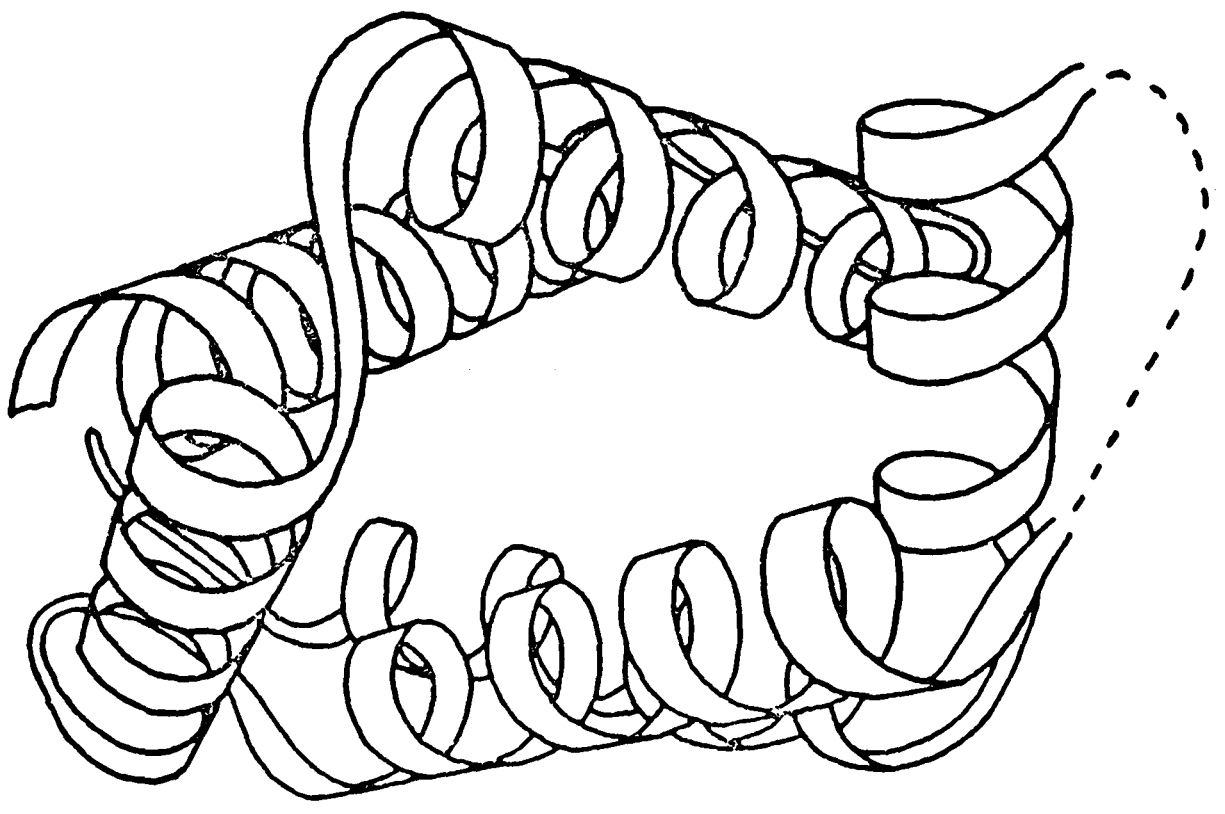
FIGURE 1.12

An αα protein



Hemoglobin β subunit: *side view*

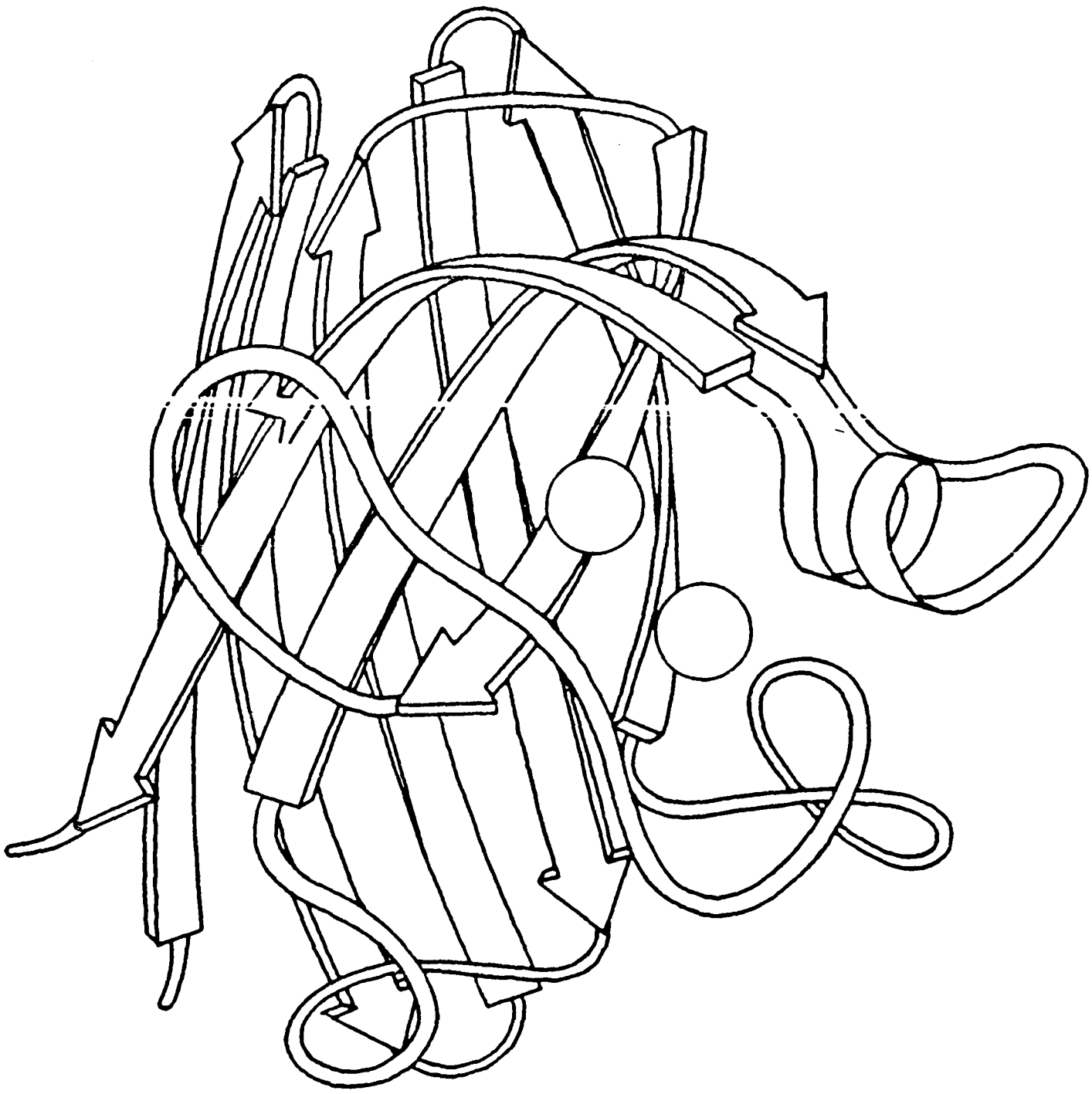
end view



This Figure is taken from Richardson (1979).

FIGURE 1.13

A ββ protein

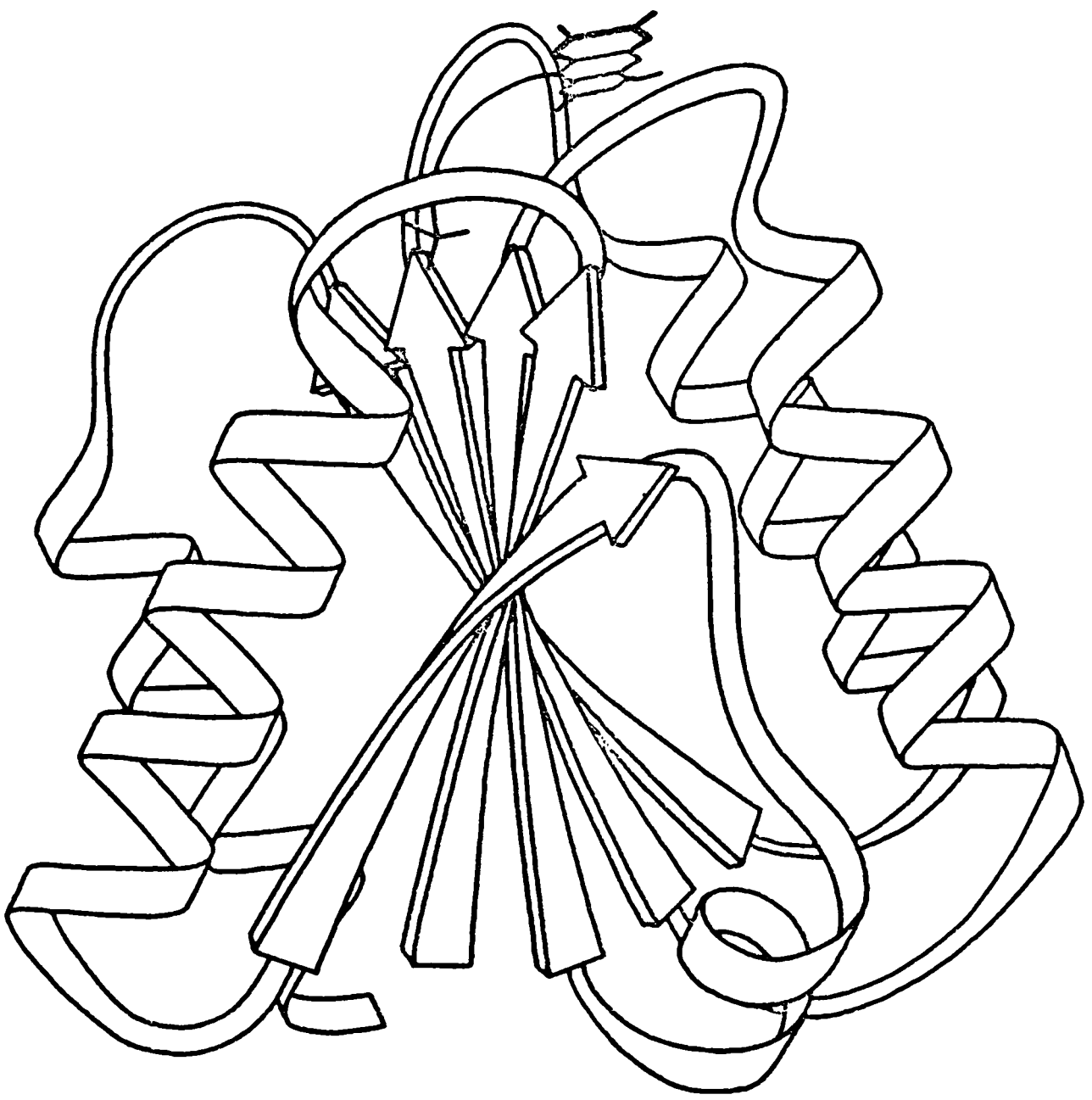


Cu,Zn Superoxide Dismutase

This Figure is taken from Richardson (1979).

FIGURE 1.14

A βαβ protein



Flavodoxin

This Figure is taken from Richardson (1979).

a helical connection between parallel β-strands had a right-handed sense in 57 of 58 known connections (see Figure 1.15). They later showed that this tendency was maintained whether the connecting loop was α-helical, a β-strand, or a coiled segment. This preference is attributed to the fact that for two strands with a -20° twist between them, the connecting loop must traverse a longer path for the connection to be left handed.

4.5.2 In pure parallel β-sheets, at most one chain reversal is observed

Richardson (1977), in an extensive survey of β-sheets, noticed that the pure parallel sheets never changed direction more than once. Thus, the topology F E D A B C would be acceptable but A C E F D B would not (see Figure 1.16).

4.5.3 A Greek key pattern is frequently observed in all-β proteins

Richardson (1977) has noticed that β-sheets frequently have the strand order i, i+3, i+2, i+1, i+4 which produces a pattern seen in Greek pottery (see Figure 1.17). Ptitsyn et al. (1979) suggest that all-β proteins must have a Greek key topology.

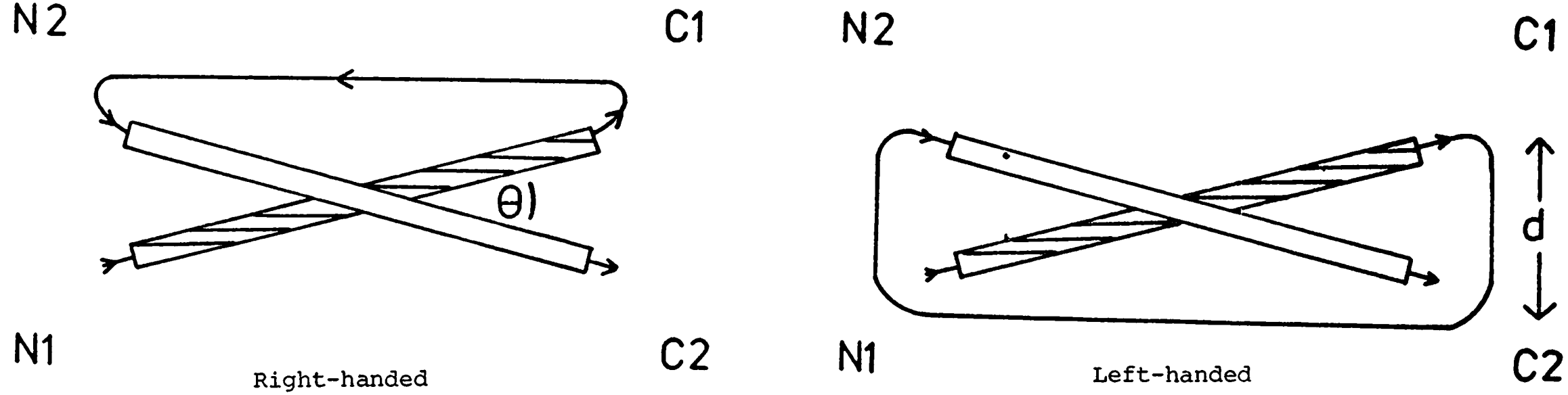
• • •

4.5.4 β-sheets exhibit a high proportion of adjacent connections

Richardson et al. (1976), Levitt & Chothia (1976) and Sternberg & Thorton (1977b) have all noticed that most neighbouring strands in a sheet are neighbours in sequence as well (see Figure 1.16). Thus, the nucleotide binding domain of lactate dehydrogenase with strand order F E D A B C has adjacency 4. The optimal adjacency for a 3, 4 or 5 stranded sheet is 2, for a 6-strand 4, and for a sheet with 7 - 10 strands 5 (Sternberg & Thorton (1977b)).

FIGURE 1.15

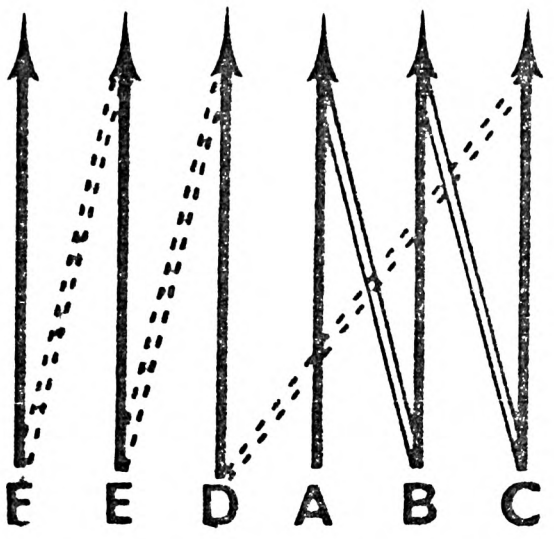
A Right and Left Handed βxβ Connection



The two possible topological connections for a βxβ unit where x is either an α-helix, a β-strand, or a coiled segment. The right-handed connection requires a shorter connection length than the left-handed connection.

FIGURE 1.16

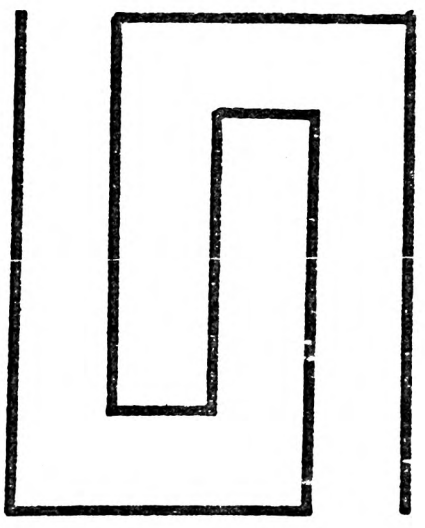
Chain Reversal and Adjacency - A Schematic β-sheet



β-strands are the bold lines with arrows. Connecting loops are double lines which are solid when in front of the sheet and dotted when behind. The chain proceeds from left to right from A to C and then reverses to right to left from C to F. The adjacency is 4 for this six-stranded parallel β-sheet since A is next to B, B to C, D to E, and E to F.

FIGURE 1.17

The Greek Key Topology



GREEK KEY PATTERN

This is Greek key pattern which was seen on ancient pottery and occurs as a common structural motif in proteins (Richardson, 1977).

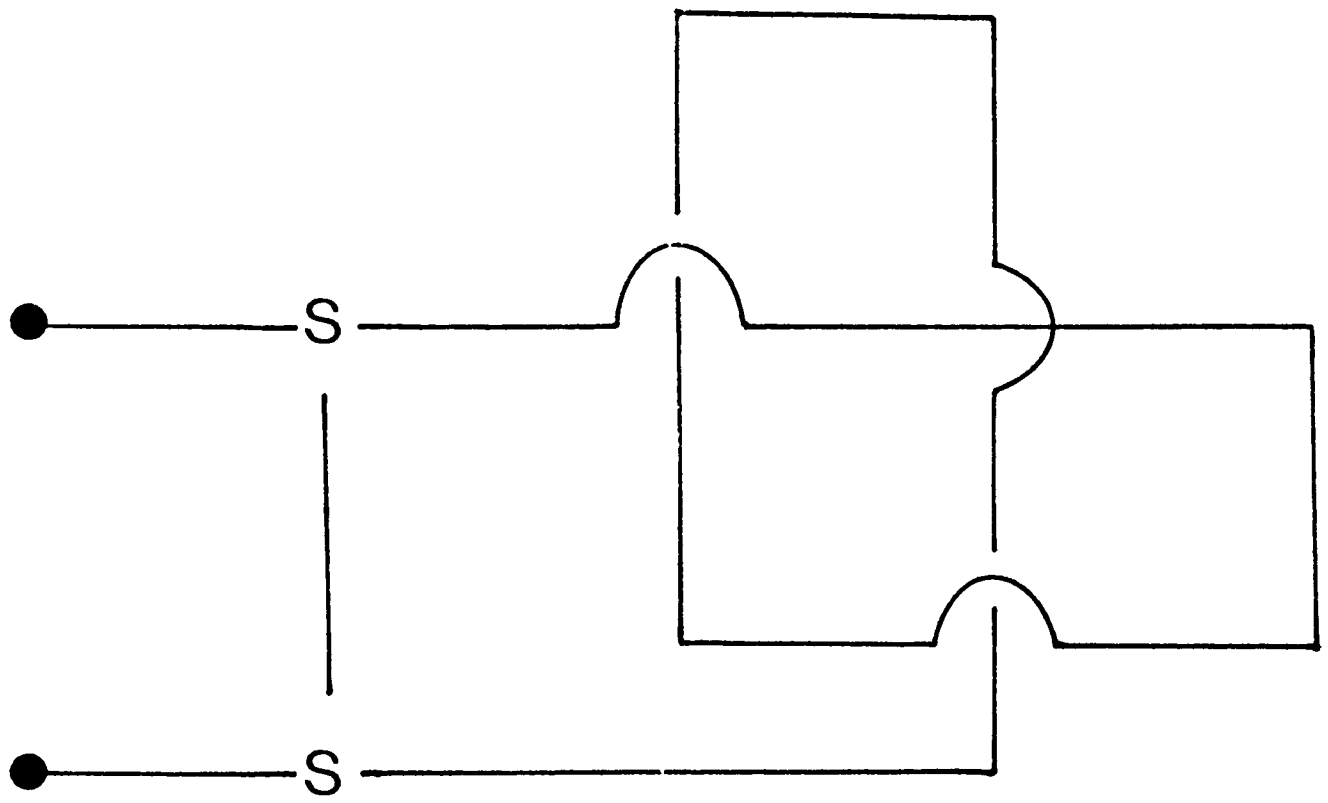
4.6 No knots are observed in globular proteins

Although knotted topologies could exist, Crippen (1974, 1975) has called attention to their notable exclusion. This is true of formal knots created by disulphide bridges which crosslink the chain (see Figure 1.18) as well as for pseudo-knots which could be formed when the N- and C-terminal ends of the protein are pulled apart (see Figure 1.19). One explanation commonly offered is that the kinetic barrier to forming knotted topologies is significantly larger than for unknotted chains. The existence of "slip-knots" has been seen in pancreatic trypsin inhibitor (Deisenhofer & Steigemann, 1974, 1975) and in carboxypeptidase (Quiocho & Lipscomb, 1971).

4.7 Similar tertiary folds result from distinct sequences

A comparison of the sequences of sperm whale myoglobin and horse haemoglobin α- and β-chains by Kendrew et al. (1965) showed a seemingly perplexing pattern of substitutions. When these substitutions were analysed in terms of the structure, the conservation of the non-polar core became apparent. Romero-Herrera et al. (1978) have shown that the sequences of over fifty globins show largely conservative substitutions and a few conserved residues. Lesk & Chothia (1980) have performed a detailed study of the structural changes induced by amino acid substitutions in nine globins. They explain how large changes in residue volume can be accommodated by the displacement of α-helical segments as long as the relative geometry of the haem pocket is maintained. Similarites between proteins with common functions, e.g. the nucleotide binding domain in the dehydrogenases (Rossmann et al., 1975) are also not uncommon. However, one domain of asparate transcarbamylase has the same topology as lactate dehydrogenase but does not bind a nucleotide in the same position (Honzatko et al., 1979). As the fold of a protein is guided by thermodynamic principles, it is quite reasonable to expect similar structural motifs and

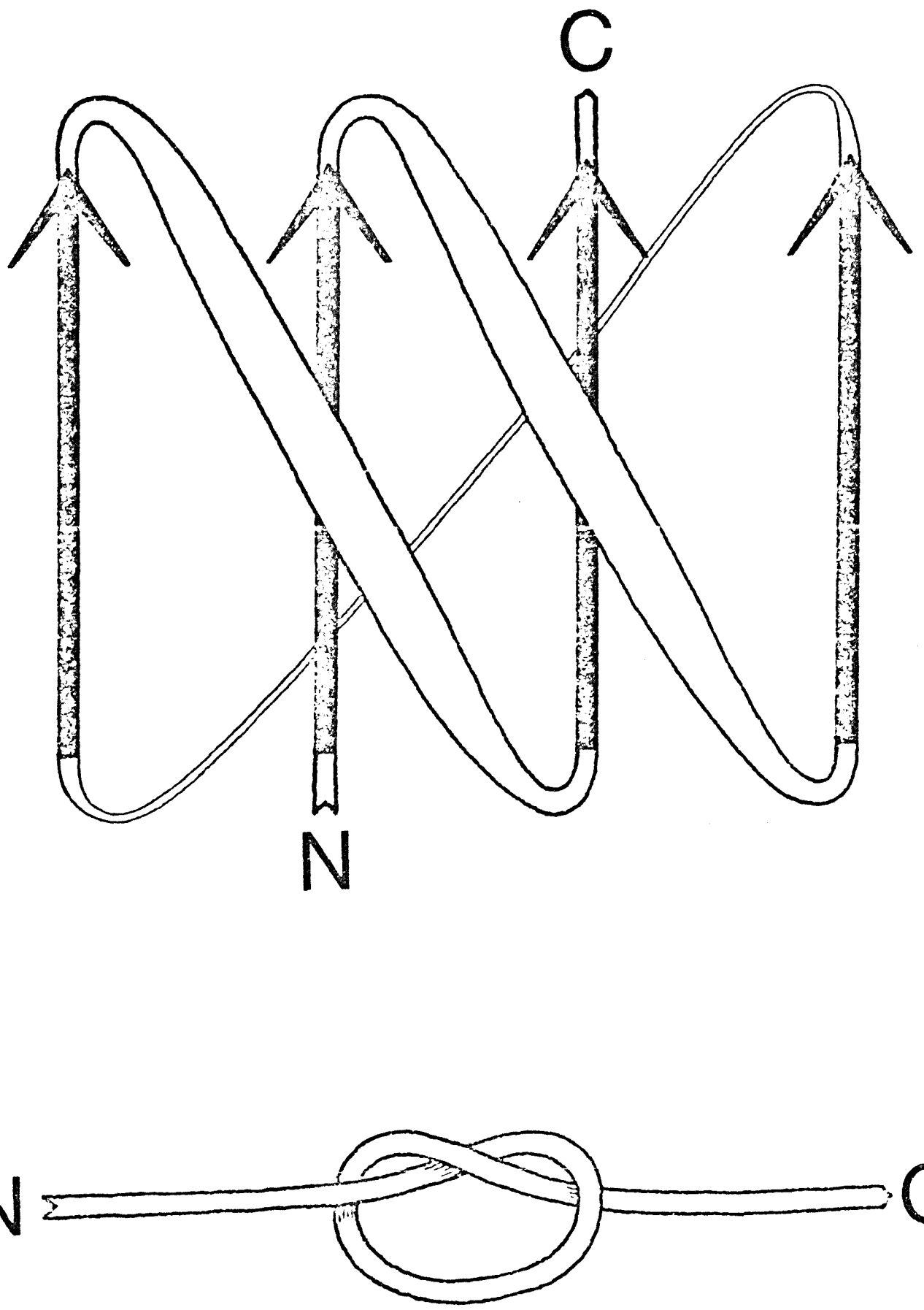
FIGURE 1.18 A Knotted Topology



In this schematic of a polypeptide backbone, an intrachain disulphide forms a knot in the structure. Such a structure has never been observed in proteins (Crippen, 1974, 1975).

FIGURE 1.19

A Knotted ß-Sheet



A pure parallel β-sheet topology which forms a knot when the N and C terminal ends are pulled. This type of knot is never seen in proteins (Richardson, 1977). Although such knotted topologies are rare for hypothetical β-sheets with less than 5 strands, they are a large percentage of the set of all possible topologies for sheets with 8 or more strands.

topologies to be shared by proteins of distinct sequence and function as well as homologous proteins. This is certainly true of superoxide dismutase and the immunoglobin domain (Richardson et al., 1976). These issues are discussed in greater detail in Chapters 4 and 5 where a structural basis is presented for sequence homology in certain key positions of proteins in similar folding classes. Questions of convergent and divergent evolution are raised by these enquiries. Objective methods for sorting between these two alternatives have been proposed (e.g. McLaughlin, 1971) but this issue remains only partially resolved.

5. Experimental Approaches to the Protein Folding Problem

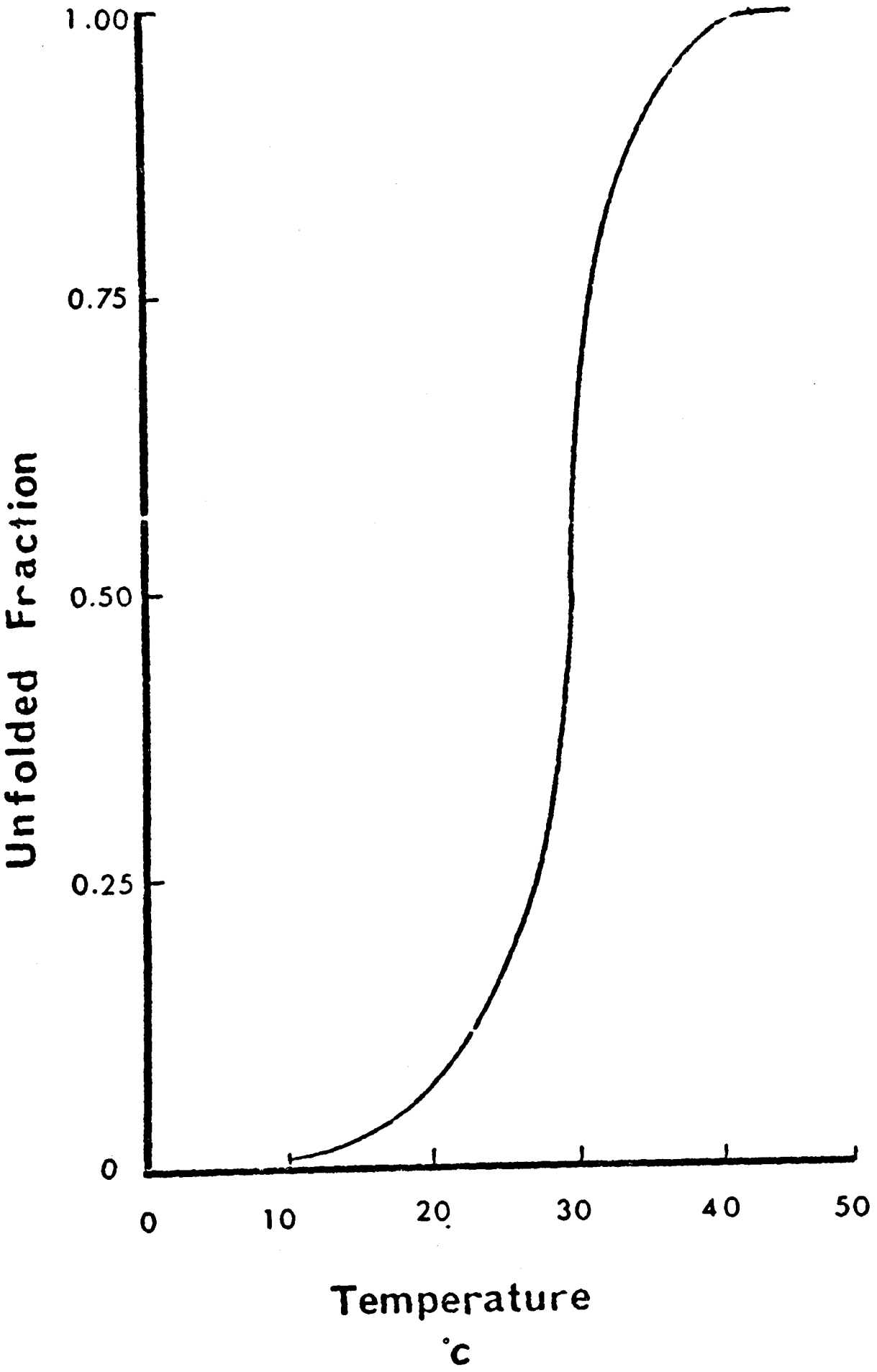
Though X-ray crystallographic methods can determine the three dimensional structure of a protein, they offer little insight into the folding path. A variety of physico-chemical measurements on the folding and unfolding of proteins have led to a preliminary picture of the kinetics of protein folding.

5.1 Sequence determines structure

The classic experiments of Anfinsen et al. (1961) on ribonuclease demonstrated that the protein could be reversibly denatured in vitro under the appropriate conditions. Fully reduced disulphide bridges would reform in the presence of a disulphide interchange enzyme. Thus it was concluded that the amino acid sequence contained all of the information necessary to define the unique tertiary structure. All doubt of this conclusion was removed when Hirschmann et al. (1969) succeeded in the total chemical synthesis of active ribonuclease. The renaturation of many other proteins has been demonstrated but proteins whose active form is the product of proteolytic cleavage of a precursor (e.g. insulin) will not renature

FIGURE 1.20.

Thermal Denaturation of Ribonuclease



Fraction of Unfolded Ribonuclease as a function of temperature at pH 2.19. The unfolded fraction was measured by intrinsic viscosity, optical rotation at 365nm and difference absorption at 289nm. This Figure is re-drawn from Pace (1975).

5.3 Nucleation vs. Intermediate Control of Protein Folding

5.3.1 Nucleation Control

In order to explain the rate of protein folding, a variety of generic initial events which nucleate the transition between the unfolded and native state have been proposed. These include a cluster of α-helices (Lim, 1978), a β-hairpin (Ptitsyn et al., 1979), or a cluster of hydrophobic residues (Matheson & Scheraga, 1978). These models all suggest that a certain obligatory intermediate (I*) must be formed for folding to proceed. However folding follows so quickly from the formation of I* that it never accumulates.

$$U \rightleftharpoons I^* \rightleftharpoons N \tag{1.3}$$

$$\frac{dI^*}{dT} = 0 \tag{1.4}$$

This equation follows from a steady state approximation where the concentration of I* is negligible.

Labhardt & Baldwin (1979) have examined the refolding of ribonuclease S (subtilisin cleaved ribonuclease with a nineteen residue fragment, S-peptide and non-covalently linked to the remaining S-protein) and found it to be inconsistent with nucleation control. Ribonuclease S folds more rapidly than S-protein and the rate of ribonuclease S folding is concentration dependent while the rate for S-protein is not. Thus S-peptide must combine with S-protein before S-protein completely folds. If a nucleation mechanism was operating, then S-peptide binding to the nucleated species would pull the equilibrium from U to I*. If the upper bound for the second-order binding rate constant is $10^9 M^{-1} s^{-1}$, then the species binding to S-peptide must accumulate (Baldwin, 1979). This is inconsistent with equation (1.4) and so nucleation control is not possible.

5.3.2 Intermediate Control

The conclusion of Labhardt & Baldwin (1979) is that folding must

be intermediate-controlled. The nature of these folding intermediates has been discussed by Ptitsyn & Rashin (1975) for myoglobin. Three phases of folding are hypothesized:

(1) The formation of separate helices;

(2) The diffusion of two helices together in a suitable orientation; and

(3) Docking of the helices.

5.3.2.1 The formation of α-helices

The formation of α-helices is conveniently described as an initiation-elongation process. The mechanism is clearly nucleation-controlled with an initiation rate of 10^{-7} - 10^{-5}sec for some synthetic polypeptides and elongation rates in a time range of 10^{-8} - 10^{-11}sec (Schwarz, 1965; Hammes & Roberts, 1969; Cummings & Eyring, 1975). The lifetime of an α-helix is always shorter than the time for helix formation.

5.3.2.2 Diffusion of two helices together

Karplus & Weaver (1975) have considered the general diffusion behaviour of polypeptide chains. For helices, the time for diffusion of two connected helices into pairing proximity can be evaluated from Einstein's equation:

$$\langle \Delta X \rangle^2 = 2Dt \qquad (1.5)$$

If D is $10^{-6}cm^2sec^{-1}$, then the apparent diffusion coefficient for the relative motion of two linked helices might be $10^{-7}cm^2sec^{-1}$. If $\Delta X = 20\text{Å}$, $t = 2 \times 10^{-7}$sec, slightly faster than the rate of helix formation (Baldwin, 1979).

5.3.2.3 Helix docking

The docking of two α-helices should stabilize the helices involved. Pairing requires optimising the fit of the van der Waals surface subject

to electrostatic constraints (see 6.1.1 and 6.1.2). The effective concentration of proximal pairs of helices is not likely to exceed 10^{-3}M and since the rate constant for sticking will probably not exceed $10^{9}M^{-1}s^{-1}$, the time range for helix pairing is 10^{-6}sec or slower (Baldwin, 1979). However, the rate of dissociation is likely to be much slower than the association rate as the complex is stabilised.

5.4 Possible Folding Intermediates

The most complete study of kinetically trapped intermediates during protein folding is Creighton's work on pancreatic trypsin inhibitor (for a review, see Creighton, 1978 and references therein). In this work, folding was compartmentalised by varying the concentration of reagents which promote and inhibit disulphide bridge formation. He found that two one-disulphide intermediates were favoured over the fifteen possibilities. Moreover, the formation of the native-like two-disulphide intermediate which led to the native structure was dependent on the formation of a non-native two-disulphide intermediate (see Figure 1.21). The kinetic barriers for the folding of pancreatic trypsin inhibitor as well as the relative stability of various intermediates was also determined (see Figure 1.22).

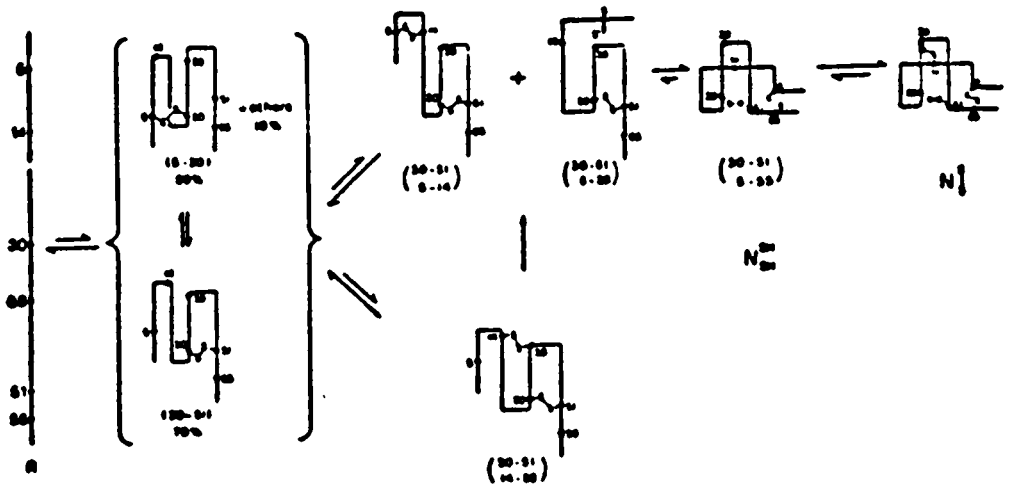
6. Theoretical Approaches to the Protein Folding Problem

There are two distinct aspects to the problem of folding a polypeptide chain: (1) the specification of the initial and final states, and (2) the nature of the path actually followed during the folding process. In principle, proper application of known chemical interactions between all parts of the full covalent structure of the peptide would supply both the specifications of states and the path(s) simultaneously. Such a

FIGURE 1.21

Schematic folding pathway of pancreatic trypsin inhibitor

(Taken from Creighton, 1978).



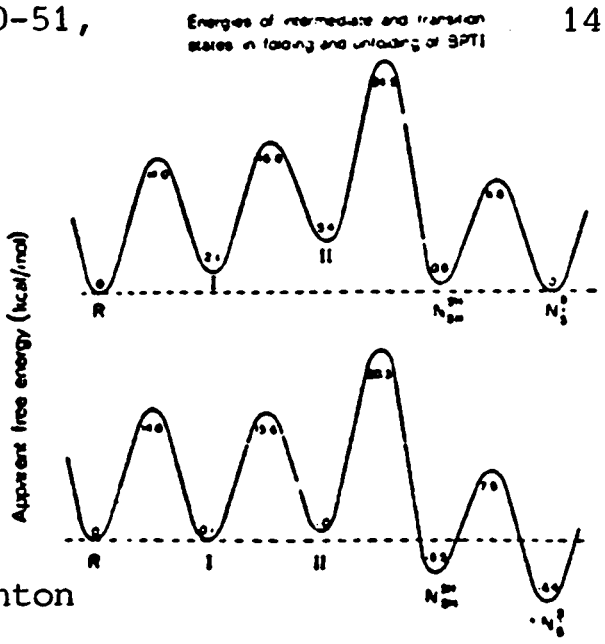
Schematic diagram of the pathway of folding and unfolding of BPTI. The solid line represents the polypeptide backbone, with the positions of the six cysteine residues indicated. The configurations of species N^{SH}_{SH} and N^{S}_{S} approximate the folded conformation of native BPTI; those of the others are rather arbitrary, except for the relative positions of the cysteine residues involved in disulphide bonds. The numbers of the cysteine residues involved in the disulphide bonds are shown below each diagram of the intermediates.

The brackets around the single-disulphide intermediates indicate that they are in rapid equilibrium; only the two most predominant species are indicated, along with the approximate composition of the normal spectrum of species. The "+" between intermediates (30-51, 5-14) and (30-51, 5-38) signifies that both are formed directly from the single-disulphide intermediates, that both are converted directly to N^{SH}_{SH}, and that either or both are intermediates in the rearrangement of (30-51, 14-38) to N^{SH}_{SH}.

FIGURE 1.22

Apparent Free Energy of the Refolding of Pancreatic Trypsin Inhibitor

(Taken from Creighton 1978).



Apparent free energy diagrams of the intramolecular transitions in the folding and unfolding of BPTI. R is fully reduced BPTI, I the normal spectrum of single-disulphide intermediates, II the mixture of (30-51, 14-38) (30-51, 5-14) and (30-51, 5-38), N^{SH}_{SH} is (30-51, 5-55) in the native-like conformation and N^{S}_{S} is fully folded BPTI with three disulphide bonds. The apparent free energies of the various states and of the intervening transition states, in kcal/mole relative to that of R, were calculated from the intramolecular rates of forming and rearranging disulphide bonds and from the rates of reduction of the disulphide bonds by DTT^{SH}_{SH} unde theoretical conditions in which N^{S}_{S} has either (a) the same energy as R, or (b) is 6.4kcal/mole more stable. The relative free energies of N^{S}_{S} and R depend upon the stabilities of the three protein disulphide bonds, which may be varied collectively by changing the ratio of disulphide to thiol reagent in the solution. In the upper diagram (a), the ratio of DTT^{S}_{S} to DTT^{SH}_{SH} is 26:1, while in the lower diagram (b) the ratio is 10^3:1, making each of the disulphide bonds 2.1kcal/mole more stable than in the first instance.

complete solution is not yet possible. Nemethy & Scheraga (1977) have provided an extensive review of the whole field of protein folding studies.

6.1 The energetics of a polypeptide chain

Unfortunately, the number of electrons and nuclei in a polypeptide chain make a complete quantum mechanical description of the energy associated with a particular conformation not feasible. Instead approximate functional forms for various contributions to the stability of a conformation have been suggested. These functional forms are constructed to mimic the potential surface of small molecules whose energy can be calculated using various quantum mechanical techniques. (For the nature of the approximations inherent in these calculations, see the review by Pullman & Pullman, 1972.) Normally, calculations of the energy of a polypeptide chain are partitioned into seven components. These components are presented in order of the relative enthalpic contributions (see Figure 1.23).

6.1.1 Bond length and angle

Semi-empirical quantum mechanical calculations can be used to evaluate the energy of a bond as a function of bond length and angle (see Figure 1.24). Given the shape of the potential well, we can approximate the energy of distortion by a Hookean function:

$$E_{dis} = (K/2)(r - r_{eq})^2$$

where r_{eq} is chosen so that:

$$\left.\frac{\partial E}{\partial r}\right|_{r=r_{eq}} = 0$$

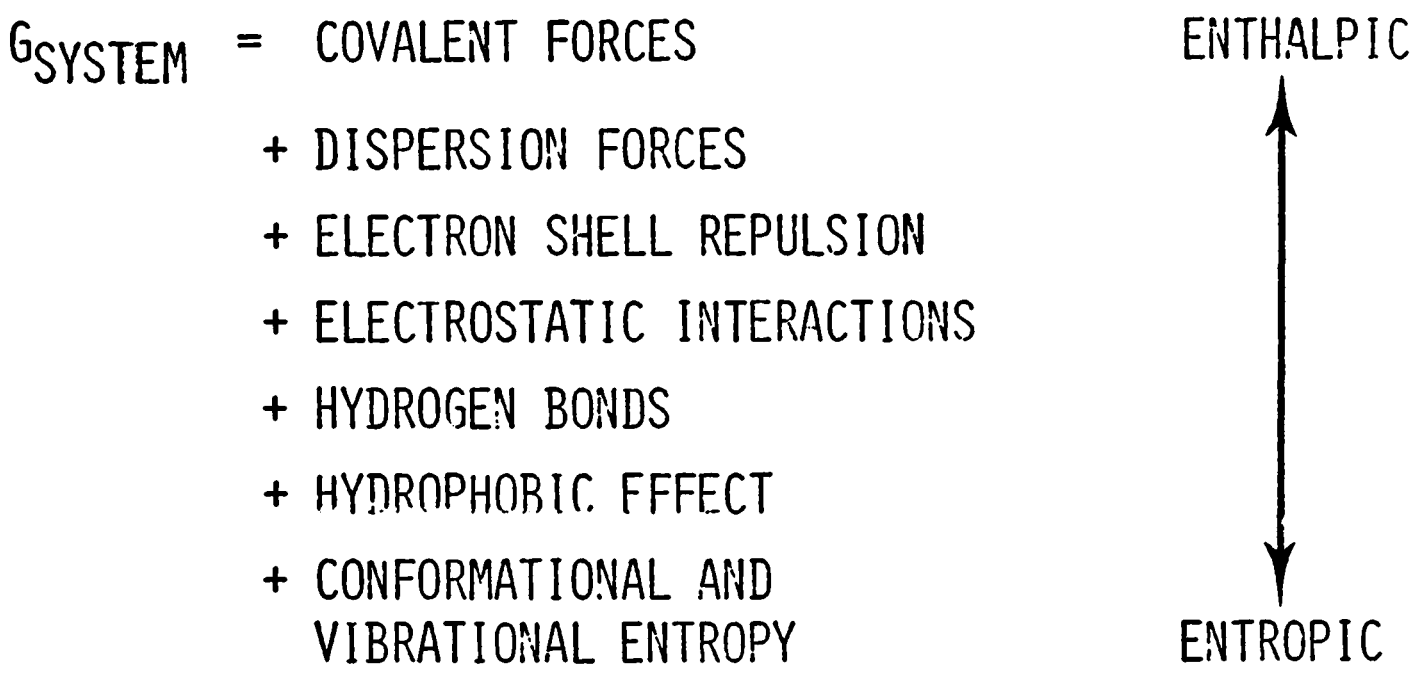
and K is computed as:

$$K = \left.\frac{\partial^2 E}{\partial r^2}\right|_{r = r_{eq}}$$

K and r_{eq} can also be determined experimentally from spectroscopy studies

FIGURE 1.23

THERMODYNAMICS OF PROTEIN FOLDING



The forces of protein folding are sorted from most enthalpic to most entropic. Section 6.1 is organised along these lines.

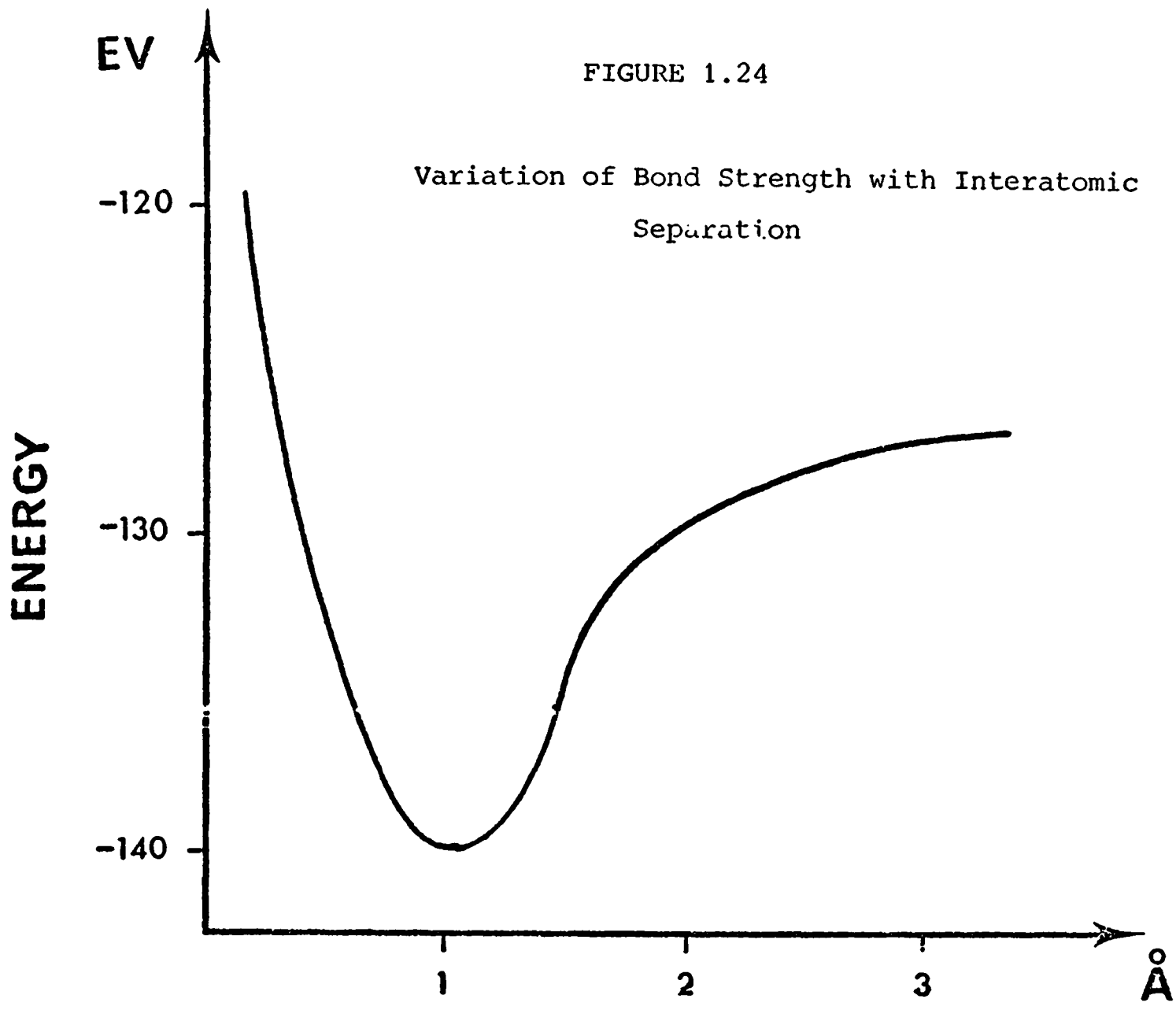


FIGURE 1.24

Variation of Bond Strength with Interatomic Separation

This is the approximate energy profile of a carbon hydrogen bond with an optimal bond length of 1Å. The well is parabolic locally about the minimum.

This Figure is re-drawn from Hopfinger (1973).

(Hopfinger, 1973).

6.1.2 Torsional

The total molecular energy of ethane $E(\chi)$ (see Figure 1.25) displays a periodic form. In general, a single bond torsional potential can take the form:

$$E(\chi) = E^*(1 \pm \cos(n\chi - b))$$

where E^* is the barrier height, n is the periodicity of the function and b is the phase angle. This form is most accurate when the groups bound to an atom are equivalent and additional corrections are required when different substituents change the relative heights of each barrier.

The actual barrier height, E^*, can be determined from temperature dependent NMR experiments (Roberts, 1955).

6.1.3 Van der Waals: Dispersion and Repulsion

A van der Waals potential often used to describe the non-bonded forces is:

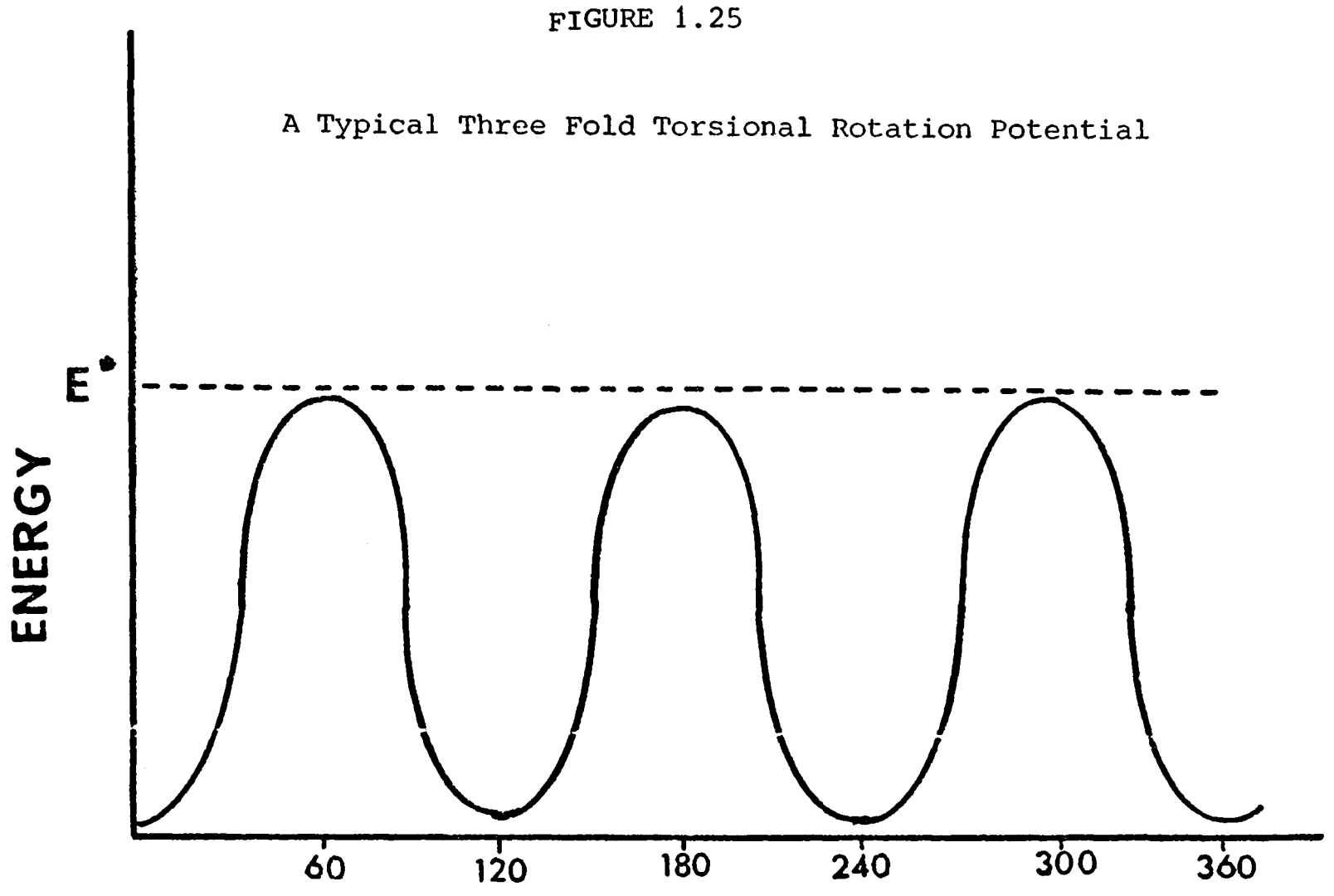
$$E = Ar^{-12} - Br^{-6}$$

where B is the coefficient of the attractive term and A, the repulsive term. B reflects the polarisability of the electron clouds and hence the favourable induced dipole-dipole interaction. A measures the steric complications of bringing one atom near another without orchestrating the overlap of bonding orbitals. This is largely a coulombic repulsion.

A sample plot of E(r) is given in Figure 1.26. Functional forms are chosen to fit the shape of the potential surface produced by quantum mechanical calculations. B can be calculated theoretically and A is constructed so that the $\partial E/\partial r\big|_{r=r_o} = 0$ where r_o is the ideal inter-atomic separation. The contribution of each atom to r_o is known as its van der Waals radius.

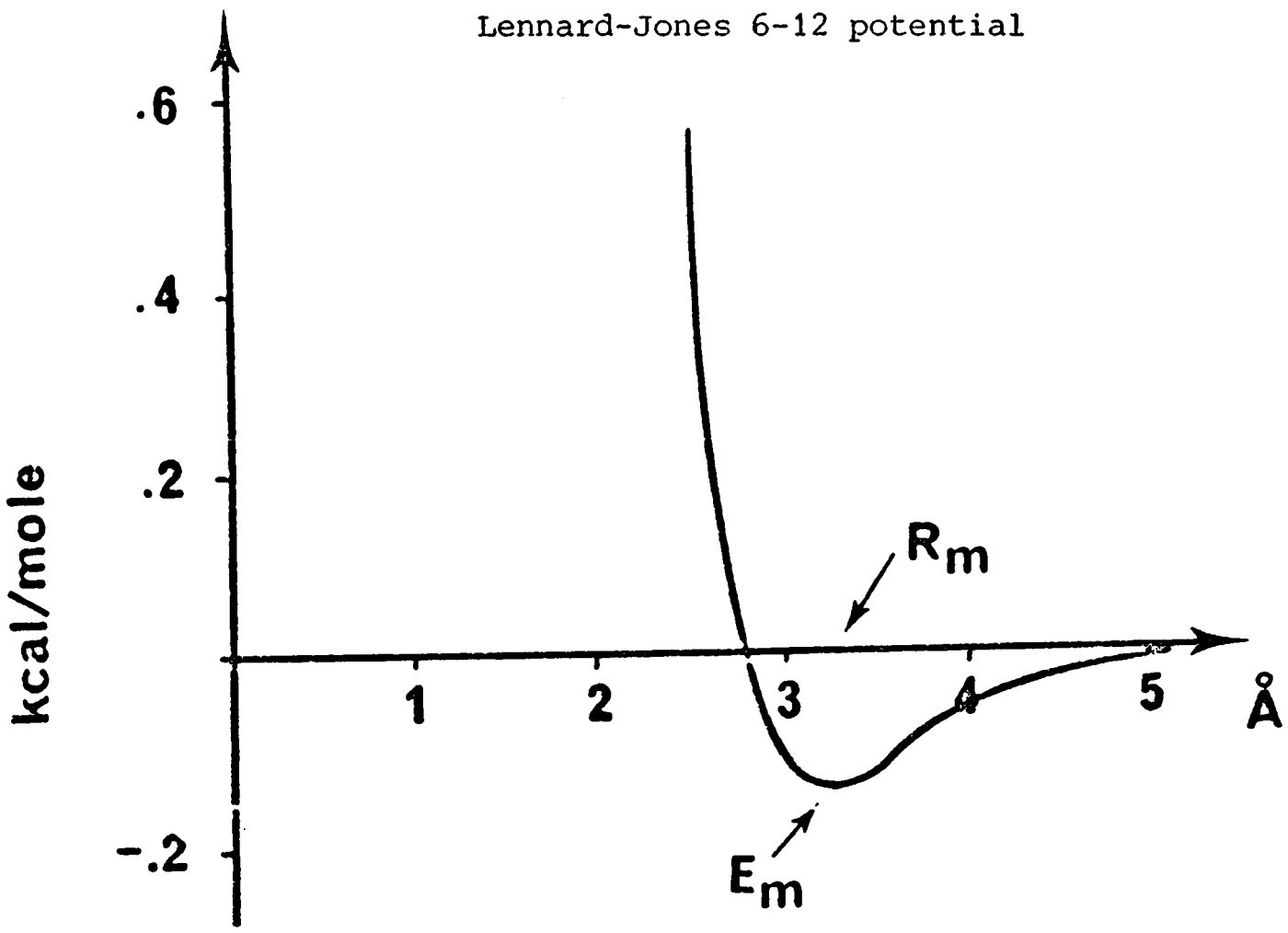
FIGURE 1.25

A Typical Three Fold Torsional Rotation Potential



This is the energy spectrum of ethane as a function of torsional rotation about the carbon-carbon bond. E* is the barrier height. As ethane has 3-fold rotational symmetry, the potential well resembles a plot of sin3θ. Equal barrier heights follow from the symmetry.

FIGURE 1.26



Lennard-Jones 6-12 potential for dispersion forces and electron repulsion (R_m = 3.24Å, E_m = -0.13kcal/mol). The two-parameter formula is given below in its computationally practical form (parameters A, B) and in its normalised form (parameters E_m, R_m). Because of repulsion between electronic shells and attraction by dispersion forces, A and B are positive, and there exists a relative minimum. With smaller B and larger A, E_m becomes smaller and the corresponding atomic distance R_m larger. The potential wall is asymmetric. Repulsion balances attraction when the atomic distance is reduced to $0.89R_m$. On the other side the attraction energy is still one-sixth of E_m at a distance of $1.5R_m$.

$$E = \frac{A}{R^{12}} - \frac{B}{R^{6}} = E_m \left\{ -\left(\frac{R_m}{R}\right)^{12} + 2\left(\frac{R_m}{R}\right)^{6} \right\}$$

$$E_m = -\frac{B^2}{4A} \qquad R_m = \sqrt[6]{\frac{2A}{B}}$$

This Figure is re-drawn from Schulz & Schirmer (1979).

6.1.4 Electrostatics

The potential function central to all electrostatic calculations is due to Coulomb:

$$E = \frac{332}{\varepsilon} \frac{q_1 q_2}{r_{12}}$$

where ε is the dielectric constant of the medium, q_1 and q_2 are partial charges relative to the fundamental charge of an electron and r_{12} is the distance between the partial charges in Å.

In isolation, atoms are neutral. In covalent bonds, the relative electronegativities of the bonded atoms imply a partial positive charge on one atom and a partial negative charge on the other. The size of the partial charges for atoms in a polypeptide have been approximated by CNDO/2 calculations (Momany et al., 1975), ab initio molecular orbital calculations on small molecules (Hagler & Lapiccirela, 1976) or by fitting observed crystal data (Hagler et al., 1974) (see Table 1.2).

Of course, polyelectronic atoms are not monopoles. When the distance between charges in an atom is similar in magnitude to the interatomic separation of charges, dipole, quadrupole and higher order terms in the electrostatic series must be added to improve the value of calculated energies.

6.1.5 Hydrogen Bonding

While the existence of a hydrogen bond is supposed whenever the separation between donor and acceptor is less than the sum of the van der Waals radii of the atoms involved, some uncertainty about hydrogen bonding geometries exists (see Figures 1.27 and 1.28). A general discussion of hydrogen bonding is provided by Pauling (1960).

Explicit quantum mechanical calculations verify that hydrogen bonded arrangements are energetically favourable. Pullman and co-workers (e.g. Dreyfus & Pullman, 1970) have shown that the hydrogen bond can be

TABLE 1.2

Partial Charges of Atoms in the Polypeptide Backbone and in Three Side Chains[a]

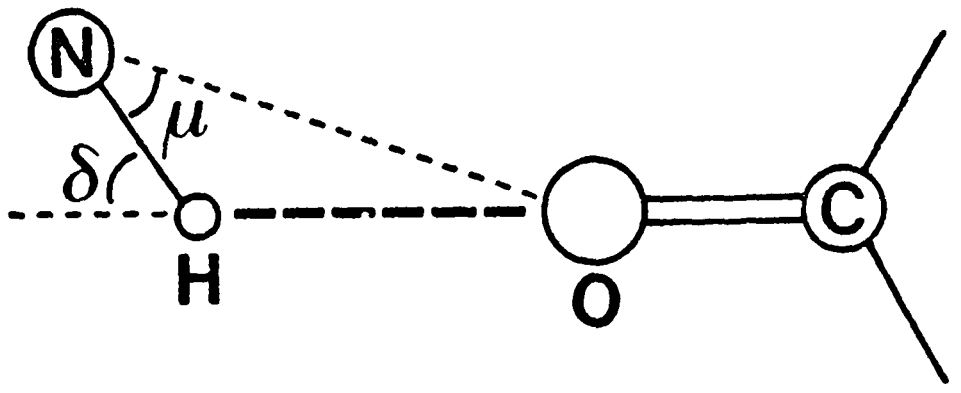
(Taken from Schulz & Schirmer, 1979)

Peptide	N	-0.36	Ser	C_β	+0.13
Peptide	H_N	+0.18	Ser	H_β	+0.02
Peptide	C_α	+0.06	Ser	O_γ	-0.31
Peptide	C'	+0.45	Ser	H_γ	+0.17
Peptide	O	-0.38			
Peptide	H_α	+0.02	Tyr	O_η	-0.33
			Tyr	H_η	+0.17
			Asn	C_β	-0.12
			Asn	H_β	+0.06
			Asn	C_γ	+0.46
			Asn	O_δ	-0.38
Cys	S_γ	+0.01	Asn	N_δ	-0.45
Cys	H_γ	+0.01	Asn	H_δ	+0.20

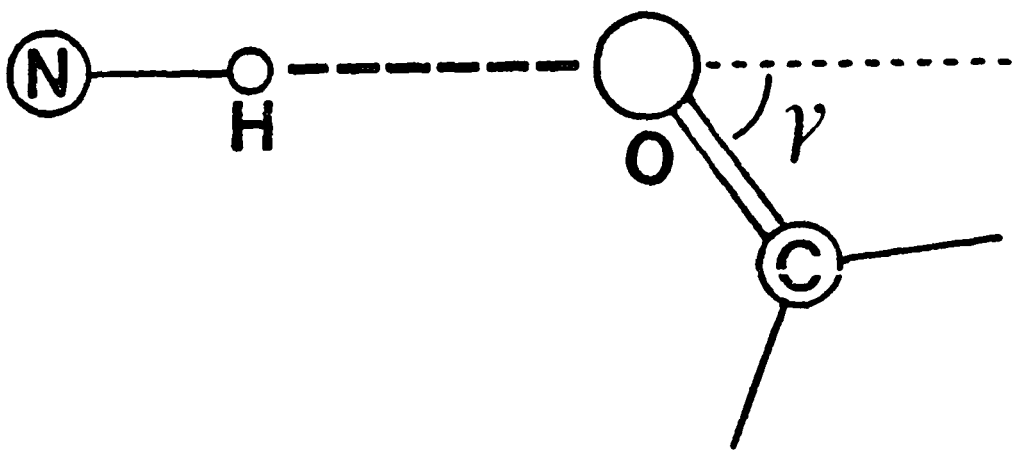
[a]Taken from the results of Momany et al. (1975) who derived the partial charges of the 20 common amino acid residues using the CNDO/2 (complete neglect of differential overlap) method. Partial charges were also derived from ab initio molecular orbital calculations of small molecules or by fitting observed crystal data. A comparison between the results shows that accuracy is low.

FIGURE 1.27

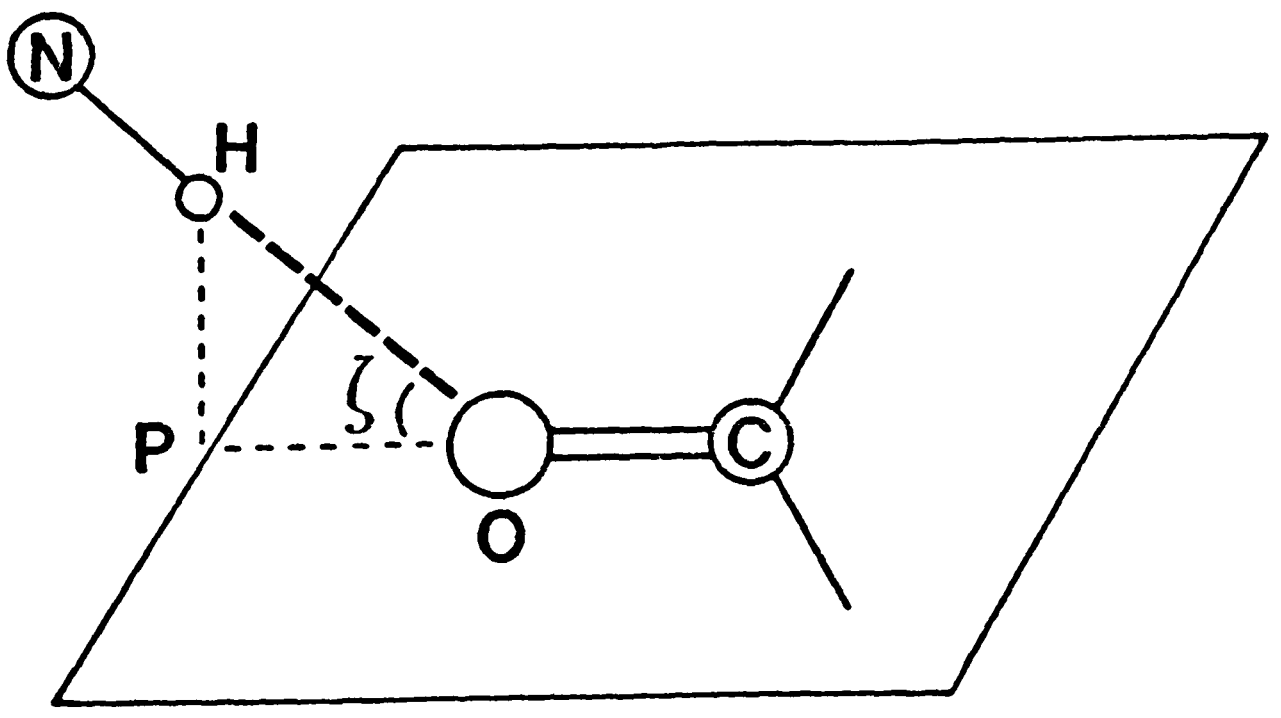
The Parameters of Hydrogeon Bonding



(a)



(b)

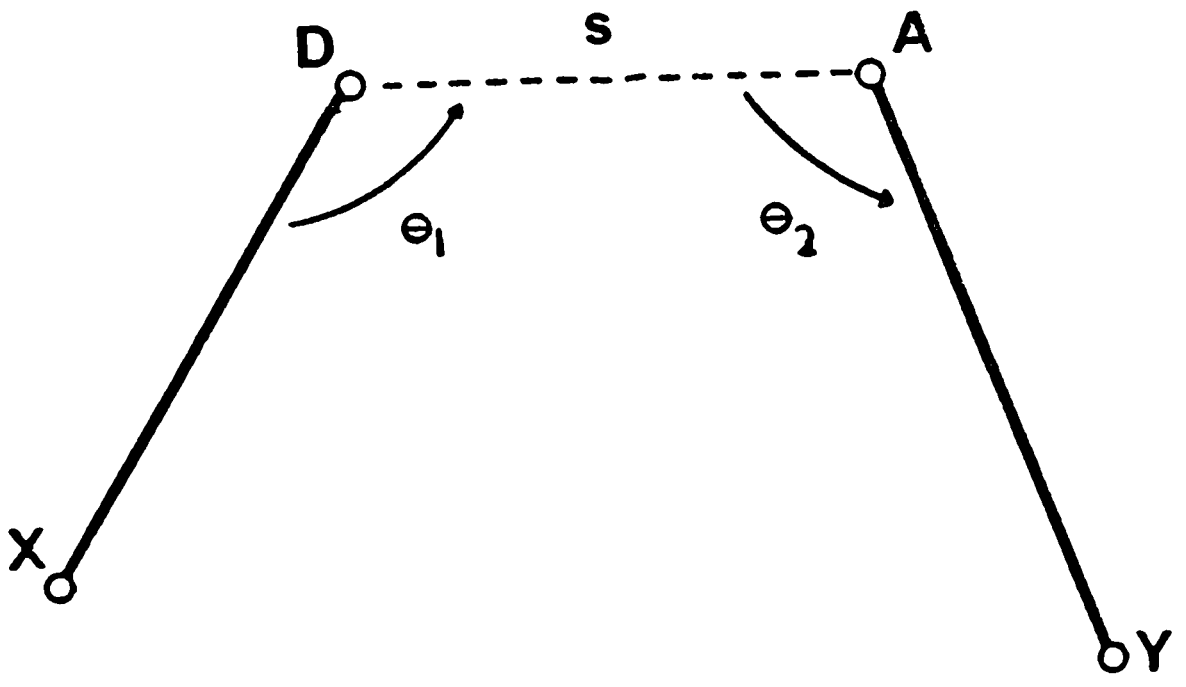


(C)

The stability of a hydrogen bond is affected by
(a) m and f, the non-linearity of the hydrogen bond
(b) ν , the deviation from linearity of the H-C-O angle
and (c) ζ , the lack of planarity of the bond.

FIGURE 1.28

Hydrogen Bonding Geometry



The geometry of the hydrogen bond. D denotes the donor and A denotes the acceptor atom. The symbol s indicates the distance between the donor and the acceptor.

This Figure is re-drawn from Hopfinger (1973).

worth -7 to -9½ kcal/mole *in vacuo*. Given the cost of the calculations, analytic potentials have been sought. Scheraga *et al*. (1967) expressed this energy as a function of the three natural geometric parameters (see Figure 1.28).

$$E(S,\theta_1,\theta_2) = -D^*\cos^2(\theta_1)\exp\left\{\frac{-m^*(R-S-r_o^*)^2}{2(R-S)}\right\} - D^*\cos^2(\theta_2)\exp\left\{\frac{-m^*(R-S-r_o^*)^2}{2(R-S)}\right\} + A\exp(-bR) - \tfrac{1}{2}A\left(\frac{R_o}{R}\right)^m \exp(-bR_o)$$

where D* is the strength of the hydrogen-acceptor interaction, m is an adjustable exponent, m* the variable parameter proportional to the ionisation potential of the hydrogen, R the donor-acceptor separation, r_o^* the equilibrium hydrogen-acceptor separation, A the strength of the donor acceptor potential, b the "hardness" of the donor acceptor potential, and R_o the equilibrium donor-acceptor separation, with $|\theta_1,\theta_2| > \pi/2$. An alternative potential form employs a van der Waals plus electrostatic term together with a small correction for the angular dependence of a hydrogen bond (Gibson & Scheraga, 1967):

$$E(S,\theta_1,\theta_2) = \frac{-a}{S^6} + \frac{b}{S^{12}} + k\frac{Q_{donor}Q_{acceptor}}{\varepsilon S} - \frac{G}{S^m}f(\theta_1,\theta_2) \quad .$$

Artymiuk (1979) has examined the distribution of hydrogen bond geometries in human lysozyme. The data is consistent with theoretical expectations about hydrogen bond lengths and deviations from planarity.

6.1.6 Hydrophobic

The propensity of proteins to restrict the motion of water molecules locally while simultaneously disrupting hydrogen bonds often leads to a partition between water and protein. This entropic effect is often invoked to explain the stability of a globular protein over its denatured

conformer. The free energy difference between these two states is often small, between 5 - 15 kcal/mole (Pace, 1975). The concept of hydrophobicity in biological systems has been discussed by Kauzmann (1959), Kuntz & Kauzmann (1972) and Tanford (1973).

Gibson & Scheraga (1967) developed a hydration shell model to approximate macromolecule-solvent interactions indirectly. The inaccuracies in this model have hampered its success in folding studies. Robson & Osguthorpe (1979) modify their van der Waals parameters to include solvent effects but the result of this work is not conclusive. The most accurate way to treat solvent effects is to include water molecules explicitly. Monte Carlo simulations of the protein solvent interface are very interesting but remain too time-consuming for a folding study (Hagler & Moult, 1976).

6.1.6.1 Non-polar accessible contact area.

Of the 20 amino acids typically seen in globular proteins, roughly half are hydrophilic and half hydrophobic. Historically this led to an "oil droplet" model of a protein with hydrophilic residues exposed to the solvent and hydrophobic residues buried within the globule (e.g. Fischer, 1964). However, as the data base of detailed protein structures increased, it became clear that the steric restrictions of the chain made a perfect partitioning of hydrophobics and hydrophilics impossible.

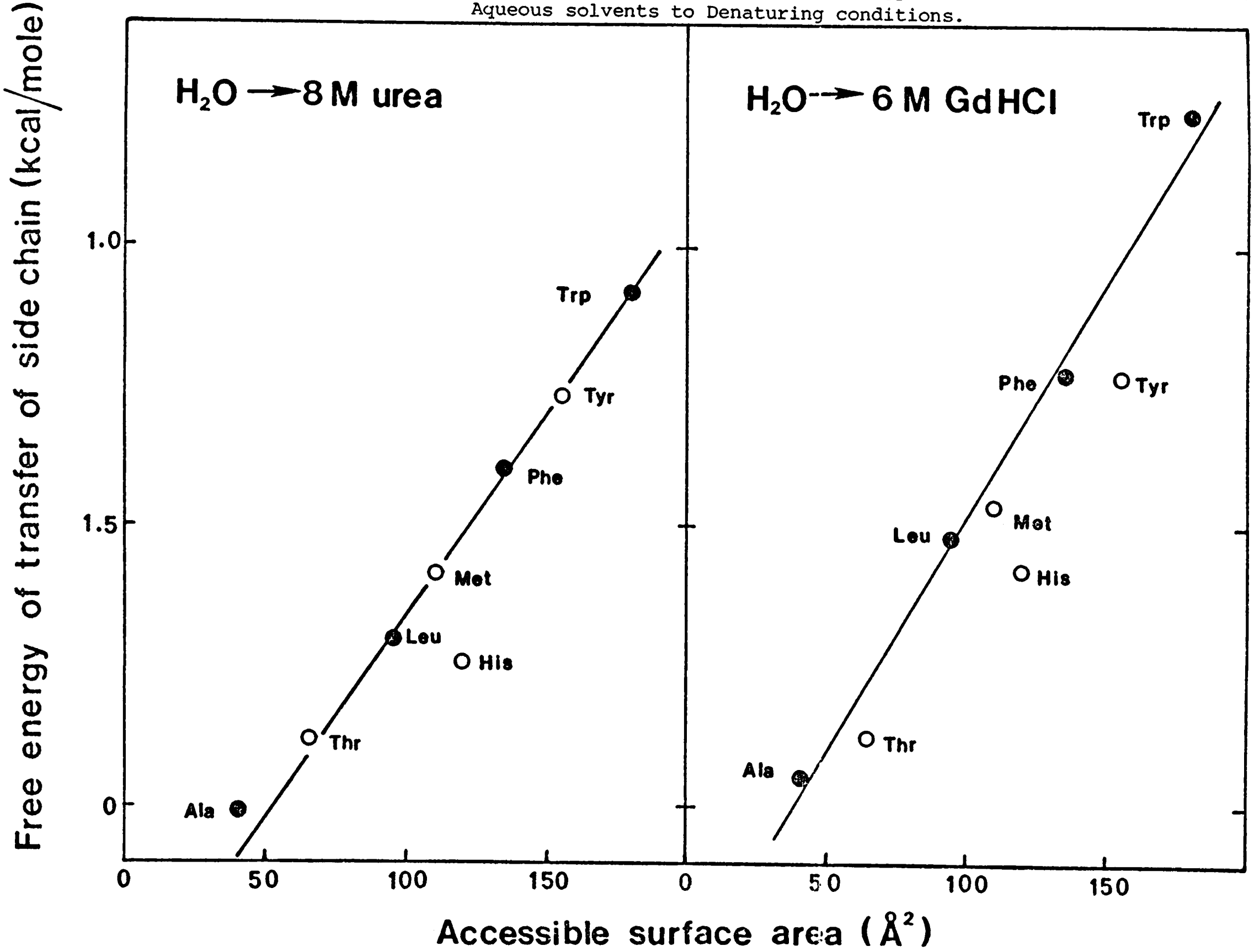
Still, the tendency for a folded protein to bury hydrophobic residues was evident. Lee & Richards (1971) quantified this tendency by measuring the area of the surface traced by the locus of the centre of a water molecule rolling along the van der Waals surface of a protein (see Figure 1.10). Richmond & Richards (1978) redefined this concept slightly by introducing the concept of accessible contact area. This is the area of the van der Waals surface of a protein which could contact a hypothetical water probe. The non-polar accessible contact area is that fraction of the accessible contact area of the protein which is due to

sulphur atoms and carbon atoms excluding the carbonyl carbon. This area is evaluated by computing the intersection of equally spaced z-sections of coordinate space with the van der Waals surface of the molecule and summing the lengths of the arcs in contact with a hypothetical water probe with a radius of 1.4Å as it rolls along this surface slice. This arc length is integrated numerically over all sections.

When a protein folds, its non-polar accessible contact area decreases. This led Chothia (1974) to plot accessible surface area of an amino acid against the free energy change for the transfer of that amino acid from an aqueous to non-polar medium measured by Nozaki & Tanford (1971). The slope of the least squares line through these points suggested the empirical relation: 23cal ≡ 1Å² (Chothia, 1974; see Figure 1.11). For non-polar accessible contact area, this relation is 80cal ≡ 1Å² (Richmond & Richards, 1978). Creighton (1979) observed that when the transfer experiments were done from a non-polar solvent to 6M guanidine hydrochloride (GdHCl), the slope of the line relating accessible surface area to transfer free energies decreased to 8cal ≡ 1Å² (see Figure 1.29). As GdHCl is a protein denaturant which is thought to interrupt the extensive hydrogen bonding of water, it is logical that the hydrophobic effect should decrease. The correlation seen by Chothia (1974) and the GdHCl plot suggest that the hydrophobic contribution to the free energy of protein folding can be approximated by the change in non-polar accessible contact area. Although surface tension effects are undoubtedly involved in creating an energetically reasonable partition between the protein and the solvent, the correlation of accessible surface area and free energy is best considered as an empirical relationship which provides a useful approximation.

In a recent development, Richmond (1980) has defined an analytic expression for accessible contact area as a function of atomic position and van der Waals radii. The derivatives of this function can be evaluated

FIGURE 1.29. The Correlation between Accessible Surface Area to the Free Energy of Transfer of Amino Acids from Aqueous solvents to Denaturing conditions.



These plots suggest a drop in the relationship Chothia (1974) observed, $1Å^2 \equiv$ 23cal to $1Å^2 \equiv$ 8cal as a result of the denaturants guanadinium hydrochloride and urea.

This Figure is redrawn from Creighton (1979).

and may prove very useful in future energy minimisation studies.

6.1.7 Entropic Contributions to the Polypeptide Potential Surface

Typically, energies calculated for various polypeptide conformations ignore the entropic contribution to the free energy of the system. Thus, searches for equilibrium conformations rely on the approximation:

$$\Delta G \simeq \Delta H$$

The Gibbs free energy is approximately equal to the enthalpy at the temperature considered. The Gibbs equation is:

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

To the extent that $|T\Delta S|$ is large, this approximation errs.

The vibrational entropy for each mode of motion is:

$$S_{vib} = R\left\{ \frac{x}{e^{x} - 1} + \ln \frac{1}{1 - e^{-x}} \right\}$$

where

$$x = h\nu_{vib}/kT$$

Most vibrational modes within proteins are high frequency (e.g. bond stretching or bond angle bending). Unfortunately, hydrogen exchange data suggests that there must be some low frequency modes. In PTI, most amide protons exchange rapidly with the solvent, but others require more than 1 second to exchange (Wutrich & Wagner, 1979). If these intervals are taken as vibration frequencies, then $h\nu$ is very small and $S_{vib} \sim 30$ e.u. at 300°K. Although this calculation probably overestimates S_{vib} for this mode, $T\Delta S$ could be 90 kcal/mole. Even if this calculation is wrong by an order of magnitude, this is large relative to the free energy of stabilisation for globular proteins, 5 - 15 kcal/mole (Pace, 1972).

Sturtevant (1977) has analysed the relative contributions of enthalpic and entropic terms from calorimetric studies of protein folding. Entropic contributions to protein free energy estimated from changes in

TABLE 1.3

Entropy Change on Protein Unfolding at 25°C.

Protein	ΔS^a cal $K^{-1}mol^{-1}$	ΔG^b kcal mol^{-1}
α-Chymotrypsin	330	11.0
Cytochrome C	6.2	8.7
Lysozyme	140	14.0
Metmyoglobin	300	12.0
Ribonuclease	215	9.5

[a] Taken from Sturtevant (1977)

[b] Taken from Privalov & Khechinashvili (1974).

heat capacity are presented in Table 1.3. At 300°K, $T\Delta S$ is 90 kcal/mol for metmyoglobin. As the free energy of stabilisation of myoglobin is only 13.3 kcal/mol (Harrison & Blout, 1965) then stability must result from the small difference, ΔG, between two large quantities, $T\Delta S$ and ΔH.

6.2 Six Strategies for the Prediction of Protein Structure

Much effort has been spent developing more and more accurate potential functions. Three approaches to the problem rely on these potentials: Energy Minimisation, Molecular Dynamics and Monte Carlo Simulations. The other two approaches ignore the specific energetics of a protein folding and rely on geometric constraints: Distance Geometry & the Combinatorial Approach. Clearly allowed geometries must be energetically reasonable but relative energies for structures produced by these techniques are not easily discernible. The final method is Statistical. This approach has enjoyed some success in predicting the location of secondary structure but has yet to offer an algorithm for predicting tertiary structure.

Throughout this discussion of approaches to the protein folding problem, the root mean square (r.m.s.) deviation will be presented as a measure of the studies' success. This value is commonly computed as:

$$(\Delta d)^2 = \sum_i \sum_j \frac{(d_{ij} - \ell_{ij})^2}{n^2}$$

where d_{ij} and ℓ_{ij} are entries in the matrix of interatomic distances. A complete discussion of structural comparisons will appear in Chapter 2.

6.2.1 Energy Minimisation

A stable equilibrium position requires that all local perturbations of the system result in an increase in energy and a tendency to return to the equilibrium position. Thus, the crystallographically determined

structure of a protein must approximate* a state where:

$$\frac{\partial E}{\partial x_i} = 0 \qquad \text{for all atoms } i.$$

The energy terms described in the preceding section constitute the potential function which is evaluated and minimised for a protein of interest. Typically pancreatic trypsin inhibitor (PTI) is chosen because it is only 58 residues long and a high resolution crystal structure is available (Deisenhofer & Steigemann, 1974).

The implicit problems with this technique are three-fold. The potential functions available are still approximations and the solvent is acknowledged in only the crudest manner or totally ignored. Due to the trignometric functions in the torsional term and the non-specificity of the van der Waals effect, many local minima distinct from the global minimum exist which confuse minimisation procedures. Finally, the number of degrees of freedom is large (e.g. >150 for PTI) so that detailed calculations are computationally not feasible.

In spite of these difficulties, several investigators have attempted to energy minimise an extended conformation of PTI and reach the crystal structure. Every study uses methods for eliminating degrees of freedom and for simplifying potential functions. A random compact structure with 58 amino acids (e.g. PTI) would have an r.m.s. deviation of approximately 7Å (Δd) from the crystal structure (Cohen & Sternberg, 1980; and Chapter 2). To date, energy minimisation studies have failed to do significantly better than random. Still, many simplifications have been suggested which ultimately may prove useful.

Levitt (1976) simplified the polypeptide chain by coupling the behaviour of the virtual bond angle formed by three consecutive alpha

* *The crystal structure is a time-averaged picture of many states and this average may not correspond to any one physically reasonable conformation, although it will be similar to many of the conformations adopted by the crystalline protein.*

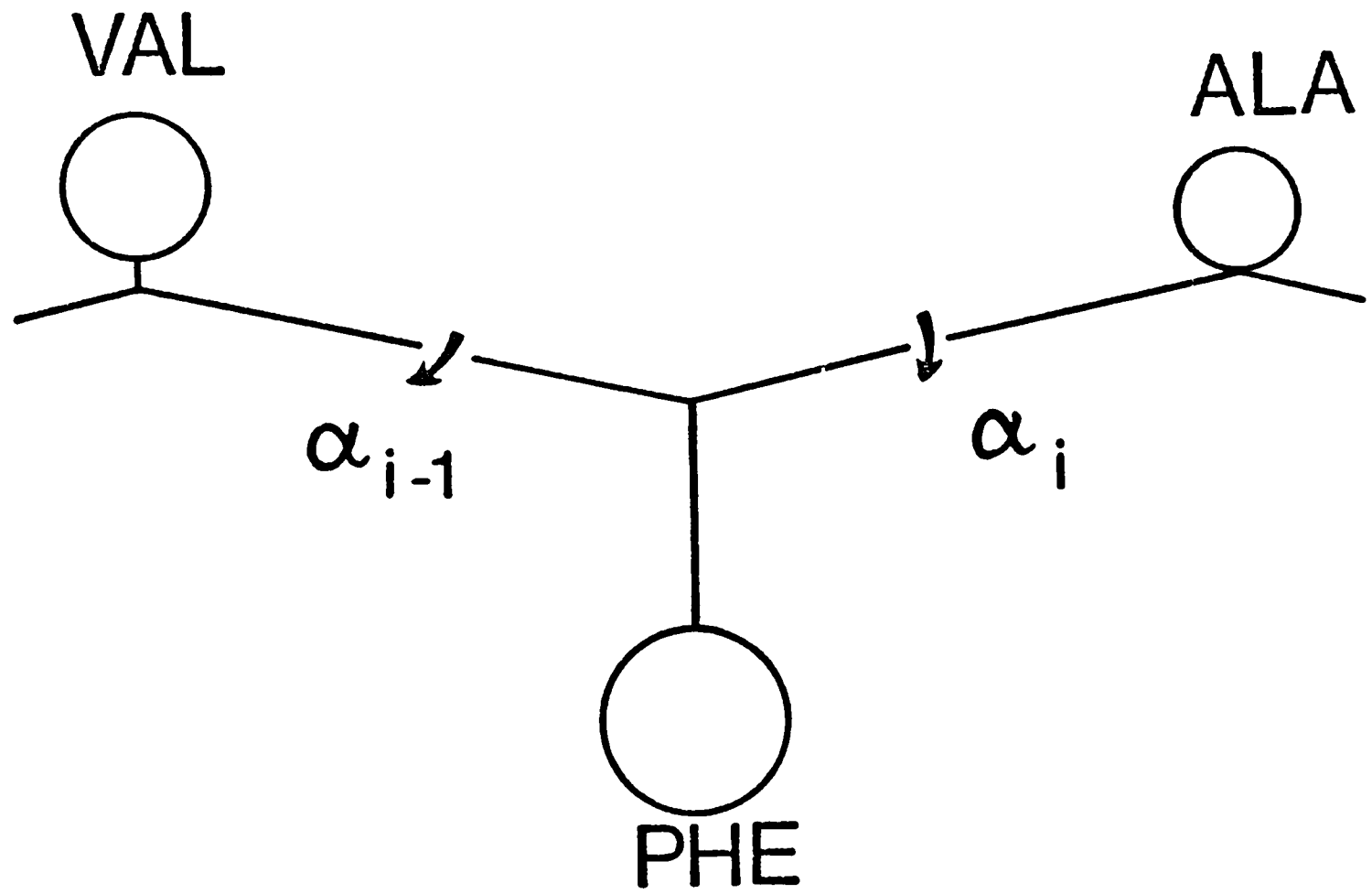
carbons, α_i, with the dihedral angle formed by four consecutive alpha carbons, τ_i. The backbone was fitted to the C_α positions and side chains were inserted according to a "lolly-pop" model (see Figure 1.30). The adjustable parameters in the potential function were determined for this simplified representation and minimisation attempted. To avoid false minima, pushing and pulling potentials as well as normal mode thermalisation were employed. Bend potentials were inserted at the flexible points in the chain. The r.m.s. deviation obtained for the minimised structure was $\Delta d = 8.5\text{Å}$. This value dropped to $\Delta d = 6.2\text{Å}$ when the helix was inserted before minimisation. The significance of this result was discounted by Hagler & Honig (1978). They achieve a comparable r.m.s. deviation by constructing a polyalanine chain with glycine residues in the positions where Levitt had introduced bend potentials.

Warshel & Levitt (1976) repeated the procedure of Levitt (1976) on carp parvalbumin (CPV). The r.m.s. deviations obtained were $\Delta d = 8.2\text{Å}$ without preset helices and $\Delta d = 7.4\text{Å}$ with preset helices. A compact random structure with 108 residues would have an r.m.s. deviation of $\Delta d = 8.3\text{Å}$ from the crystal structure (Cohen & Sternberg, 1980).

Robson & Osbuthorpe (1979) introduced a variable γ_i which linked the behaviour of the backbone dihedral angles ϕ and ψ. γ_i provides a more suitable parametrisation of backbone atom positions for sterically reasonable dipeptide conformations than α_i used by Levitt (1976). This halved the number of degrees of freedom. Solvent effects were included by modifying the van der Waals parameters to account for hydrophilic and hydrophobic tendencies. To avoid local minima, two minimisation algorithms with different convergence criteria were used. When one minimiser had converged, the other was called. Simultaneous convergence was achieved and the resulting structure had an r.m.s. deviation of $\Delta d = 6.0\text{Å}$ from the crystal structure.

Kuntz et al. (1976) chose to represent the polypeptide chain as

FIGURE 1.30. Levitt's (1976) Simplified Polypeptide Chain Geometry



The simplified chain geometry used by Levitt (1976) is a simulation of the refolding of Pancreatic Trypsin Inhibitor. The torsion angle α about the virtual bond between α-carbons is shown.

a chain of spheres with radii determined by the molecular volume of the residue. An extremely simple penalty function was developed to express the observed tendencies of different amino acids to lie near one another and to be internal or external. Minimisations were begun at a variety of extended conformations and the resulting structures showed r.m.s. deviations from $\Delta d = 4.7\text{Å}$ to $\Delta d = 6.5\text{Å}$ from the crystal structure.

6.2.2 Molecular Dynamics

An alternative to energy minimisation is to solve the equations of motion explicitly for the atoms in a polypeptide chain and follow the fluctuations of the chain from an unfolded to folded state. This implies solving the classical equations of motion:

$$\frac{\partial E_i}{\partial x_i} = m_i \frac{\partial^2 x_i}{\partial t^2}$$

$$\frac{\partial E_i}{\partial y_i} = m_i \frac{\partial^2 y_i}{\partial t^2} \qquad \text{for all atom } i$$

$$\frac{\partial E_i}{\partial z_i} = m_i \frac{\partial^2 z_i}{\partial t^2}$$

and integrating over an interval of time. The success of these calculations also depends upon the accuracy of the potential energy functions.

Currently, calculations are limited to the 10-100 picosecond range and so simulations of the real time folding of a protein are not possible. Instead Karplus and co-workers have used molecular dynamics calculations to examine the equilibrium fluctuations of proteins.

McCammon & Karplus (1979), in a 10ps simulation of the dynamic motions of PTI at 306K, studied the rotation of an internal tyrosine ring. Ring flipping events were observed in agreement with NMR findings. Karplus & McCammon (1979a) extended this simulation to 100ps and observed a variety of low frequency modes of motion: the radius of gyration of

the fluctuating state was 10.22 ± 0.1Å, significantly more compact than the crystal structure r_g = 10.96Å; the density increased to ρ_{100ps} = 0.63 daltons $Å^{-3}$ from $\rho_{crystal}$ = 0.60 daltons $Å^{-3}$; and the motions of the N and C termini were coupled. This study also confirmed the matrix damping effect of the protein on ring flipping.

Molecular dynamics seems to be the most reasonable representation of actual motion within proteins. Motions seen in these calculations are consistent with NMR and crystallographic data. Unfortunately, computation time is measured in hours even for picosecond simulations and protein folding requires at least a millisecond time scale.

6.2.3 Monte Carlo Simulations

Another procedure, closely related to Molecular Dynamics, is Monte Carlo simulation. The goal of the calculations is to reproduce the macroscopic properties of macromolecules by generating an ensemble of conformations representative of all possible states.

Monte Carlo calculations usually follow the procedure developed by Metropolis et al. (1953). This iterative algorithm contains 4 steps:

(1) Generate an initial state, i

(2) Include i in the weighted average $$\langle X \rangle = \frac{\sum X \exp(E_i/kT)}{\sum \exp(E_i/kT)}$$

where X is a macroscopic property and E_i is the energy of state i

(3) Introduce a small perturbation of state i to create state i+1

(4) a) If $E_{i+1} \leq E_i$, then i+1 becomes the current state. Return to step (2).

b) If $E_{i+1} > E_i$, then generate a positive random number ρ

b1) If $\exp(E_{i+1}/E_i) < \rho$ then state i+1 becomes the current state. Return to step (2).

b2) If $\exp(E_{i+1}/E_i) \geq \rho$ then state i+1 is rejected and i remains the current state. Return to step (2).

The resulting set of structures follow a Boltzman distribution of energy and there exists a Markov chain between states.

Hagler & Moult (1978) used the Metropolis Algorithm (Metropolis et al., 1953) to simulate the location of water molecules along the protein-solvent interface of lysozyme. They found highly ordered water near the protein surface with positions similar to the bound water molecules found in the crystallographic analysis. The first hydration shell, the most tightly bound water molecules, extends 3.8Å from the surface. Additional solvation shells are more difficult to assign.

Gō and co-workers (e.g. Gō & Taketomi, 1979) have conducted Monte Carlo simulations of protein folding to study the transition between the folded and unfolded state. Protein geometries are represented by two or three dimensional cubic lattices where vertices are C_α positions. The relative positions of atoms follows quickly from the indices of an array which contains the location of each atom. The simplified potential function is a correct list of nearest neighbour residues with 1 unit for each correct pairing. The temperature dependence of the kinetics of protein folding is simulated well with the two dimensional model but the three dimensional model which should be a more reasonable approximation fails to exhibit the desired behaviour.

Monte Carlo calculations seem to suffer from the same problems seen with Molecular Dynamics: calculation times are large because more than 100,000 conformations are often required to avoid statistical bias and the results can be very sensitive to changes in the form of the potential function.

6.2.4 Distance Geometry

Probable conformations of atoms can be defined without reference

to specific energy functions. Kuntz and co-workers (e.g. Kuntz et al., 1978) have exploited the properties of the matrix of interatomic distances to contain simultaneously experimental and theoretical information for studies of macromolecular conformations. In its exact form, the interatomic distance matrix is equivalent to a list of three dimensional coordinates. However, most experimental and theoretical methods for quantifying distances can only offer upper and lower bounds. The mathematical core of the Distance Geometry approach is the algorithm by Crippen (1977) for determining a set of coordinates which is consistent with a set of upper and lower bounds on the entries in the matrix of interatomic distances.

The algorithm follows 6 steps:

(1) Set up the upper and lower bound matrices $[u_{ij}]$ and $[\ell_{ij}]$

(2) Use the triangle inequality to make the boundary matrices consistent

(3) Randomly choose a matrix $[d_{ij}]$ such that $u_{ij} \geqslant d_{ij} \geqslant \ell_{ij}$ for all i and j

(4) Calculate the matrix $[g_{ij}]$ where $g_{ij} = \frac{1}{2}(d^2_{i0} + d^2_{j0} - d^2_{ij})$ and

$$d^2_{i0} = \frac{1}{n}\sum_{j=1}^{n} d^2_{ij} - \frac{1}{n^2}\sum_{j=2}^{n}\sum_{k=1}^{j-1} d^2_{jk}$$

(5) Calculate the 3 largest eigenvalues, λ_1, λ_2 and λ_3 together with their corresponding eigenvectors ω_1, ω_2 and ω_3. Then compute the new list of coordinates:

$$x_i = \lambda_1^{\frac{1}{2}}\omega_{1i}$$

$$y_i = \lambda_2^{\frac{1}{2}}\omega_{2i}$$

$$z_i = \lambda_3^{\frac{1}{2}}\omega_{3i}$$

(6) Compute $[d_{ij}]$ from the new list of coordinates and refine the positions to fit the constraints.

Havel et al. (1979) have used this technique to examine the effect

of various experimental and theoretical constraints on the protein folding problem. With PTI as an example, a random structure would be $\Delta d = 7.1$Å r.m.s. from the crystallographic coordinates. They found that adding the exact location of α-helices and strands still produced structures which averaged $\Delta d = 7.0$Å r.m.s. from the crystal structure. A complete specification of the distance of all C_α atoms from the centroid, which a simple hydrophobic potential might produce, left an average r.m.s. deviation of $\Delta d = 5.0$Å. When the number of specified interatomic distances was equal to the number of residues, the r.m.s. deviation averaged $\Delta d = 3.0$Å. Their final finding was that the extensive but qualitative assignment of all interatomic distances being greater or less than 10Å forced the generated structures to have average r.m.s. deviations of $\Delta d = 1.1$Å.

Kuntz et al. (1979) used the distance geometry procedure on PTI with purely theoretical information. They created a turn prediction algorithm and added a hydrophobic potential to construct an interatomic distance matrix. With the S-S bridges intact, the r.m.s. deviations from the crystal structure ranged between $\Delta d = 4$Å and 8Å with an average of $\Delta d = 6.6 \pm 0.6$Å.

Goel & Ycas (1979) have developed an alternative approach to the folding problem which is essentially similar to the distance geometry calculations of Kuntz and co-workers. A hydrophobic term which defines the distance from the centroid of various residues is coupled with local geometry (average interatomic distances between 5 consecutive residues) and disulphide bridges to yield a penalty function which reproduces the observed statistical distributions of certain specific classes of interatomic distances. For PTI, an r.m.s. deviation of approximately $\Delta d = 5.5$Å is obtained. Certain biases in the starting coordinate sets complicate a reasonable assessment of the quality of this approach in the other examples cited, lysozyme and staph nuclease. Moreover, the coefficients for the penalty functions were derived from the crystallographic coordi-

nates of the protein to be modeled. Although some novel ideas are presented, it is difficult to assess the meaning of this modified distance geometry algorithm.

Distance Geometry calculations offer a unique method for including theoretical and experimental information. This technique may provide a powerful method for adjusting the results of certain theoretical calculations to be consistent with specific experimental evidence.

6.2.5 Statistical Approach

If the proteins whose crystallographic coordinates are known form a representative group of all possible protein structures, then the statistical preferences of amino acids for specific conformations may be useful in the prediction of protein structure. The most successful applications of this approach have been in the area of secondary structure prediction, the location along the chain of α-helices, β-strands, and perhaps β-turns.

There are two major statistical approaches to secondary structure prediction: singlet and multiplet propensities. Some approaches to the secondary structure prediction problem are based on stereochemical rather than statistical considerations. However, these methods can often be recast in statistical terms with multiplet information. These include Kuntz (1972) for turns, Schiffer & Edmundson (1967, 1968) and Palau & Puigdomenech (1974) for helices, and Lim (1974a,b) for α-helices and β-structure. They allow only 1 and 0 probabilities. Singlet propensities reflection the probability of an α-helical (P_α), β-strand (P_β), β-turn (P_T) conformation for a particular amino acid while multiplet propensities estimate these same probabilities for an amino acid in position i with other amino acids j_1, j_2, ... j_m at a certain spacing from i.

6.2.5.1 Stereochemical Analysis

Lim (1974a) developed a secondary structure prediction algorithm based upon the stereochemical properties of the amino acids. He conjectured that secondary structure is determined in conjunction with tertiary structure and so only sequences which are compatible with the close packing of preformed units of secondary structure promote the formation of α or β structure. Thus, he concluded that the creation of a hydrophobic globule shielded by hydrophilic residues determines protein structure.

A variety of rules about sequences which promote and inhibit secondary structure formation were devised from model building studies. An algorithm which synthesized these structural rules was applied to a variety of sequences of proteins with known tertiary structure. As these rules are quite convoluted, they cannot be enumerated briefly. The results of this algorithm were comparable to those obtained by statistical techniques. This success suggests that there are very special local sequence-structure relationships which follow from a consideration of the steric and hydrophobic requirements of individual amino acids. Although there have been no successful attempts to improve Lim's prediction algorithm, a secondary structure prediction scheme rooted in stereochemical considerations seems to have the greatest chance for success.

6.2.5.2 Singlet Propensities

Many investigators have studied the correlation between amino acid type and backbone dihedral angles. These studies have led to a variety of secondary structure prediction algorithms (Chou & Fasman, 1974; Lewis et al., 1971; Beghin & Dirkx, 1975; Dirkx, 1972; Ptitsyn & Finkelstein, 1970a,b; Finkelstein & Ptitsyn, 1971). The simplest and most popular of these methods is due to Chou & Fasman (1974). This method follows a four step procedure:

(1) Rank the preference of each residue for each of the three

major conformational substates: α-helix, β-strand or coil. For the helical case, these are typically H for a residue which is frequently helical $P_\alpha > 1.05$; I for a residue with no preference $0.9 < P_\alpha < 1.05$; and B for a helix breaker $P_\alpha < 0.9$. An analogous approach is used for β-structure and coiled segments.

(2) Search for nucleation of secondary structure. A helix is formed when four H_α's occur in a hexapeptide and a strand is formed when three H_β's occur in a pentapeptide. A turn is formed when the product of the P_T for each residue in a tetrapeptide is greater than 5×10^{-5}.

(3) Helix and strand nucleation units are elongated until a cluster of helix, B_α, or strand, B_β, breakers are encountered.

(4) When a region is predicted to have two distinct conformations, the region with the highest average propensity is favoured.

Although this method seems straightforward, ambiguities exist in the assignment of residues at the end of a helix or strand. This has hampered the development of a computer algorithm to apply the Chou & Fasman method.

6.2.5.3 Multiplet propensities

Although the data base of known protein structures is large relative to a singlet contingency table, the size is comparable to the number of entries in a doublet contingency table. Thus, the statistical quality of all but restricted doublet studies must be in doubt. Periti (1974) has developed a prediction algorithm based solely on doublet frequencies. Robson and co-workers (e.g. Robson & Pain, 1974; Robson & Suzuki, 1977) have partially resolved the statistical problem by combining singlet and limited doublet propensities in an information theory approach with adjustable α, β, and turn potentials. In another variation Nagano (1973, 1974, 1975, 1977) and Wu & Kabat (1971, 1973, see also Kabat & Wu 1973a,b, 1974) have used homologous sequences to augment their data

base.

A variety of procedures have been developed to predict the location of secondary structure along a polypeptide chain. Although the accuracy is typically greater than 50% and perhaps as high as 80%, there is a great need for improvement, especially in the prediction of β-structure. The conclusion is supported by the success of various prediction schemes on adenyl kinase when the sequence was available but the structure was unknown (Schulz et al., 1974). No method successfully predicted the location of all α-helices and β-strand but the results were better than expected by chance alone. The predictions for helical segments were much better than for regions of β-structure. Thus structures involving interactions between local pieces of chain are more successfully predicted than those interactions between distant parts of the chain. It appears that additional statistical information will not improve the accuracy of these algorithms so other types of structural information must be used. It is clear, however, that a complete knowledge of secondary structure from the amino acid sequence will explain very few features of tertiary structure (Havel et al., 1979) unless a procedure for packing preformed secondary structure is available (e.g. Cohen et al., 1979).

6.2.6 Combinatorial Approach

Instead of relying on an empirical potential function, several researchers (e.g. Sternberg & Thorton, 1977; Richardson, 1977; Chothia et al., 1977; Levitt & Chothia, 1976) have studied the geometrical patterns frequently observed in a variety of proteins. A geometric approach to the allowed packing motifs for secondary structures suggests a variety of alternative tertiary structures distributed through the conformation space of a polypeptide chain. A path search (e.g. Ptitsyn & Rashin, 1975) or a global search (e.g. Cohen et al., 1979) of conformation space is possible and empirical rules to sort among the alternatives can

be developed.

Ptitsyn & Rashin (1975) developed a procedure for deducing the structure of myoglobin from a knowledge of the approximate secondary structure. Myoglobin, a predominantly helical protein, was modeled as a 'sausage and string' and hydrophobic interaction sites along the helices of varying strengths were located. Conformation space was then searched manually. Helices were packed together to cover these hydrophobic sites and only the most energetically favourable intermediates on the basis of a crude hydrophobic potential were allowed to continue. Steric and connectivity restrictions were consequences of the real three-dimensional blocks used for helices. One of the two best structures bore a physical resemblance to the globin fold, but no coordinates were available. This procedure also suggested a path for the folding of myoglobin. Unfortunately, this procedure was not automated.

A computer algorithm similar in essence to the Ptitsyn & Rashin approach was developed by Richards and co-workers. Richmond & Richards (1978) developed an algorithm for locating strong helix-helix interaction sites along the myoglobin sequence. Cohen et al. (1979) used these sites to construct 10^8 candidates for the structure of myoglobin. Of these, only twenty were consistent with certain steric and connectivity constraints. Cohen & Sternberg (1980a) showed that only two of these twenty can accommodate the heme group so that the bound iron atom is between His 64 and His 93. Moreover, one of these structures closely resembles myoglobin. The r.m.s. deviation is $\Delta d = 3.6\text{Å}$ from the crystal structure when a value of $\Delta d = 8.0\text{Å}$ would be expected for a random compact structure. This work will be discussed in detail in Chapter 3.

Ptitsyn and co-workers have continued the path approach in studies of immunoglobulins (Zav'yalov, 1977) and other all-β proteins (Ptitsyn et al., 1979). Only Greek key topologies are accepted and those topologies which force crossover connection are rejected.

The most probable folding paths are examined. Although this approach remains conceptually intriguing, the precise relationship between the predicted topologies and a list of coordinates remains unclear.

Cohen et al. (1980a and b) have extended their work to all-β proteins (see Chapter 4) as well as βαβ proteins (see Chapter 5). These algorithms have been applied to 20 different proteins with reasonable success. The best approximation to the crystal structure is typically one of one hundred alternative structures and has an r.m.s. error of approximately $\Delta d = 2.5\text{Å}$ from the X-ray coordinates.

These methods have succeeded in producing a good prediction of the crystallographically determined protein coordinates. Of course, this approach hinges on a reliable procedure for identifying secondary structure and is only useful for proteins comprised largely of α-helices and/or β-strands. Ptitsyn and co-workers have failed to automate their investigations and quantitatively assess the relationship between the predicted and crystal structures. The global search avoids the multiple minima problem but produces more than one possible structure. It is the inability to select between the remaining alternatives that is the major shortcoming of this work. Perhaps energy minimisation procedures will be able to sort rapidly between these folded alternatives.

7. Scope of the Thesis

Since the number of parameters in an all atom representation of a protein structure is large, most attempts at structure prediction start with a simplified or idealised representation of the actual polypeptide chain (e.g. Levitt & Warshel, 1975; Robson & Osguthorpe, 1979). Pseudo-potential functions compatible with these idealised representations were developed and an energy minimsation procedure was then applied to cause the extended chain to collapse to a compact structure. Although the final

structure is the main interest, this general approach may result in geometrically difficult or impossible intermediate stages causing the procedure to stop short of the final goal. False minima are an intrinsic difficulty of all direct energy minimisation approaches even though these methods have a sound theoretical basis. The problem appears to stem from the fact that the smooth energy-conformation surface on which the minimisation is carried out is a macroscopic concept applicable to an ensemble of molecules while the actual calculation is carried out by specific geometrical adjustment of a single molecule with only a limited number of attempts at alternate paths on the surface. Thermal energy is inserted only in an ad hoc fashion in these procedures.

In contrast to this fundamental method, this thesis explores an alternate approach for the early stages of protein folding. Energy is not considered in an explicit potential function. Instead, probable polypeptide chain geometries are derived from studies of known structures. Structural elements are simply placed in space; sequential changes by small increments are not required. Such procedures will usually have a computational "path" but this need bear no relation to a physically reasonable pathway. The basic rules depend heavily on the analyses of known protein structures that have been carried out in many laboratories over the past few years (see reviews by Richardson, 1980, and by Richards, 1977).

Chapter 2 is devoted to understanding the root mean square deviation of the atomic positions in a predicted structure from the native crystallographic coordinates as this is a common measure of success in protein folding studies. The deviation expected for a random prediction which only requires that the chain be compact is evaluated for 12 proteins with a wide range of molecular weights. An estimate of the number of compact conformations for a protein of a specific chain length is proposed.

The rest of this thesis concentrates on understanding the inter-

actions of secondary structure units and how these interactions can be useful in a combinatorial approach to the protein folding problem. Chapter 3 is devoted to helix-helix interactions, Chapter 4, sheet-sheet interactions and Chapter 5, the interactions of helices and sheets. The beginning of each of these Chapters is devoted to an analysis of the geometrical and topological properties of the various structural classes. Then, the structures are dissected and the nature of the changes in non-polar accessible contact area on unpacking is monitored to reveal how the burial of hydrophobic residues is facilitated by the packing of secondary structure units. Constellations of hydrophobic residues are found which mediate the interaction of pieces of secondary structure and rules to locate them from a knowledge of sequence and secondary structure are developed. Then all possible combinations of secondary structure which are suggested by the interaction geometry and hydrophobic packing requirements are constructed. Those that are consistent with topological and steric constraints are filtered from the list of all combinations and compared to the crystallographic structure. Typically, the number of structures which survive these sieves is small yet a reasonable approximation to the native is always included. Conclusions about the ultimate utility of a combinatorial approach and future work are discussed in Chapter 6.

CHAPTER II

THE SIGNIFICANCE OF THE ROOT MEAN SQUARE DEVIATION

1. Introduction

Many investigators, through a variety of techniques, have simulated the folding of extended polypeptide chains and produced compact globular forms. As a measure of their success, they have usually reported the r.m.s.* deviation of the atomic positions in their model from those in the crystallographically observed native structure. This deviation offers some measure of the relative effectiveness of two procedures for folding the same protein. However, little effort has been devoted to establish a standard to determine the absolute success of a procedure or to relate quantitatively the success of various methods on different proteins.

Havel *et al.* (1979) in their studies of PTI and CPV have obtained r.m.s. deviations for randomized structures using a distance geometry approach and have discussed in detail the effects of various constraints on folding simulations. Hagler & Honig (1978) assessed the quality of energy minimisation studies of PTI by folding a chain composed solely of alanine and glycine residues.

In this chapter, a general standard for assessing the quality of folding studies is developed. A series of random compact structures was generated and compared to the crystallographic coordinates for each of twelve proteins. The r.m.s. deviation was found to be directly related

* In this Chapter, the following abbreviations will be used: root mean square (r.m.s.); pancreatic trypsin inhibitor (PTI); b_5-cytochrome (B5C); Bence-Jones protein (REI); Ribonuclease S (RNS); Flavodoxin (FXN); Staphlococyl nuclease (SNS); Super oxide dismutase (SOD); Myoglobin (MBN); Adenylate kinase (ADK); Concanavalin A (CNA); Triose phosphate isomerase (TIM); Carp parvalbumin (CPV); Thermolysin (TLN); T4 phage lysozyme (T4L).

to the length of the protein chain and a model to explain this correlation is presented. This information is then used to assess the quality of various predictive studies. A similar random walk procedure has been developed by Schulz (1980) to assess the evolutionary significance of structural comparisons.

2. The Relationship between the r.m.s. Deviation computed by the Rotation and Interatomic Distance Methods

There are two procedures commonly used to compute the r.m.s. deviation: the rotation and interatomic distances methods. The first is used in investigating questions of evolutionary similarity, and the second quantity is normally evaluated in folding studies. The relationship between the r.m.s. deviation calculated using the rotation and interatomic distance method is considered first.

The r.m.s. deviation based on the rotation method is computed as:

$$\overline{\Delta r} = \sqrt{\sum_{i=1}^{m} (x_i - y_i')^2 / n} \qquad (2.1)$$

with $\{x_i\}$ the coordinates of the crystal structure and $\{y_i\}$ the generated structure where the primed frame is the rotation which minimised $\overline{\Delta r}$. The r.m.s. deviation based on the interactomic distances method is computed as:

$$\overline{\Delta d} = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{m} (d_{ij} - e_{ij})^2 / n^2} \qquad (2.2)$$

$$d_{ij} = [(x_i - x_j)^2 + (y_i - y_j)^2 + (z_i - z_j)^2]^{\frac{1}{2}}$$

where $\{d_{ij}\}$ are the interatomic distances between atoms in the crystal structure and $\{e_{ij}\}$ are the interatomic distances between atoms in the generated structure (Levitt, 1976; Nishikawa et al., 1972).

Levitt (1976) suggested that the two r.m.s. deviations could be

related by the relationship:

$$\overline{\Delta d} = \sqrt{3/2}\ \overline{\Delta r} \qquad (2.3)$$

This cannot always be true because $\overline{\Delta d}$ is invariant under reflection and $\overline{\Delta r}$ is not. What relationship holds if a reflection as well as rotation is allowed in computing $\overline{\Delta r}$ was then investigated.

A series of structures with a continuous range of r.m.s. deviations was constructed by perturbing the PTI structure and regularising the resulting coordinate set. Each atom was moved by a fixed increment and the coordinates were adjusted so that each virtual bond between consecutive alpha carbons was 3.81Å*. A graph of the values obtained for the rotation method versus the interatomic distances method is shown in Figure 2.1a. The least squares line is:

$$\overline{\Delta d} = 0.75\overline{\Delta r} + 0.19 \qquad (2.4)$$

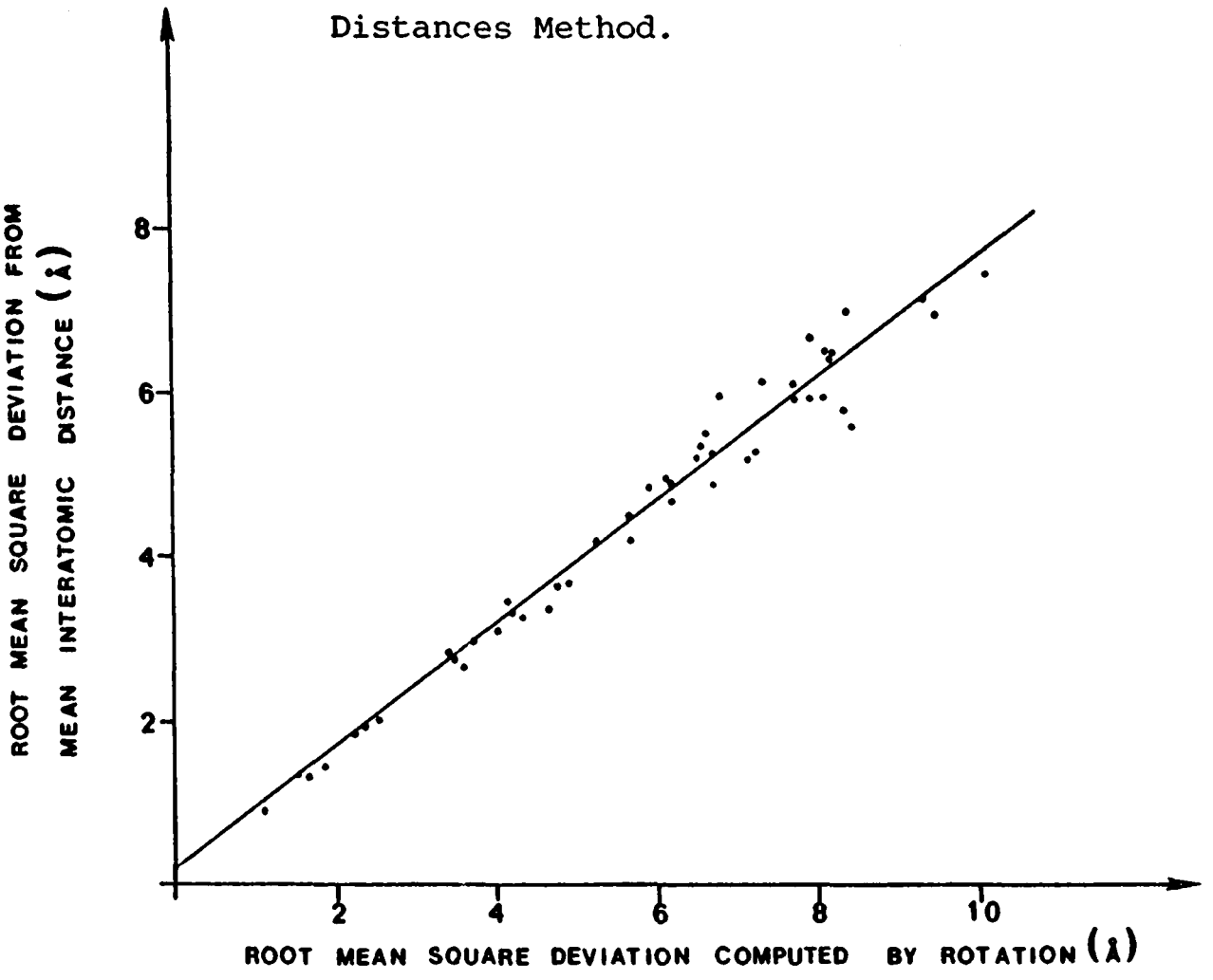
with a product moment correlation coefficient of 0.99. The fit to the lower half of the curve is better than to the upper half where a slight spreading occurs. A slope of 0.82 would have been expected from Levitt's relationship. If this computation is repeated, but before perturbing the alpha carbons, the x-coordinate is replaced by -x, the graph changes markedly. This reflection shifted the y-intercept and dropped the correlation coefficient to 0.69 (see Figure 2.1b). The scatter of points in the region with interatomic distance r.m.s. deviations (Δd) between 6 and 8Å corresponds with the scatter observed in this region in Figure 2.1a. Perhaps structures in this region are truly random as they have lost their original topological handedness.

Clearly, the r.m.s. deviation computed by rotation is a better method of assessing structural relatedness as structures with the wrong

* A copy of the Fortran program INCREMENT used in this calculation is in Appendix 1.

FIGURE 2.1a

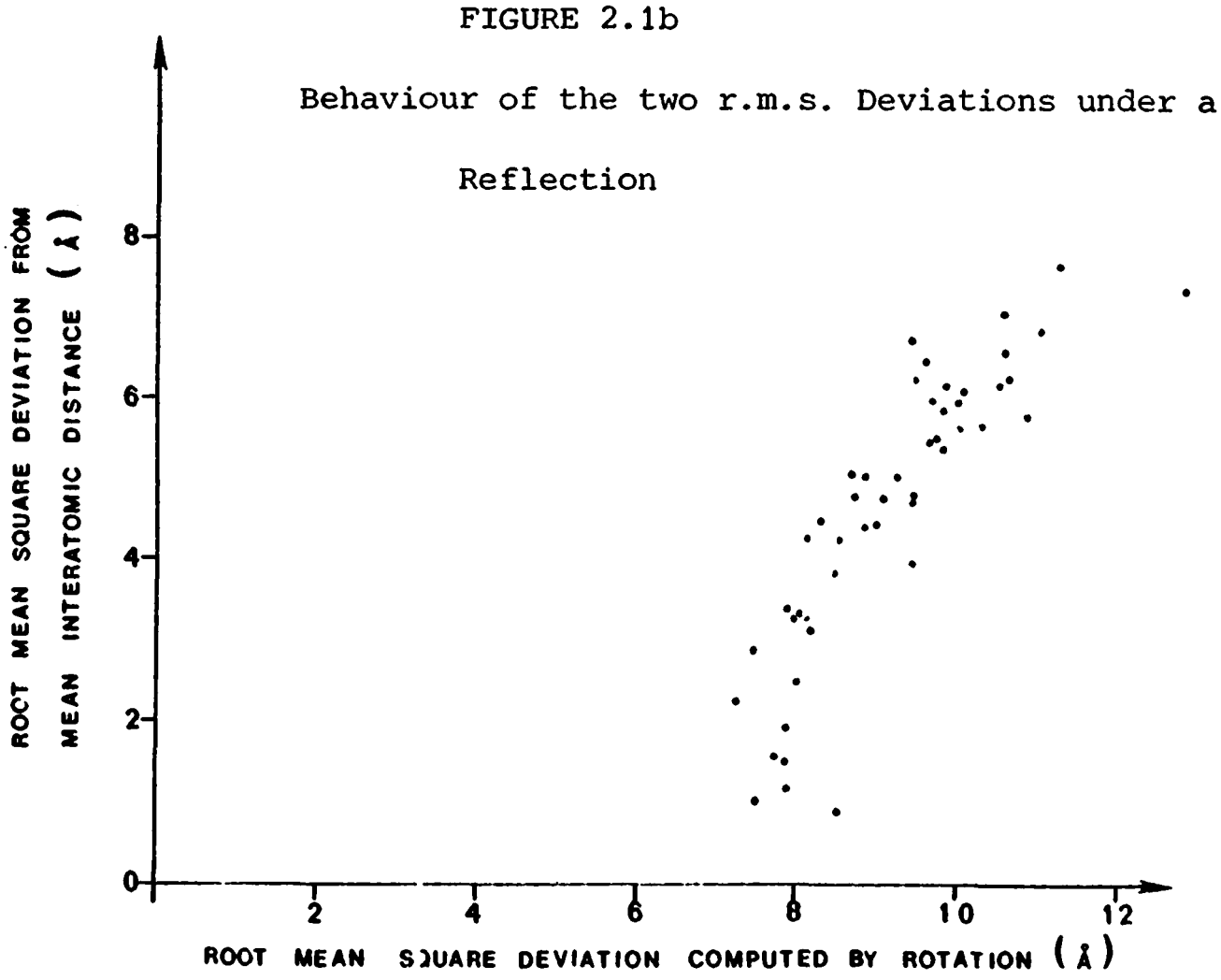
The Linear Relationship between the Rotation and Interatomic Distances Method.



The linear least squares equation through these points is $\overline{\Delta d} = 0.75\overline{\Delta r} + 0.19$, with a product moment correlation of 0.99.

FIGURE 2.1b

Behaviour of the two r.m.s. Deviations under a Reflection



The same ordinate and abscissa as in Fig. 2.1a are used to show that the interatomic distances method is invariant under reflection ($x \rightarrow -x$), but the rotation method is not.

hand will not produce a low r.m.s. deviation. Since many protein folding studies have computed r.m.s. deviations using interatomic distances, both methods will be considered in the work that follows. However, it is suggested that in any future folding studies, only the r.m.s. deviation by rotation ($\overline{\Delta r}$) should be quoted as this will prevent confusion.

3. Self Avoiding Random Walks

3.1 Algorithm

Protein folding studies often report the r.m.s. deviation of the "best-predicted" structure from the crystal structure as a measure of success. All current folding procedures generate compact structures as a result of a van der Waals attractive term and/or a hydrophobic term in their potential or pseudo-potential functions. However, no measure of the r.m.s. deviation of a randomly generated compact structure from the crystallographic coordinates exists. This hinders an interpretation of the success of the folding study. To obtain this standard, a procedure to generate random compact polypeptide chains was required.

Monte Carlo simulations have been useful in understanding the physical chemical solution properties of many homopolymers (see Flory, 1969). More recently, these computations have been applied to answer questions about the nature of the protein-solvent interface and the properties of small polypeptides (Hagler & Moult, 1978). However, as the size of the system increases, the computation time increases dramatically. Increasing the number of atoms considered produces an exponential increase in the number of steric clashes. This is the excluded volume problem.

For long polypeptide chains constrained to be compact, a method of improving the efficiency of the generating algorithm is needed. Either the set of allowed conformations must be limited or the standard procedure

of rejecting the entire structure when one point fails the imposed restrictions must be replaced by a back-trace procedure. Levitt (1977) chose the first option in his studies on PTI and restricted the allowed virtual bond angles between consecutive alpha carbons to 90° or 120°. Speed is achieved as all possible alpha carbon positions are the vertices of a cubic lattice and space filling criteria are violated when the same lattice point is used twice. Unfortunately, the best fit of a crystal structure to this lattice results in an r.m.s. deviation ($\overline{\Delta r}$) of 2.6Å.

The second alternative was chosen and random structures were generated with a limiting sphere. A back-trace procedure is used to resolve steric clashes. The algorithm is divided into the four calculations described below:

3.1.1 Compute the radius (r) of the limiting sphere.

$$r = \beta[(3)(110)n\bar{v}/4\pi 6.022]^{1/3}\text{Å} \qquad (2.5)$$

$\bar{v}$ = 0.73 g/ml

n = number of residues

In this equation, 110 is taken to be the molecular weight of an average amino acid residue and $\bar{v}$ is the partial specific volume. β is a constant varied in this study between 1.0 and 1.5 to allow particles generated in this study to have axial ratios between 1.0 and 1.84.

3.1.2 Place the first alpha carbon.

This first point is chosen randomly from all points in the sphere so that the radial distribution of starting points is uniform. Since the r.m.s. deviation computed from either the rotation method or the interatomic distance method is invariant under rotation, a uniform radial distribution was chosen.

3.1.3 Place the remaining alpha carbon.

Additional points are chosen randomly so that the $i+1^{th}$ atom lies

3.81Å from the i^{th} atom and so that the bond angle between three consecutive alpha carbons is between 90° and 140°. Space filling information is incorporated by ensuring that the distance between non-consecutive alpha carbons was greater than 4.5Å. These values represent the minimum and maximum observed quantities in PTI (Deisenhofer & Steigemann, 1975). Each point was also required to lie within the constraining sphere.

3.1.4 Back-trace procedure.

Since the randomly chosen points were only required to lie 3.81Å from the previous point with a reasonable bond angle, the space filling and limiting sphere constraints were frequently violated. When either of these constraints was not satisfied, 9 additional attempts to find a suitable point were made. If these 10 tries proved unsuccessful, another attempt to place the previous point was made. This back-trace procedure was applied iteratively.

3.2 The validity of the algorithm*

First the fairness of this algorithm must be justified by considering the validity of the back-trace procedure and assessing the effect of varying the size of the constraining sphere (i.e. the constant β). In an "unbiased" random walk, a back-trace procedure should not be used. Instead, the entire structure should be rejected when a space-filling or the limiting-sphere constraint is violated. Therefore an "unbiased" random walk calculation on PTI with β = 1.5 was performed. The r.m.s. deviation for 50 structures was 13.55 ± 1.58Å. This compares reasonably well with the "biased" value 12.57 ± 1.54Å for 400 structures, a 7% decrease. The advantage of the "biased" procedure is that the computer time decreased from 50seconds/structure to 1 second/structure on an ICL 2980 computer. This increase in computational speed will be crucial when larger structures are considered.

* The Fortran coding for this algorithm, WALK, is in Appendix 1.

The other feature of the algorithm that was introduced to reduce the time for the calculation is the value of β which dictates the size of the constraining sphere. For PTI, decreasing β from 1.5 to 1.0 resulted in a ten-fold increase in the computation time. We have therefore developed a mathematical model to relate the r.m.s. deviation of random structures generated with $\beta = 1.5$ to the deviation of structures of $\beta = 1.0$.

If the bond-length, bond-angle and space-filling constraints are neglected, then the deviation ($\overline{\Delta r}$) should be the average distance between points in two superimposed spheres. Thus:

$$\left(\frac{\overline{\Delta r}}{\rho}\right)^2 = \frac{\int_0^{\beta}\int_0^{\pi}\int_0^{2\pi}\int_0^{1}\int_0^{\pi}\int_0^{2\pi} \{r_1^2 + r_2^2 - 2r_1r_2\{\sin\phi_1\sin\phi_2\cos(\theta_1-\theta_2) + \cos\phi_1\cos\phi_2\} \times r_1\sin\theta_1 r_2\sin\theta_2\}dr_1\, d\theta_1\, d\phi_1\, dr_2\, d\theta_2\, d\phi_2}{\frac{4}{3}\pi\frac{4}{3}\pi\beta^3} \qquad (2.6)$$

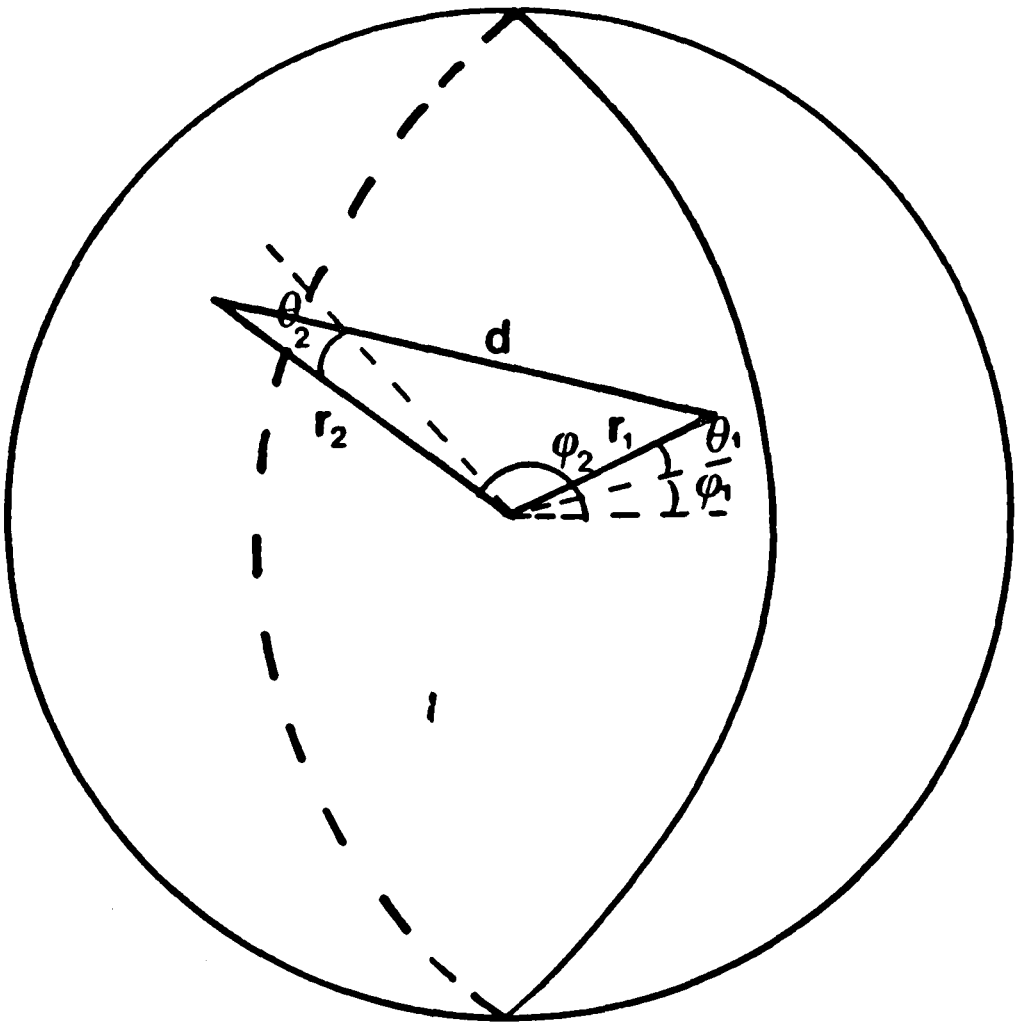
The numerator is the integral over the entire sphere of the distance between two points in spherical polar coordinates for each point. The denominator is the volume squared or the total number of points counted. β relates the radius of the constraining sphere to that of the crystal structure and ρ is the radius of the sphere (see Figure 2.2). Upon integration, equation 2.6 reduces to:

$$\overline{\Delta r}/\rho = \{0.3(1 + 1/\beta^2)\}^{\frac{1}{2}} \qquad (2.7)$$

A list of values for the r.m.s. deviation using the rotation method for fifty randomly generated structures compared to PTI for each of six values of β is presented in Table 2.1. Table 2.2 contains the theoretical and actual values of $\overline{\Delta r}/\rho$. Although the trend is correct, the agreement is not satisfactory. In this first model, no attempt was made to include the space-filling constraint. Thus, interatomic distances less than 4.5Å were included in the integral but are not consistent with a space-filling model.

FIGURE 2.2

Geometric Construction for the Integrand of Equation 2.6.



The distance between two points in spherical polar coordinates is:

$$d^2 = r_1^2 + r^2 + 2r_1r_2\{\sin\phi_1\sin\phi_2\cos(\theta_1 - \theta_2) + \cos\phi_1\cos\phi_2\}$$

TABLE 2.1

PTI RANDOM WALKS WITH CONSTRAINING SPHERES OF VARIOUS SIZES

β	Radius (Å)	r.m.s. deviation-rotation method	
		Mean (Å)	Standard deviation (Å)
1.0	12.25	10.86	1.42
1.1	13.48	10.91	1.19
1.2	14.70	11.70	1.33
1.3	15.93	12.01	1.44
1.4	17.15	12.13	1.24
1.5	18.38	12.58	1.33

TABLE 2.2

THE RATIO OF THE R.M.S. DEVIATION TO THE

SIZE OF THE CONSTRAINING SPHERE FOR PTI

β	Radius (ρ)	Mean r.m.s. Deviation ($\overline{\Delta r}$) Rotation Method (Å)	$\overline{\Delta r}/\rho$		$\overline{\Delta r}/\rho$ + 2.25	
			Theoretical	Empirical	Theoretical	Empirical
1.0	12.25	10.86	.77	.89	.72	.75
1.1	13.48	10.91	.74	.81	.69	.69
1.2	14.70	11.70	.71	.80	.68	.69
1.3	15.92	12.01	.69	.75	.66	.66
1.4	17.15	12.13	.67	.71	.65	.63
1.5	18.38	12.58	.66	.68	.64	.61

Space-filling information can be incorporated in an approximate fashion by increasing the effective size of the constraining sphere by the radius of an alpha carbon. Since the minimum allowed C_α-C_α distance was 4.5Å this radius was taken as 2.25Å. This has the desired effect of balancing out the short distances which were counted in the integral, but not allowed in the sum by the space-filling constraint. The values of the ratio $\overline{\Delta r}/\rho+$ where $\rho+ = \rho + 2.25$ can be found in Table 2.2. Thus the values found with $\beta = 1.5$ in all subsequent studies to relate to values with $\beta = 1.0$ by a factor of $(0.64/0.72) = 0.89$. All future r.m.s. deviations reported in this study are adjusted to the value for $\beta = 1.0$.

3.3 Random walks on twelve proteins

Fifty random structures with lengths corresponding to the proteins listed in Table 2.3 were generated for each protein. The time required to generate a structure and compute its r.m.s. deviation from the crystallographic coordinates varied between 1.0 second for PTI and 12.0 seconds for TIM. Table 2.3 also presents the mean and standard deviation for the r.m.s. deviation computed by the rotation and interatomic distances method. This information was plotted against the number of residues (see Figure 2.3) and the equations of the two least squares lines were found to be:

$$\overline{\Delta r} = 0.0468(\text{number of residues}) + 9.25 \qquad (2.8)$$

$$\overline{\Delta d} = 0.0271(\text{number of residues}) + 5.56 \qquad (2.9)$$

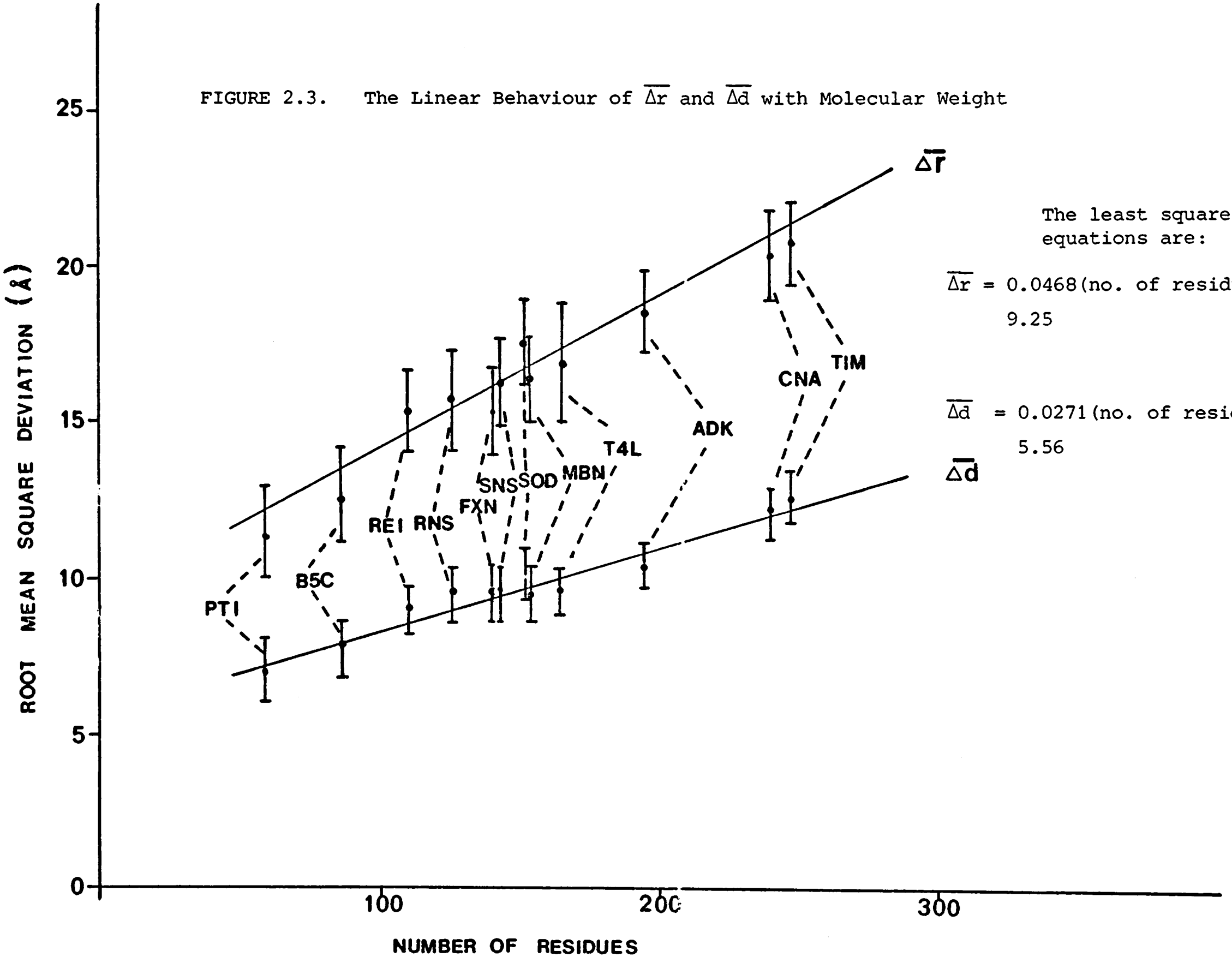
with correlation coefficients of 0.95 and 0.97 respectively. The value of $\overline{\Delta r}/\rho+$ for these twelve proteins is 0.67 ± 0.04 which agrees reasonably well with 0.64, the value expected theoretically. It is troubling that the y-intercepts of equations (2.8) and (2.9) are far from the origin as the value of $\overline{\Delta r}$ or $\overline{\Delta d}$ should be zero with only two residues. Apparently, the approximations used preclude the linear extrapolation of these data

TABLE 2.3

STATISTICS ABOUT THE R.M.S. DEVIATION FOR FIFTY RANDOM ANALOGUES OF TWELVE PROTEINS FROM THE CORRESPONDING CRYSTAL STRUCTURES

Protein		Number of Residues	r.m.s. deviation rotation method ($\overline{\Delta r}$)		r.m.s. deviation interatomic distances ($\overline{\Delta d}$)	
			Mean (Å)	Standard deviation (Å)	Mean (Å)	Standard deviation (Å)
Pancreatic Trypsin Inhibitor	(PTI)	58	11.20	1.42	6.90	.82
B_5 Cytochrome	(B5C)	84	12.40	1.58	7.53	.84
Bence-Jones Variable Portion	(BEI)	107	15.17	1.39	8.97	.78
Ribonuclease-S	(RNS)	124	15.74	1.77	9.47	.75
Flavodoxin	(FXN)	138	15.17	1.52	9.32	.88
Staphylococcal Nuclease	(SNS)	142	16.16	1.65	9.43	.92
Super Oxide Dismutase	(SOD)	151	17.48	1.40	10.10	.99
Myoglobin	(MBN)	153	16.21	1.42	9.48	.93
T4 Lysozyme	(T4L)	164	16.35	2.04	9.46	.84
Adenylate Kinase	(ADK)	194	18.44	1.66	10.27	.71
Concanavalin A	(CNA)	237	20.26	1.55	12.08	.87
Triose Phosphate Isomerase	(TIM)	247	20.59	1.61	12.43	1.00

FIGURE 2.3. The Linear Behaviour of $\overline{\Delta r}$ and $\overline{\Delta d}$ with Molecular Weight



The least squares equations are:

$\overline{\Delta r}$ = 0.0468(no. of residues) + 9.25

$\overline{\Delta d}$ = 0.0271(no. of residues) + 5.56

to polypeptides smaller than PTI.

3.4 Self avoiding random walks with preset helices

In addition to attempting to fold a protein from an open chain conformation, several researchers have commenced their folding simulations with the secondary structure preset. This includes the work of Levitt (1976) on PTI, Warshel & Levitt (1976) on CPV, Kuntz et al. (1976) on PTI, Robson & Osguthorpe (1979) on PTI and Cohen et al. (1979) on MBN.

The procedure used to generate random structures was extended to allow α-helices to be inserted at the appropriate points in the sequential construction of random chains. When the N-terminal alpha carbon of an α-helix was reached, a random vector was selected as the helical axis and a second random vector was chosen perpendicular to the axis to phase the helix. Alpha carbon positions were selected to fit the standard parameters of an α-helix: 3.6 residues/turn and a pitch of 1.5Å/residue, with a distance of 2.29Å between the alpha carbon and the axis. The same space-filling, bond-length and bond-angle constraints were imposed*.

Random walks with preset helices were performed on CPV, MBN and the helical domain of TLN. The crystallographic assignments of α-helices were used. Results obtained in these studies (see Table 2.4) were strikingly similar to those for walks with no secondary structure specified. The mean r.m.s. values for the random structures with α-helices average 0.8Å less than the value expected from equation (2.8). For the interatomic distance r.m.s. deviation, the corresponding value is 0.3Å. This distance, though consistently negative, in only one half of a standard deviation from the mean for a random structure. Thus, the knowledge of only the location of helical structure along the polypeptide chain does not significantly increase the chance of success of a theoretical protein folding study. This is

* The Fortran coding for WALKER is in Appendix 1.

TABLE 2.4

RANDOM WALKS WITH PRESET HELICES

Protein	Number of Residues	r.m.s. deviation rotation method ($\overline{\Delta r}$)		r.m.s. deviation interatomic distances ($\overline{\Delta d}$)	
		Mean (Å)	Standard deviation (Å)	Mean (Å)	Standard deviation (Å)
Carp Parvalbumin (CPV)	108	13.41	1.97	7.98	.86
Myoglobin (MBN)	153	16.22	1.69	9.58	.93
Thermolysin α domain (TLN)	186	16.78	1.48	10.48	.93

not surprising as a complete knowledge of α-helical structure in myoglobin specifies only 5% of the matrix of interatomic distances. A different approach is required to examine the effect of the knowledge of the position of β-structure or disulphide bridges on reducing the r.m.s. deviation as these arrangements require an algorithm which ensures the spatial proximity of sequentially-distant parts of the polypeptide chain. No attempt has been made to include topological features of tertiary structure (for example see Sternberg & Thornton (1977), Richardson (1977)) as a goal of our work is to provide an r.m.s. deviation for a random structure. The success of a protein-folding study to incorporate these topological properties into a predictive scheme will then be evident.

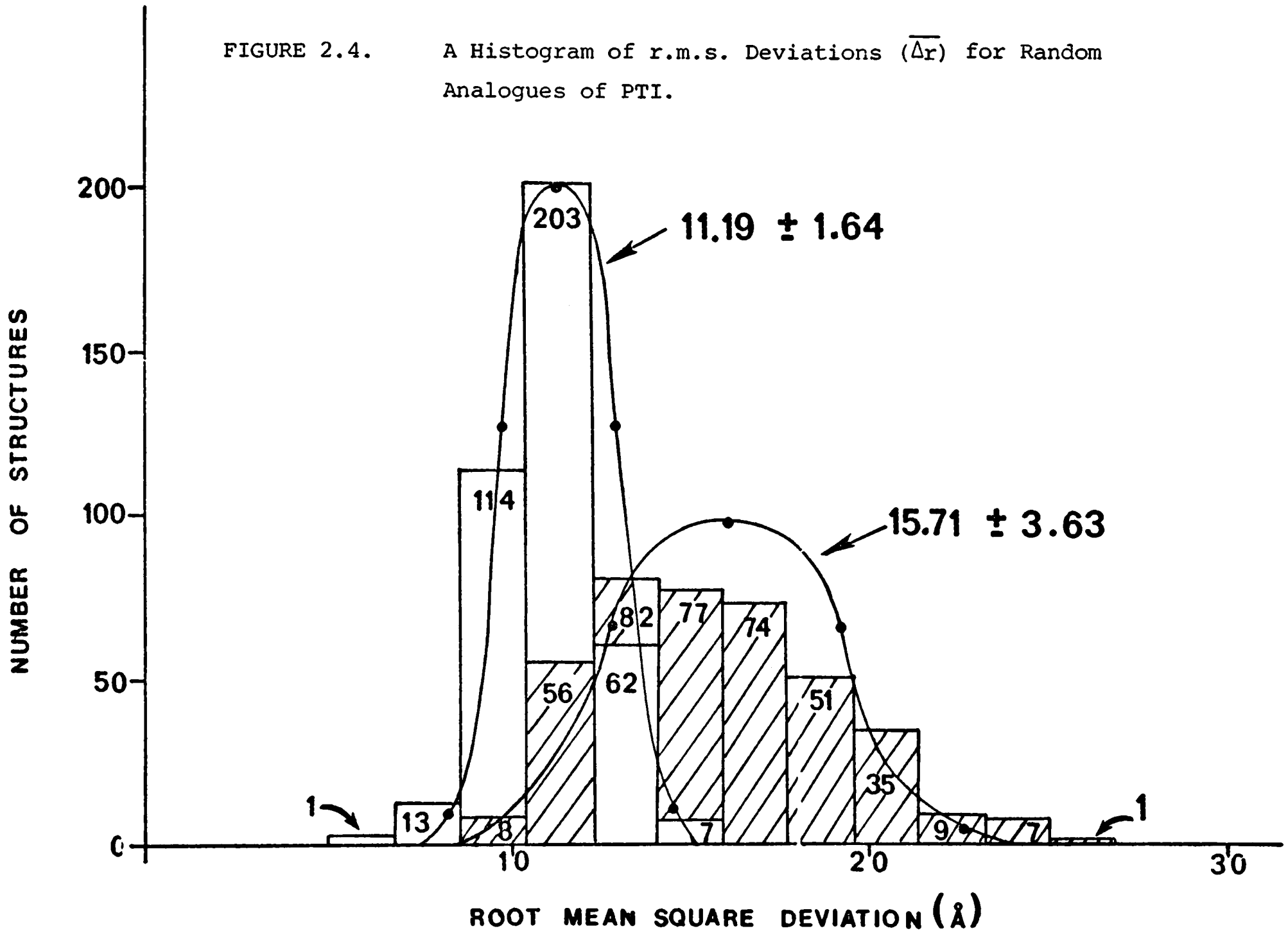
4. Deducing the Total Number of Compact Globular Structures for a Given Polypeptide Chain Length

A standard estimate of the total number of conformations available to a polypeptide chain taken from Levinthal (1966) is:

$$\text{number of conformations} = 10^{\text{number of residues}} \qquad (2.10)$$

The results of the random walks enable an estimate of the total number of compact globular golds for C_α positions. If the distribution of r.m.s. deviations for a random structure compared to a crystal structure were Gaussian, then all information about the distribution is contained in the mean and standard deviation. A histogram of r.m.s. deviations for 400 analogues of PTI and the corresponding normal curve is shown in Figure 2.4. With the compactness constraint, the agreement of the two distributions is good. When this constraint is removed, the distribution shifts to the right as expected. However, since this new distribution is skewed, the number of structures cannot be evaluated by the method suggested. Since only one structure has an r.m.s. deviation of 0.0Å, one can normalise

FIGURE 2.4. A Histogram of r.m.s. Deviations ($\overline{\Delta r}$) for Random Analogues of PTI.



800 random analogues of PTI were generated, half with the limiting sphere compactness constraint (in white) and half without this constraint (shaded). These were compared to the native structure and a histogram of the results was constructed. Also illustrated are the normal curves corresponding to the mean and standard deviation for the two sets.

the distribution by ensuring that:

$$\frac{N}{\sqrt{2\pi}} \int_{-\infty}^{-\mu/\sigma} e^{-t^2/2}\, dt = 1 \qquad (2.11)$$

where N is the total number of points in the sample space, μ is the mean and σ is the standard deviation for a set of random compact structures, and the integral is the standard error function. Thus, the total number of compact conformations should be:*

$$N = \sqrt{2\pi} \Big/ \int_{-\infty}^{-\mu/\sigma} e^{-t^2/2}\, dt \qquad (2.12)$$

From this formula, the empirical analogue of equation (2.10) for the number of compact structures was found:

$$\text{number of compact conformations} = 10^{0.125(\text{number of residues}) + 5} \qquad (2.13)$$

(see Table 2.5). If a molecule can sample one conformation in 10^{-13} sec, the period of a molecular vibration, then a polypeptide with 64 residues could sample all of the compact conformations in 1 second. Perhaps Levinthal's argument for the existence of specific kinetic pathways requires review in the light of equation (2.13).

5. Evaluation of Folding Simulations

Equation (2.12) suggests a method for assessing the quality of various predictive schemes on the conformations of proteins. In these studies, $\overline{\Delta d}$ has been computed. The average standard deviation for this quantity for the twelve proteins in this study (see Table 2.3) is 0.86 ± 0.09Å. A reasonable measure of the quality of a study is the ratio of the number of structures with an r.m.s. deviation smaller than the predicted

* This integral is evaluated by the Fortran program PROB in Appendix 1.

TABLE 2.5

AN ESTIMATION OF THE NUMBER OF COMPACT GLOBULAR STRUCTURES FROM THE GAUSSIAN APPROXIMATION

Protein	$\frac{\overline{\Delta r}}{\sigma}$ [a,b]	Number of structures N [c]
PTI	7.57	$.54 \times 10^{14}$
B5C	8.34	$.27 \times 10^{17}$
REI	9.02	$.11 \times 10^{20}$
RNS	9.52	$.12 \times 10^{22}$
FXN	9.93	$.69 \times 10^{23}$
SNS	10.05	$.22 \times 10^{24}$
SOD	10.32	$.35 \times 10^{25}$
MBN	10.38	$.65 \times 10^{25}$
T4L	10.70	$.20 \times 10^{27}$
ADK	11.59	$.43 \times 10^{31}$
CNA	12.87	$.30 \times 10^{38}$
TIM	13.16	$.13 \times 10^{40}$

a $\overline{\Delta r}$ is evaluated from equation (8)

b σ is the average standard deviation from Table 3

c N is computed from equation (12)

structure to the number of structures with an r.m.s. deviation smaller than a random structure. To produce a more reasonable scale, the ratio of the logarithms of the number of structures (N) is used giving a quality index (Q) where:

$$Q = 1 - (\log(N \text{ predicted})/\log(N \text{ random}) \qquad (2.14)$$

where N is computed from equation (2.12) with μ equal to the r.m.s. deviation for the best predicted structure and $\sigma = 0.86$ when the interatomic distance r.m.s. is used, and $\sigma = 1.58$ when the rotation method is used.

The values of Q for several predictive studies are given in Table 2.6. In the range of values considered here, a difference of 0.05 in Q is approximately a ten-fold reduction in the number of structures. The results given in Table 2.6 indicate that only Kuntz et al. (1976) and Cohen et al. (1979) have succeeded in reproducing a significant fraction of the tertiary structural motifs seen in PTI and MBN respectively. In future folding studies, we suggest that the quality index (Q) as well as the r.m.s. deviation be reported as a quantitative standard for success.

6. Conclusion

Efforts to understand more about the nature of structural comparisons based on the r.m.s. deviation have led to several conclusions:

(1) The rotation method provides a much more effective and significant method of comparing structures than the interatomic distance method.

(2) The random walks presented here overcome the excluded volume problem without introducing too much bias. This method may prove useful in other simulation studies such as those of Gō and Taketomi (1979) on thermal denaturation.

(3) The significance of any structural comparison should be judged in

TABLE 2.6

AN EVALUATION OF PAST PROTEIN FOLDING STUDIES

Study	Number of Residues	r.m.s. deviation reported for predicted structure $\overline{\Delta d}$ (Å)	r.m.s. deviation for a random structure with m residues $\overline{\Delta d}$ (Å)	Q[a]
Levitt (1976) PTI with helices	58	6.2	6.86	.20
Levitt (1976) PTI	58	8.5	7.13	.28
Warshel & Levitt (1976) CPV with helices	108	7.4	8.15	.17
Warshel & Levitt (1976) CPV	108	8.15	8.48	.08
Kuntz et al. (1976) PTI with disulphides	58	4.7-6.5[c]	6.86	.09-.50
Cohen et al. (1979) MBN with helices	99[b]	3.6	7.97	.76
Robson & Osguthorpe (1979) PTI with helices	58	6.0	6.86	.22

a computed from equation (14)

b only the alpha carbon positions of the helical residues are predicted

c a range of r.m.s. deviations for all of the PTI folding simulations was reported

the light of the value expected for a random structure of the same size.

(4) The location of helical structure along the polypeptide chain does not significantly restrict the number of reasonable conformations for a compact globule.

(5) The difference in these figures can be used to assess the relative merits of various methods for predicting the structure of a protein from its amino acid sequence.

CHAPTER III

HELIX-HELIX INTERACTIONS

Crystallographers and theoreticians alike have noticed that the structure of many proteins is largely influenced by the relative arrangement of α-helices, β-strands, or both. The focus of this Chapter is the helix-helix interaction. Three different properties of helix-helix interactions are investigated:

(1) Their importance in determining the structure of all α-helical proteins;

(2) Their possible functional roles; and

(3) Their possible importance as intermediates in the folding transition between extended and native states.

1. Structural Role of Helix-Helix Interactions

1.1 Prelude - Myoglobin

Based on a study of some observed helix-helix interations, Richmond and Richards (1978) have proposed tentative rules for predicting possible interaction sites from the amino acid sequence. It is necessary to know, or to assume, that given portions of the chain are in a helical conformation. Site probabilities are calculated from possible changes in solvent contact area for each appropriate cluster of residues along the helix

surface. Although changes in solvent accessibility may also be useful in secondary structure prediction (Richards & Richmond, 1977; Rose, 1978), the emphasis in this Chapter is on the further assembly of preformed secondary structural units. The predicted interaction sites specify normals to the helix axis which are defined both as to position along the helix axis and angular orientation around the axis. The helices are connected by peptide chains which are described only by length. This idealised form of the chain, a sausage and string model with identified interaction sites, is then folded by forming site pairs according to some simple rules and geometrical restrictions. With myoglobin as an example, the following restrictions are imposed:

(1) The distance between the end points of consecutive α-helices is no greater than could be spanned by the number of residues in the polypeptide chain joining them.

(2) There are less than 13 contacts of less than 7.5Å between helix axis points on different α-helices. The helix axis points are the projection of the C^{α} atoms onto the helical axis.

(3) The structures are compatible with the incorporation of a heme ring between the distal and proximal histidines.

All possible pairings are investigated to yield a list of acceptable structures. Results of this procedure on myoglobin and preliminary results on other all-helical proteins are presented. The extent to which this procedure can provide a reduced conformation space which might be of practical use as input for more detailed packing or energy minimisation programs is assessed.

1.2 Definitions

1.2.1 Input Data

The sequence of sperm whale myoglobin together with the location

of helical segments as defined by Watson (1969) were used (see Table 3.1).

1.2.2 Structural Data

All protein crystallopgrahic data were taken from the files of the Protein Data Bank at the Brookhaven National Laboratory (Bernstein et al., 1977). The specific file used for deoxymyoglobin was 3MBN (Takano, 1977).

1.2.3 Description of an idealised α-helix

The representation of helices as cylinders of close packed spheres has been described by Richmond and Richards (1978). Each sphere represents one residue. The α-helix can be closely simulated by such a representation with spheres of diameter 6.39Å placed with their centres 4.12Å from the helix axis. The pitch is about 4.5Å as required for an actual α-helix. The helix is described in cylindrical coordinates (z,r,θ) with the Z axis collinear with the helix axis and positive in the direction of the C terminus of the peptide chain. The side chains are best enclosed by the spheres when the radius through the centre of the β carbon atom passes through the centre of the sphere which is then about 1Å outside the C_β position. The intersection of this radius with the helix axis defines the helix axis point for this residue. Such points are spaced 1.50Å along the axis and consecutive radii have an angular separation of 100° as in an ideal helix. From known positions in the peptide sequence, the relative positions of two residues are described by the Z coordinates and angular separation of the two radii representing these residues. In this Chapter, no other parameters of the actual residues enter into the geometrical description of a single helix, but side chain character will affect the packing of two helices.

TABLE 3.1

Interaction Site Prediction Data from the Myoglobin Sequence[a]

Helix Designation	Residues in Helices: Sequence Numbers of Termini N	C	Number of Residues in helix	between helices	Predicted Number of Possible Strong Sites	Potential Central Residue[b]	Interaction Class[c]
A	3	18	16		2	10	II or III*
				1		13	II
B	20	35	16		2	25	I
				0		28	II
C	36	42	7		0	—	—
				8			
D	51	57	7		0	—	—
				0			
E	58	77	20		3	65	I
				8		68	II or III*
						69	III*
						71	II
F	86	94	9		1	90	II
				5			
G	100	118	19		4	107	II or III*
						108	II* or III*
				6		110	II
						111	II* or III
						112	II
						114	II*
H	125	147	23		3	134	II or III*
						135	III
						142	II

[a] See Richmond & Richards (1978). Designation of helical residues is taken from Watson (1969).

[b] The following potential sites which have contact area changes above the cut-off values shown in Fig. 8 of Richmond & Richards (1978) were not included in the list given above. One or more of the surrounding residues in each of these cases is not in an actual helical segment. If secondary structure were unknown or uncertain, some or all of these would have to be tested: 17(II or III), 21(III), 29(III), 72(III), 75(II or III), 101(II), 104(III), 114(III), 127(II or III), 131(III), 142(III).

[c] Entries marked with an asterisk were not used in the site list for the calculations reported here. These deletions are explained in the text.

1.2.4 Helix Pairs

The geometrical relation of two helices is specified by four parameters: the perpendicular distance between the two helix axes, referred to subsequently as the contact normal; the dihedral angle between the two helix axis directions, the helix axis angle; and two angles relating the internal angle coordinate for each helix and the contact normal, the skew angles (see Figure 3.1).

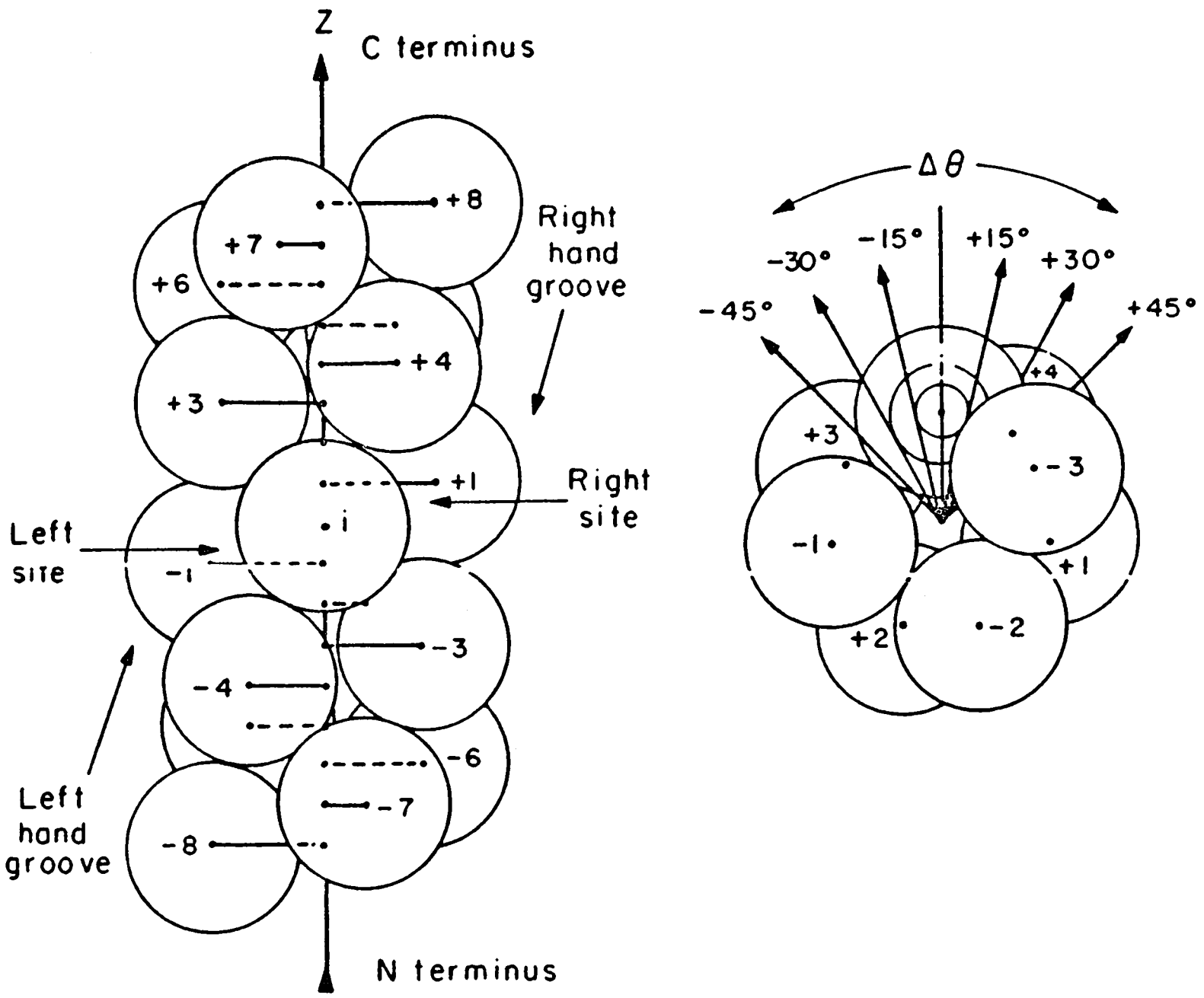
When two helices pack together, the residues of one helix will tend to fit in between the residues of the other helix in one of several characteristic arrangements (Crick, 1953; Chothia et al., 1977). With an α-helix modelled by a helical net (see Figure 3.2), Chothia et al. (1977) concluded that helix-helix packing was due to the interdigitation of ridges on one helix (i) into grooves on the other helix (j) and vice versa. They identified two ridges and associated grooves, one formed from residues i, i±4, i±8, and the other formed from i, i±3, i±6. Packing of helices then fell naturally into three categories with associated helix axis angles: i±3 ridges into j±3 grooves, -82°; i±4 ridges into j±4 grooves, -60°; and i±4 ridges into j±3 grooves, +19°. In the representation used here the radii simulating the two closest residues, i and j, on the two helices, the central residues, will not be collinear with the contact normal but will have angular displacements which are related to the approximate size of the residues, the displacement being larger for the large residues. This is shown schematically in Figure 3.1, for the skew angles assumed in this study. Following Richmond & Richards (1978), the types of helix-helix interactions are considered to fall into 3 classes which are related both to the size of the central residues, the length of the contact normal, and the helix axis angles. The assumed mean values are listed in Table 3.2.

1.3 Prediction of Potential Interaction Sites

The prediction scheme for strong helix-helix interaction sites

FIGURE 3.1

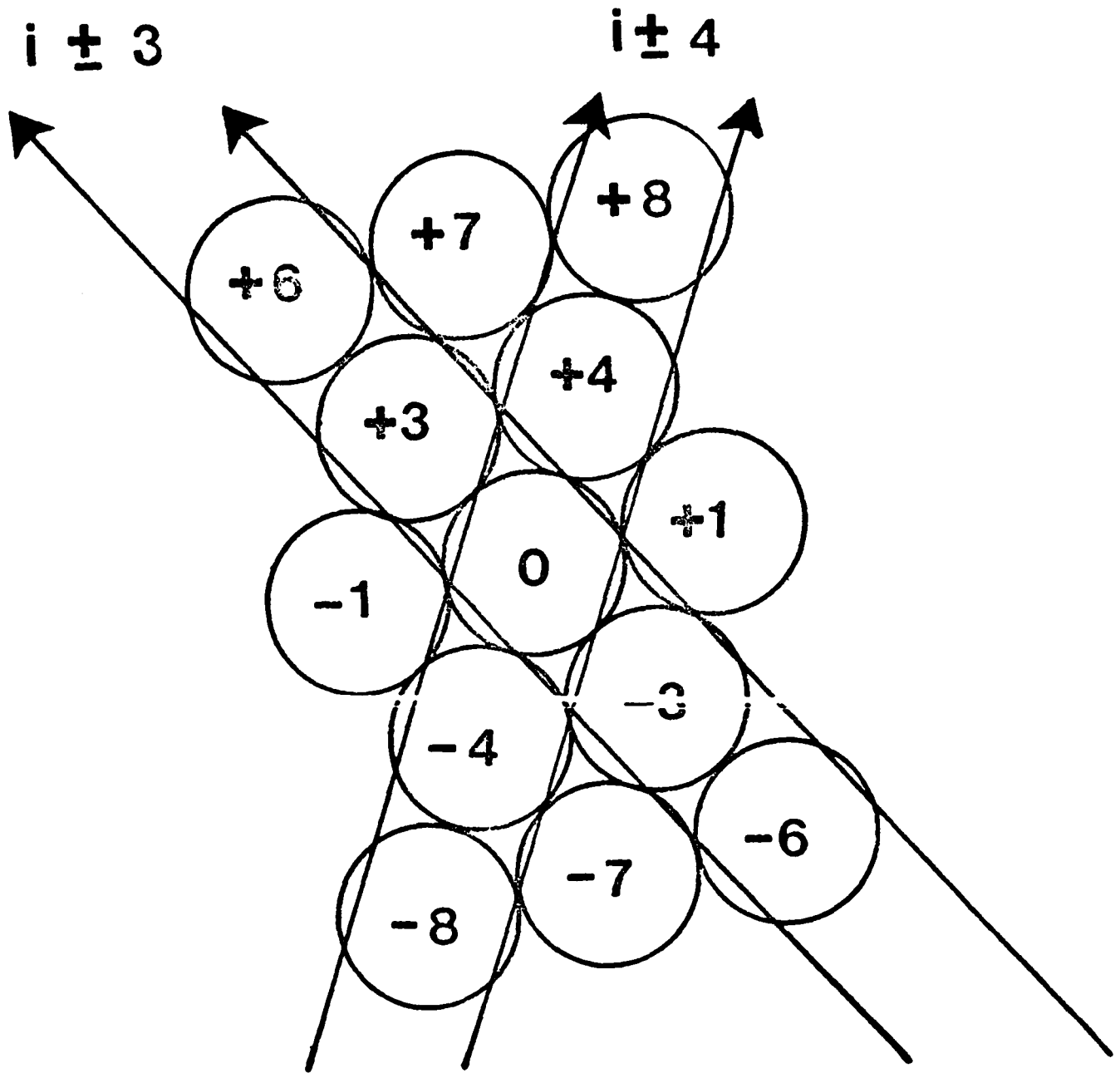
A Close-packed spheres model for an α-helix



Sketch of a helix of close packed spheres approximating an α-helix. The left panel is the view perpendicular to the helix axis along the radius through the centre of a central residue i. The helix axis points for the other spheres are shown at the end of the projected radii. The right and left grooves and the probable packing sites for the central residue of a second helix are shown. The right panel is a view along the helix axis and shows the directions of the contact normal of residue i for various interaction classes. The positive and negative skew angles shown are related to the assumed size of the central residue as it affects packing in the left or right site.

FIGURE 3.2

Ridges and Grooves on the Surface of an α-Helix



The residues on one surface of the α-helix are represented by circles. The grooves between residues i±3 and between residues i±4 are shown. Residue 0 is the central residue.

TABLE 3.2

Assumed Parameters for the Three Helix-Helix Interaction Classes

Interaction Class	Central Residue Types	Assumed Length of Contact Normal (Å)	Acute Helix Axis Angle[a]	Skew Angles
I	Gly only	7.5	-80°	±15
II	Ala, Val, Ile, Ser, Thr, Cys	8.5	-60°	±30
III	Class II + Leu, Met	10.5	19°	±45

[a] Adapted from Chothia *et al.* (1977).

proposed by Richmond & Richards (1978) is based on the probable magnitude of the change in solvent accessible area in going from the separated helices to the helix pair. Given the primary structure and information (given or assumed) that the helix actually exists, the regions of the helix that might be involved in such contacts can be rank ordered on potential area change. Each site consists of a central residue i and a subset of the nearest neighbours (±4, ±3, ±1 for Class II for example, see Fig. 3.1). Prediction lists are prepared independently for each class. One can decide on the basis of other criteria how far down these rank ordered lists to go in setting up the search procedure (see Figure 3.3). The helix data and list of potential sites used in this study are given in Table 3.1.

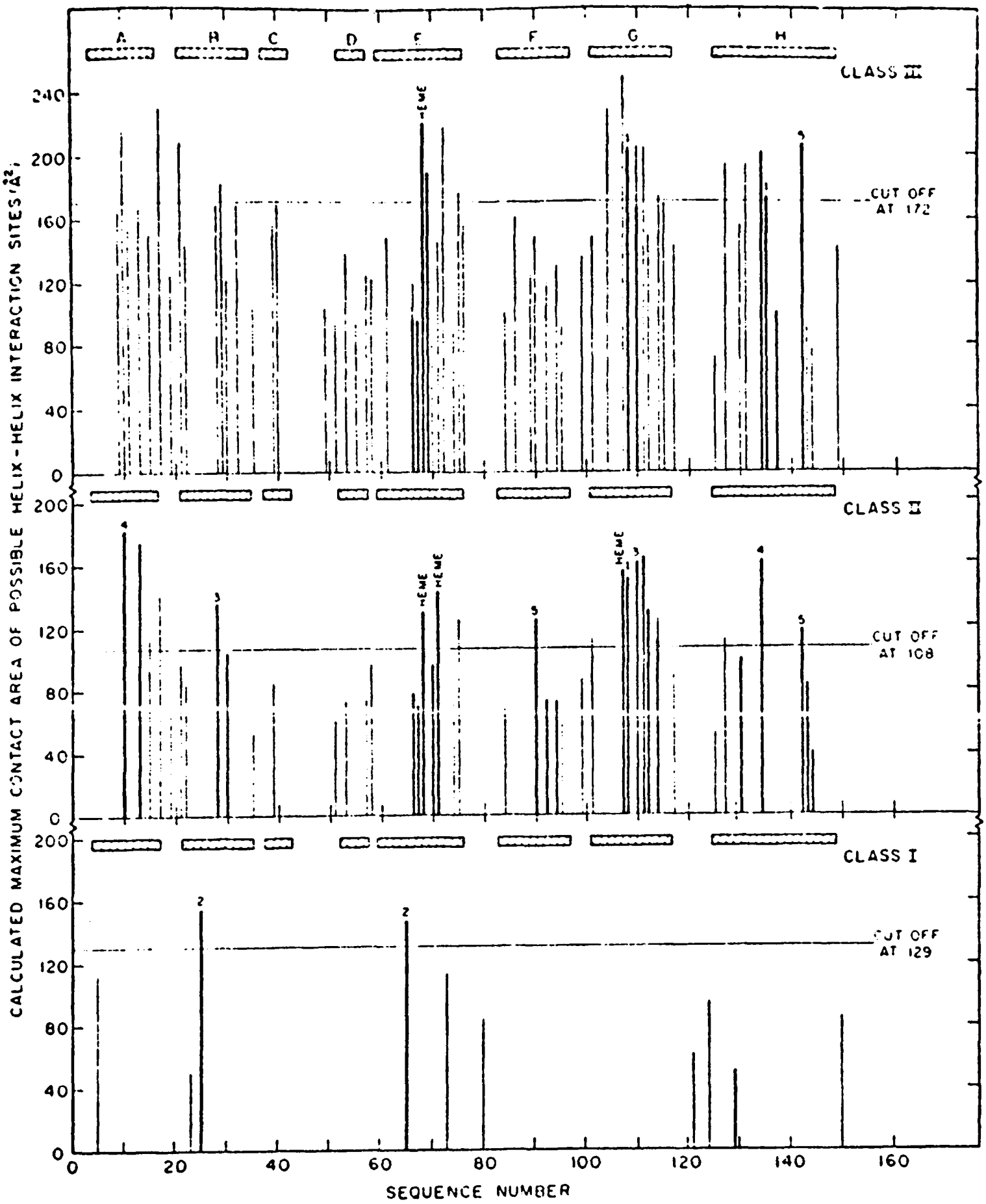
Nine sites were eliminated from the complete list of predictions due to steric conflicts when several sites occurred on one helix. Two rules were developed to handle these complications based on the geometry of the constellation used to assess the patch strength:

(1) If i is a central residue in a Type I or III interaction, residues i+3 and i+4 or i-3 and i-4 cannot also be used as Type I or III sites.

(2) If i is a central residue in a Type III interaction, residues i+4 or i-4 cannot also be used as Type II sites.

Since residue 65 was required as a Type I site to simplify the calculation, residues 68 and 69 cannot be Type III sites. Residues 107, 108 and 111 form the centre of a Type III constellation which can best be represented by residue 111. Residue 107 and 108 cannot be the central residues of other Type III sites, but 107 still can be used for a Type II site. Residues 107, 110 and 112 were kept as Type II sites thereby excluding 111, 114 and 108. Residues 10 and 134 were chosen as Type II since the per residue average contribution to the patch strength was larger for Type II than III and no contingency in the present algorithm permits a residue to be central to two different classes of interaction. It is important to note that no

FIGURE 3.3. Helix-Helix Interaction Sites in Myoglobin



Site predictions for all possible central residues by class. The predicted maximum area change for any site is shown on the ordinate. The suggested area cut-offs to be used to isolate only the potential strong sites are shown by the horizontal lines. The site identification numbers for the positions actually used are shown over the arrow. The horizontal bars show the positions of the actual helices in the sequence. Taken from Richmond & Richards (1978).

hydrophobic patches are ignored through these simplifications. Instead, two central residues which are involved in the same patch are represented by one or the other. The translational flexibility of the algorithm discussed below prevents the exact choice from being critical.

1.4 The Assembly Algorithm

The first program, FOLD, produces lists of all possible site pairs based solely on combinatorial sums subject only to the following criteria:

(1) Separate sites on the same helix cannot be paired.

(2) No two helices can form more than one pair.

(3) All helices must appear in each list.

(4) Site pairs are made only between sites of the same class.

A copy of the coding for FOLD is given in Appendix 2.

From the site data in Table 3.1 and always including the B25-E65 class I pair, FOLD produced 10,370 lists of pairings acceptable by these criteria, required 30 minutes of CPU time on a PDP 11/90. If the B-E pair was omitted and the sites were left open, FOLD produced an additional 9600 lists of pairs. The B-E interaction was fixed to reduce the computation time. In a more complete study, this too would have been variable. Each list showed 5 pairs between helices A,B,E,F,G,H, and 5 unused sites. Helices C and D do not appear in the lists as they contain no predicted sites. Fifteen sites randomly paired without regard to class or helix position and leaving five unpaired would have given 135,135 combinations.

The second program, BUILD, is designed to test which of the lists from FOLD can lead to stereochemically acceptable structures. As described earlier each site pair actually represents eight possible structures. The acute helix angle is defined by the Class, but there are two possible N → C directions for the second helix axis. Each central residue has 2 possible skew angles whose numerical value is again defined by class. The total

permutations to be tested are thus 8. Even with only 5 pairs of sites the total number of permutations to be tested is 8^5 or 32,768. Each of the 10,370 lists, in principle, must be processed through this number of trial structures, for a grand total of 3.4×10^8, unless early termination of the search occurs.

Only two approximate selection criteria, or filters, have been used in BUILD:

(1) The end-to-end distance of helices consecutive in the sequence had to be less than a maximum permissible number; and

(2) Only a limited number of close contacts were permitted in an accepted structure.

The coding for BUILD is in Appendix 2.

The permissible end-to-end distance between the C terminus of one helix and the N terminus of the next helix in the chain was set equal to, or less than, 3(m+1) + 3, where m is the number of non-helical residues between the two helices. The unit distance of 3Å was chosen as a mean value for the effective length of a residue in the irregular chain conformation assumed to connect the two helices. The additional 3Å is added to represent the distances from the helix axis to the inner edge of the helix cylinder. The further processing of a particular branch in the search was discontinued when any distance failed this test.

Distances between the axis points of neighbouring helices were checked. Close contacts (bad contacts) were scored when this distance was less than 7.5Å. When the number of such contacts exceeded 50 the structure was discarded. Fifty bad contacts corresponds to 10 misplaced axis points (triangular matrix check). This rejection number is, of course, a variable parameter to be adjusted on the basis of experience. Helices which are paired are automatically placed such that there are no bad contacts between them. Helices not actually paired may collide, and it is these which are checked for collision. No test was made for collision of the connecting

residues although this could occur, in principle, even with no bad contacts between the helices themselves.

In this particular study, a list from FOLD consisted of 5 pairs of residues in an ordered sequence. The first residue and the contact normal defined a coordinate system. The other helix of the pair was then placed in this reference frame with the correct distance and helix angle for that site class. The third helix was then placed in a vacant site of one of the first two as specified by the list. The end-to-end distance check on segment lengths between consecutive helices was then performed. This procedure was repeated with succeeding helices in the list. The calculation terminated whenever the distance check failed. All subsequent lists with the same entries up to that level were also discarded. The collision check was made only on structures satisfying all endpoint restrictions. To limit the computation, the structure of the BE pair was fixed in the antiparallel arrangement and the skew angle for B25 was set in the + position, giving a four-fold reduction to about 10^8 possible structures.

1.5 Examination of the Trial Structures

At the end of the computation by BUILD, out of the vast possible total, only 121 structures survived the two rejection criteria. This set of structures was sorted on the value of the sum of the lengths of the helix end point distances. Sixty-five of the structures had the G and H helices in the parallel orientation thus contributing a long connecting link. Although formally satisfying the end point length criterion, construction of the six residue segment with Lab-quip models failed to produce a sterically acceptable connection. Such an error would easily be found at the next level of structure refinement, but might be eliminated at this stage with an appropriate re-definition of the end point criterion. This has not yet been tried.

Of the remaining 56 structures, 36 would be eliminated if the acceptable bad contact criterion were reduced from 10 axis points to 5. Although satisfactory in this case, such a requirement might be too stringent in general. Further experience with other structures will be required to determine the most useful value.

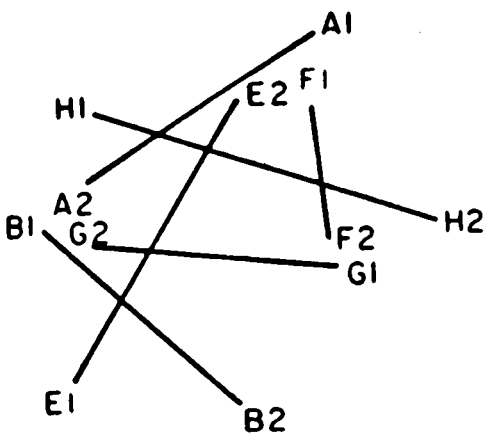
The remaining 20 structures fall into four natural groups where the A,B,E,G and H helices are identically paired in each group but the F helix may pair with one or more of the remaining open sites. One example from each groups is shown in Fig. 3.4. In the first, the distinguishing feature is that B is paired with G and G with H such that the left handed groove of residue 110 and the right handed groove of residue 111, both in helix G, are used. In group 2, the left handed groove is used in both instances. Groups 3 and 4 resemble the groove choices of groups 1 and 2 respectively but B is paired with H.

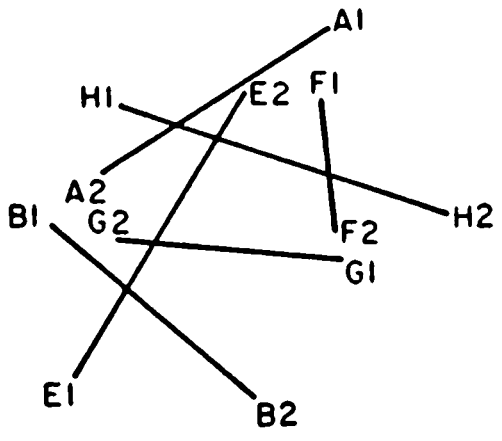
Since the helices are ideal, the helix axis representations created by BUILD can be transformed in C_α representations. Atoms are placed 2.29Å along a vector perpendicular to the helix axis emanating from each helix axis point. The relative phasing of these vectors is 100° between consecutive residues. This creates an α-helix with a pitch of 1.5Å and 3.6residues/turn. The absolute phasing of a helix follows from the position of the contact normal. REALSPACE (see Appendix 2) takes the axis coordinates from BUILD and generates a complete set of C_α coordinates for all residues by interpolating positions for the non-helical residues. The link between a helical axis representation and its C_α representation can be seen by comparing Figures 3.5 a and b.

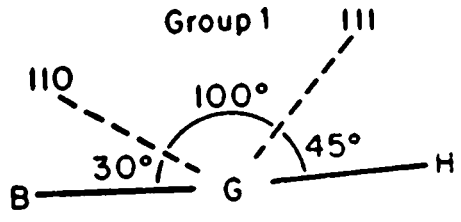
The various properties of the twenty remaining structures were then considered. They are catalogued in Table 3.4. The structures differ in the number of close contacts, C^α-C^α separations, which violate van der Waals radii. Varying agreement with the crystal structure defined by the root mean square deviation, is evident. In each structure, there is a close

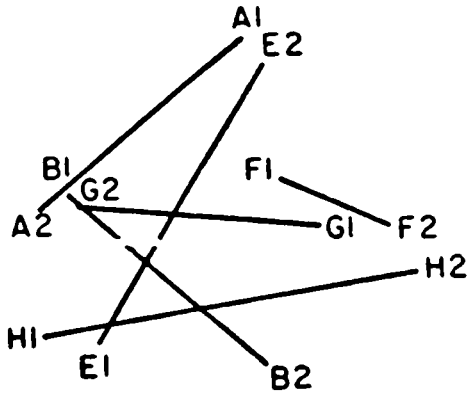
FIGURE 3.4

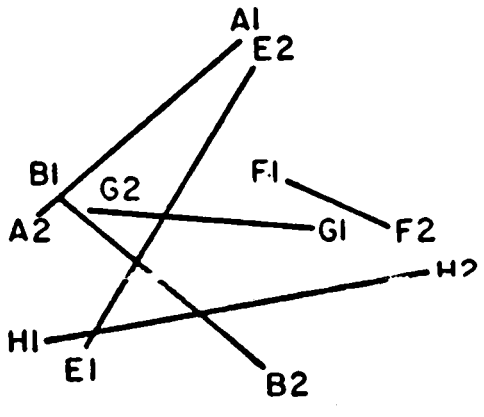
Representatives of the Four Classes of Structures Produced by BUILD on Myoglobin

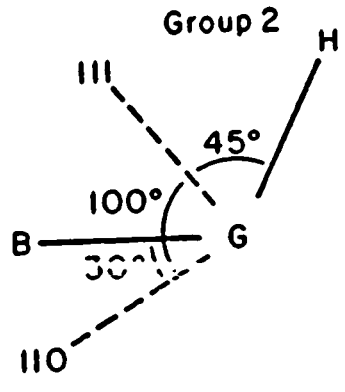


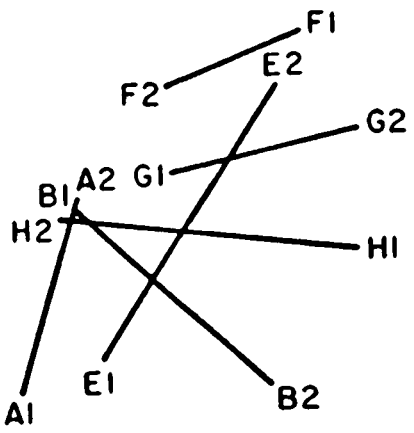


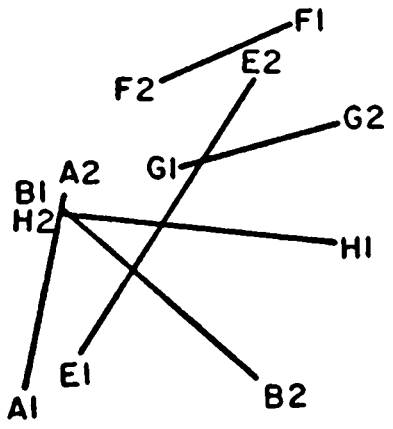


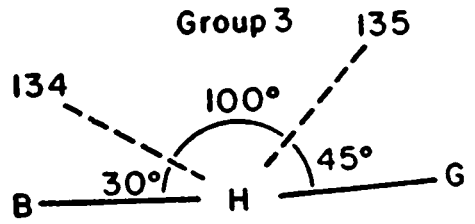


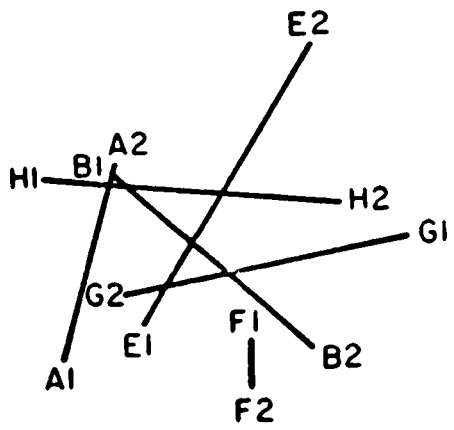


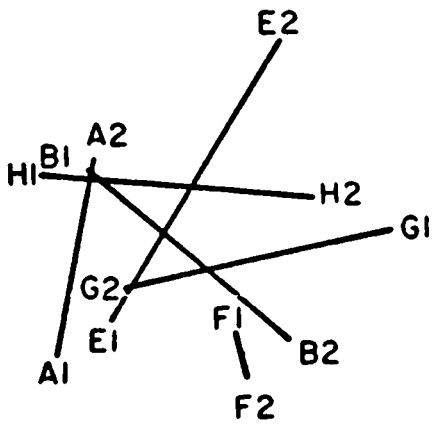


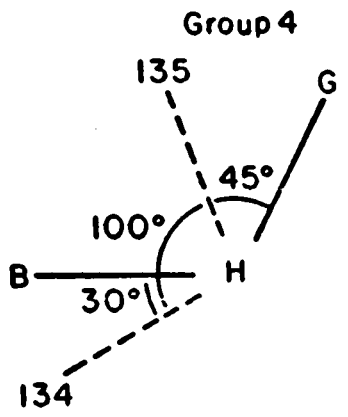












Stereo diagrams of one member of each of the four groups which together comprise the final 20 predicted structures from the assembly program. Within each group, the A,B,E,G and H helices are identically related, only the position of F differs. The angular relations between the B, G and H helices which characterise the different groups are shown to the right of each stereo pair. The helices are identified by letter. The number 1 identifies the N terminus of the helix, 2 the C terminus.

TABLE 3.3

Characteristics of the Predicted Structures

Group No.	Number of Structures in Group	Number of bad contacts between axis points of adjacent helices (distances <7.5Å)		R.m.s. Deviation of helix axis points between actual and predicted structures (Equation 3.1)	
		Mean	Dev.	Mean (Å)	Range (Å)
1	9	5.6	6.6	3.6	3.4 - 4.2
2	2	13.0	0.0	5.4	5.4 - 5.5
3	6	11.0	0.0	7.4	7.2 - 7.6
4	3	11.7	1.2	7.2	7.1 - 7.4

TABLE 3.4.

Close contact data on 20 candidates for Myoglobin structure

Group	Structure Number	Contacts of < 3.6Å between alpha carbons			R.M.S. deviation from crystal structure (Å)
		$C\alpha_1$	$C\alpha_2$	Distance Å	
I	97	25 Gly	65 Gly	3.21	4.53
		93 His	101 Ile	3.20	
		94 Ala	101 Ile	2.19	
	88	25 Gly	65 Gly	3.21	4.49
		92 Ser	100 Pro	2.73	
		93 His	104 Leu	2.91	
	87	25 Gly	65 Gly	3.21	4.51
	86	25 Gly	65 Gly	3.21	4.48
		87 Lys	104 Leu	2.91	
		88 Pro	100 Pro	2.73	
	95	25 Gly	65 Gly	3.21	4.92
		86 Leu	101 Ile	2.14	
		87 Lys	101 Ile	3.20	
	85	25 Gly	65 Gly	3.21	4.55
	98	25 Gly	65 Gly	3.21	4.57
	100	25 Gly	65 Gly	3.21	4.72
	93	25 Gly	65 Gly	3.21	6.05
II	62	16 Lys	23 Gly	3.11	8.83
		18 Glu	20 Asp	2.85	
		25 Gly	65 Gly	3.21	

TABLE 3.4, continued.

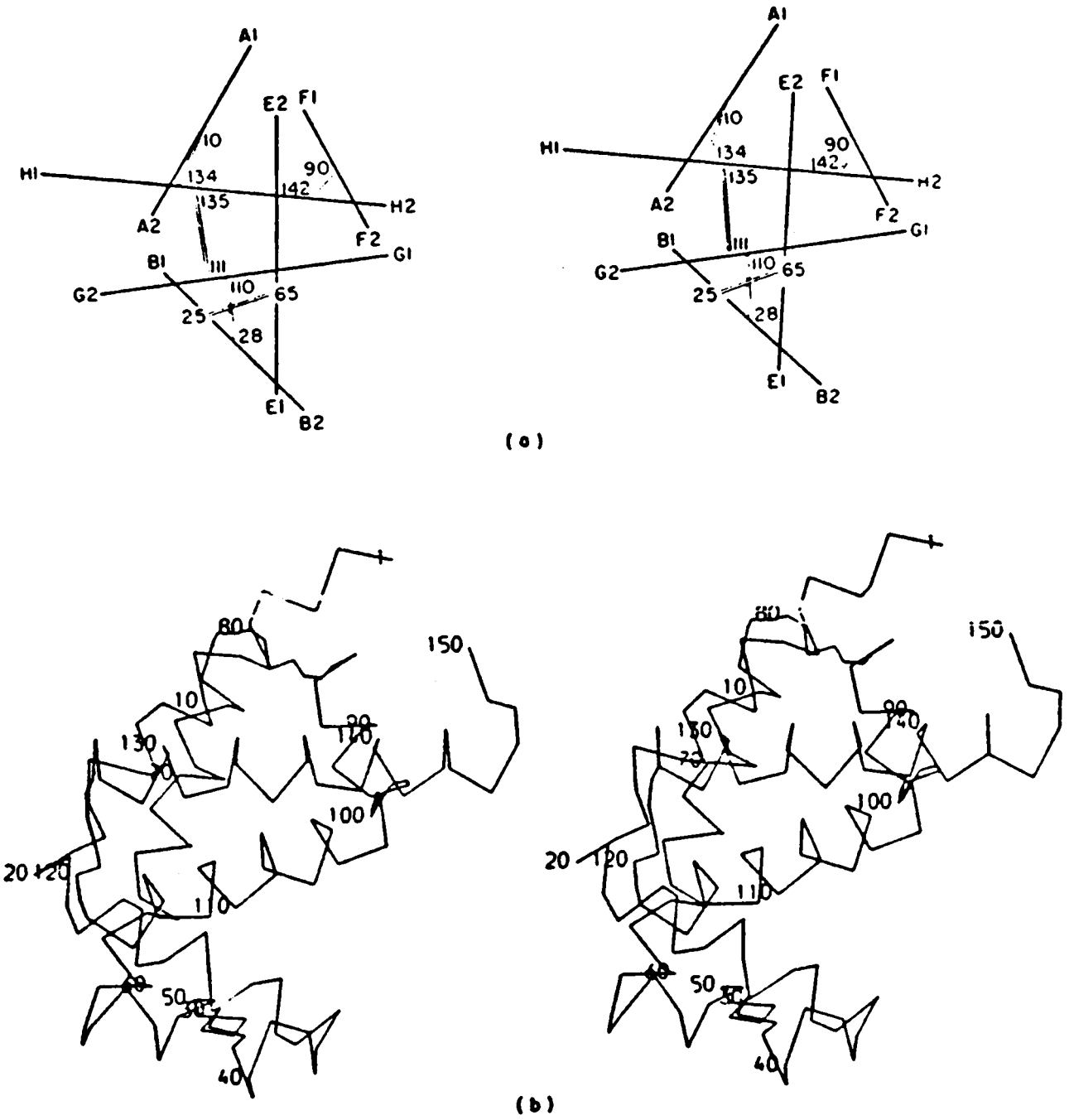
Group	Structure Number	Contacts of < 3.6Å between alpha carbons			R.M.S. deviation from crystal structure (Å)
		$C_{\alpha 1}$	$C_{\alpha 2}$	Distance Å	
II	61	16 Lys	23 Gly	3.11	8.83
		18 Glu	20 Asp	2.85	
		25 Gly	65 Gly	3.21	
	56	16 Lys	23 Gly	3.11	8.89
		18 Glu	20 Asp	2.85	
		25 Gly	65 Gly	3.21	
	76	16 Lys	23 Gly	3.11	9.09
		18 Glu	20 Asp	2.85	
		25 Gly	65 Gly	3.21	
III	158	16 Lys	23 Gly	3.11	9.24
		18 Glu	20 Asp	2.85	
		25 Gly	65 Gly	3.25	
	157	16 Lys	23 Gly	3.11	9.31
		18 Glu	20 Asp	2.85	
		25 Gly	65 Gly	3.21	
	159	16 Lys	23 Gly	3.11	9.16
		18 Glu	20 Asp	2.85	
		25 Gly	65 Gly	3.21	
	160	16 Lys	23 Gly	3.11	9.11
		18 Glu	20 Asp	2.85	
		25 Glu	65 Gly	3.21	
	167	17 Val	20 Asp	2.42	9.20
		18 Glu	20 Asp	1.54	
		25 Gly	65 Gly	3.21	

TABLE 3.4, continued.

Group	Structure Number	Contacts of < 3.6Å between alpha carbons			R.M.S. deviation from crystal structure (Å)
		$C\alpha_1$	$C\alpha_2$	Distance Å	
IV	14	14 Trp	20 Asp	3.25	6.92
		17 Val	20 Asp	3.24	
		25 Gly	65 Gly	3.21	
	31	19 Trp	20 Asp	3.25	6.69
		17 Val	20 Asp	3.24	
		25 Gly	65 Gly	3.21	

FIGURE 3.5

Two Representations of a Predicted Myoglobin Structure



(a) Helical axis representation of a candidate for the myoglobin fold. On inspection, this structure most resembles myoglobin. Helices are labelled alphabetically with 1 corresponding to the N-termini and 2 to the C-termini. The double lines represent the contact normal used in the construction of this model. The numbers of the residues central to the helix-helix interaction mark the ends of the contact normals.

(b) Alpha carbon representation of the structure shown in (a). The idealised geometry of an α-helix, a pitch of 1.5Å per residue with 3.6 residues per turn so that each alpha carbon lay 2.29Å from the helix axis, is used. The relative phasing of residues on interacting helices follows from the position of the contact normal and the magnitude of the skew angle. Residues in the chain joining 2 helices were placed so that the distance between consecutive α-carbons is 3.81Å.

contact of 3.2Å between Gly 25 and Gly 65 because the contact normal joining these two residues is required to be 7.5Å in length. However, the small volume of glycine residues, 66.9Å^3 (Richards, 1974) enables the two α-helices to close pack. Gly-Gly interactions in other proteins including various hemoglobins can have contact normals as close as 6.5Å with C^{α}-C^{α} separations below 3Å. With the exception of the Gly-Gly interaction, the close contacts are localised in the FG corner (I:97, I:88, I:86, I:85)* and the AB corner (Groups II, III and IV). It is interesting to note that the use of a C^{α} model did not aid in reducing the number of allowed structures.

The r.m.s. deviations between all possible pairs of the 20 structures are presented in Table 3.5. Structures considered are grouped into the four categories previously described, with each member of a group having the same relative arrangement of the B,E,G and H helices. Thus, the matrix of r.m.s. deviations has small off-diagonal elements corresponding to the deviation between members of the same group. It should be emphasised that this grouping provides nothing more than a convenient nomenclature for considering the 20 candidates.

1.6 The heme group restriction

The aim of any protein folding scheme is to determine the correct fold from a consideration of all the chemical interactions. The work previously described considered the fold of apomyoglobin. Since the interaction of the heme ring with the polypeptide chain plays a role in determining the myoglobin fold, the effect of the heme group must be examined (I.D. Kuntz, personal communication). Instead of considering the energetics of the binding of the heme ring to the twenty candidates for the myoglobin fold, we have relied on biochemical information which provides estimates of the distance between the heme iron and specific residues along the poly-

* A "close contact" is scored when two Cα atoms lie within 3.6Å. This is 90% of the sum of their van der Waals radii.

TABLE 3.5

The R.M.S. deviation between all possible pairings of the twenty trial structures

	88																				
	93	6.9Å																			
	95	3.3	5.9																		
	87	2.8	7.5	4.0																	
	97	.3	5.5	2.5	4.1																
Group 1	100	6.1	3.1	4.2	6.7	4.2															
	86	2.7	6.6	3.5	1.2	3.4	5.7														
	98	5.6	3.6	4.2	7.1	4.2	2.5	6.1													
	87	1.5	7.7	4.1	2.5	4.0	7.2	2.7	6.7												
	61	11.8	9.9	11.1	12.5	11.2	10.2	12.0	9.9	12.4											
	62	11.4	9.9	11.0	12.0	10.9	10.3	11.5	16.8	11.9	1.3										
Group 2	56	12.4	11.1	11.5	13.0	11.6	10.9	12.5	10.6	12.8	3.6	4.0									
	76	12.1	10.7	10.8	12.6	10.8	10.5	12.2	10.4	12.4	4.6	4.8	4.9								
	157	16.0	a	14.3	14.7	a	13.8	17.8	17.9	14.8	a	a	13.2	13.2							
	158	14.5	a	a	14.6	a	13.7	14.4	13.8	14.6	a	a	13.2	13.2	1.3						
Group 3	159	17.4	14.0	a	14.8	14.3	13.7	17.9	13.8	14.7	14.2	a	13.2	13.1	2.5	2.7					
	160	14.5	a	a	14.6	a	13.8	14.4	13.7	14.6	a	a	13.2	13.2	2.7	2.5	1.3				
	167	15.0	14.0	14.9	a	14.7	17.3	15.0	16.8	a	16.5	a	12.3	13.3	4.9	5.1	4.7	4.9			
	14	8.3	8.7	7.2	8.6	7.5	7.6	8.2	7.7	8.7	12.2	12.1	11.7	12.1	13.6	13.5	13.8	13.7	13.9		
Group 4	31	8.3	8.5	7.5	8.7	7.2	7.7	8.1	7.6	8.6	12.3	12.2	12.8	12.0	13.7	13.7	13.6	13.5	13.8	2.5	
		88	93	95	87	97	100	86	98	87	61	62	56	76	157	158	159	160	167	14	31

[a] convergence was not achieved as the derivative of the rotation matrix was singular. In practice, this indicates that the R.M.S. deviation was > 15Å.

peptide chain. Distance constraints are available for many proteins concerning the spatial proximity of specific residues and Kuntz and co-workers have found that these constraints provide a useful method for limiting the number of macromolecular conformations (e.g. Havel et al., 1979).

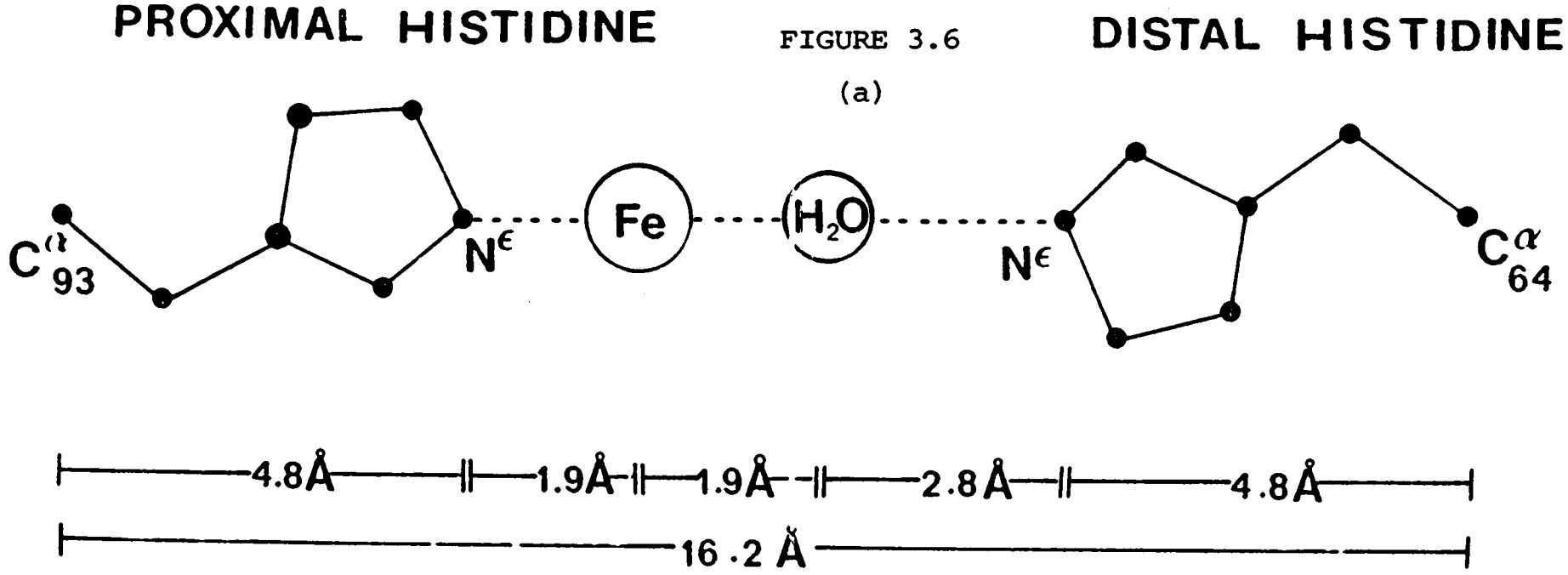
The specific distance information used in this study is the proximity of His 64, the distal histidine, and His 93, the proximal histidine, to the heme iron. Various spectroscopic studies, small molecule crystallography, and sequence information (Brill & Sanberg, 1967; Vuk-Pavolic & Siderer, 1977; Antonini & Brunori, 1971; Romero-Herrera et al., 1978; Johnson et al., 1978; Crute, 1959; Pauling, 1960) can be combined to show that the maximum C^{α}-C^{α} distance between the two critical histidines is 16.2Å (see Figure 3.6a). An additional 3.5Å is allowed to take into account imperfections in the model and conformational differences between holo and apo myoglobin. To be a candidate for the myoglobin fold, the distance between C^{α}_{64} and C^{α}_{93} was required to be less than ∿20Å. The actual distance between the C^{α}'s in the refined crystal structure of sperm whale myoglobin from the Protein Data Bank is 14.7Å.

When C^{α}_{64} and C^{α}_{93} are constrained to lie with 20Å, only 7 of the 20 structures remain (see Table 3.6). All of the structures in classes II, III and IV as well as two of the structures in class I fail to satisfy this essential criterion. The choice of 20Å as a cut-off does not seem to be too critical as the largest separation allowed was 18.35Å and the smallest rejected was 21.08Å. Considering that the diameter of the particle is only ∿35Å, the requirement that two specific C^{α} atoms be closer than 20Å is a remarkably effective constraint.

Not only must His 64 and His 93 lie close together, a cavity must exist to allow at least the iron atom to bury itself in the structure. An iron atom was placed at the midpoint of the line joining the two C^{α} atoms. The space filling constraint was generated by considering the iron atom to be liganded to four nitrogens from the heme ring as well as the proximal

FIGURE 3.6

(a)



(a) UV, visible and IR spectroscopy of myoglobins and hemoglobins have been used to identify an imidazole ring as the iron-protein or proximal ligand (Brill & Sandberg, 1967). Electron spin resonance studies of myoglobin in $H_2{}^{17}O$ indicate that a water is bound to the porphyrin ring (Vuk-Pavolic & Siderer, 1977) whose titration behaviour suggests a hydrogen bond to a histidine N_{ϵ} (Antonini & Brunori, 1971). Sequence studies demonstrate that His 93 must be the proximal histidine and His 64, the distal histidine. These assignments are confirmed by the work of Johnson et al. (1978) on leghemoglobin. 4.8Å is the maximum distance between the C_{α} and N_{ϵ} of a histidine. The separation between proton donor and acceptor in a hydrogen bond is 2.8Å (Pauling, 1960) and the length of an iron ligand is approximately 1.9Å from crystallographic studies of nickel etioporphyrin by Crute (1959).

(b) The octahedrally liganded heme iron is shown with two of the four nitrogens (N) from the porphyrin ring and the nitrogen of the histidine imidazole (N^{ϵ}). The ligand length is 1.9Å. The van der Waals radii of the C^{α} (2.0Å) and the half thickness of the aromatic rings (1.7Å) allows a C^{α} to approach within 4.0Å of the iron atom when 90% of the appropriate van der Waals radii are used. Thus the C^{α} atom of the histidine is 3.33Å from the nitrogens.

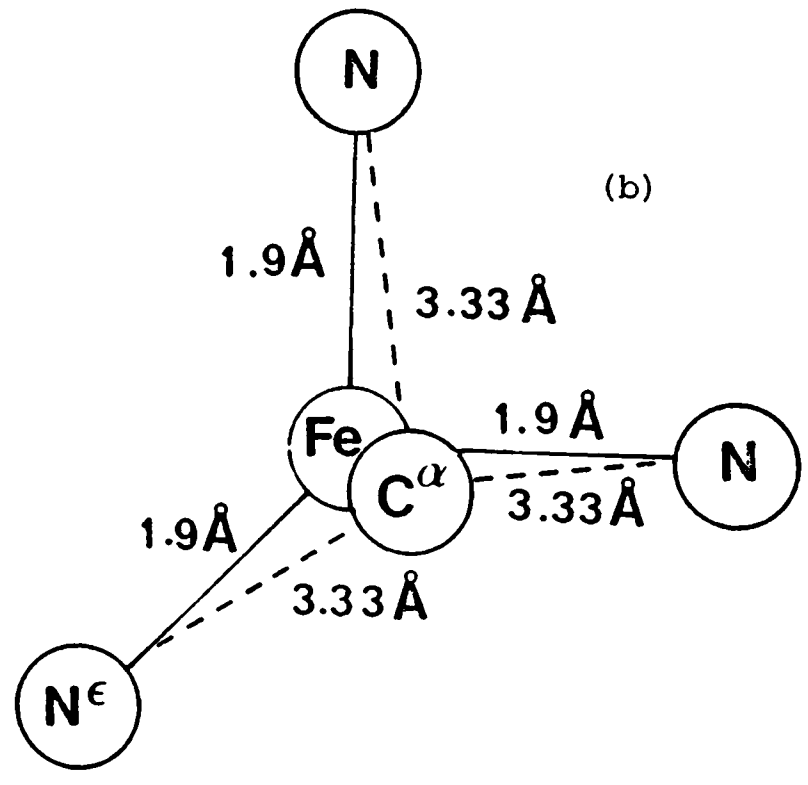


TABLE 3.6

Heme Group Restriction Data

Group	Structure Number	Distance Between $C\alpha^{64}$ and $C\alpha^{93}$ (Å)	Alpha Carbons Within 4.0Å of the Midpoint of the Line Joining $C\alpha^{64}$ to $C\alpha^{93}$	
I	97	16.98	-	
	88	13.08	-	
	87	13.44	Residue 67	3.20Å
	86	14.15	67	3.06
	95	23.35	*	
	85	14.95	67	3.22
			95	2.96
	98	28.04	*	
	100	18.35	106	3.77
			107	2.57
	93	15.95	28	2.13
			29	3.24
			31	3.77
II	61	27.91	*	
	62	24.67	*	
	56	30.48	*	
	76	31.50	*	
III	158	23.07	*	
	159	26.90	*	
	159	27.94	*	
	160	25.18	*	
	167	28.93	*	
IV	14	24.21	*	
	31	21.08	*	

* If the distance between the alpha carbons corresponding to His 64 and His 93 was greater than 20Å, close contacts to the midpoint of the line joining these alpha carbons were not computed as these two residues could not possibly co-operate in binding the heme ring.

histidine N^{ε} and the oxygen atom of the bound water. The ligands form a protective boundary about the iron atom prohibiting other atoms from approaching too close. If an iron-ligand distance is 1.9Å (Crute, 1959), then the shortest acceptable C^{α}-Fe separation would be approximately 4Å if 90% of the appropriate van der Waals radii, 1.7Å for the van der Waals radius of nitrogen plus 2.0Å for a tetrahedral carbon, are used (see Figure 3.6b).

This constraint reduced the list of candidates for the myoglobin fold from seven to two, I:97 and I:88 (see Table 3.6). The largest C^{α}-Fe distance that the cavity constraint rejected was 3.2Å. This corresponds to being within 75% of the van der Waals sphere of one of the atoms liganded to the iron.

1.7 The General Value of Distance Constraints

The importance of a specific distance constraint in reducing the number of theoretically predicted structures in a protein folding study depends on four factors: the sequential proximity of the residues involved; the accuracy with which the probe or physicochemical method can limit a particular interatomic distance; the nature of the approximations used in constructing the trial structures; and finally, the structural similarities and distinctions between the trial structures. The heme-histidine constraint, involving the E and F helices, is particularly well-suited to the myoglobin problem as the greatest variability seen in the twenty trial structures was in the F helix. Moreoever, a variety of biochemical and spectroscopic studies produced a consistent estimate of the distance constraint between the distal and proximal histidine.

To assess the value of other distance constraints in selecting the correct myoglobin structure, one residue was chosen arbitrarily on each of the six helices placed in the trial structures. This generates fifteen

distance constraints between helical pairs. The upper limit on the acceptable distance was the actual distance between the alpha carbons plus 3.5Å. The twenty trial structures were scanned to determine the number allowed by each hypothetical distance constraint. The most effective constraint is the heme-histidine constraint E64-F93 and three of the six most selective distance constraints involve the F helix (see Table 3.7). In every case, I:97 satisfied the constraint. If a pseudo cavity constraint was constructed in precisely the same manner as before, a sphere of 4Å about the midpoint of the line joining the two critical residues, it is clear that most trial structures are purged from the remaining list. Thus, the knowledge of specific cavities in a protein and the relative orientation of this cavity to a set of markers on the polypeptide is most useful. In practice these cavities in the apo enzyme might be the binding site of essential ions, prosthetic groups, or the substrate.

The disulphide bridge also provides an excellent distance constraint. Cohen et al. (1980a,b; Chapter 4) have applied this restriction in investigating methods for predicting the structure of all β-sheet proteins including the immunoglobulins. They found that this one specific distance often ruled out more than half of the predicted structures.

1.8 Comparisons of I:97 and I:88 to the native Myoglobin structure

A stereo diagram of I:97 with the structure of I:88 superimposed in dotted lines is presented in Figure 3.7. The structures are obviously very similar; the r.m.s. deviation between the two is 0.3Å. The difference arises from the fact that F90 is docked to H142 in I:97 and F90 is docked to E68 in I:88. Given the different contact normals used to place the F helix, it is coincidental that the F helix of I:88 is so close to the F helix of I:97. The existence of two such similar structures is to be expected since all possible combinations of potential central residues were used to generate

TABLE 3.7

The relative merits of various hypothetical distance and space filling constraints

Constrained Pair		Upper Limit on inter-atomic distance (Å)	Number rejected by distance constraint[a]	Number rejected by hypothetical space filling constraint[b]	Allowed Structures
A10	B26	23.7	0	19	1D
A10	E64	26.6	0	20	OD
A10	F93	28.0	1	15	4D
A10	G112	14.6	9	11	OD
A10	H138	16.6	2	17	1D
B26	E64	11.8	0	20	OD
B26	F93	23.6	8	11	1D
B26	G112	17.3	9	11	OD
B26	H138	24.3	0	20	OD
E64	F93	18.3	14	4	2DS
E64	G112	21.9	9	11	OD
E64	H138	22.9	0	20	OD
F93	G112	24.9	0	18	2D
F93	H138	17.1	8	10	2D
G112	H138	17.4	0	13	7DS

[a] The constraint was chosen to be the observed distance between the alpha carbons of interest +3.5Å.

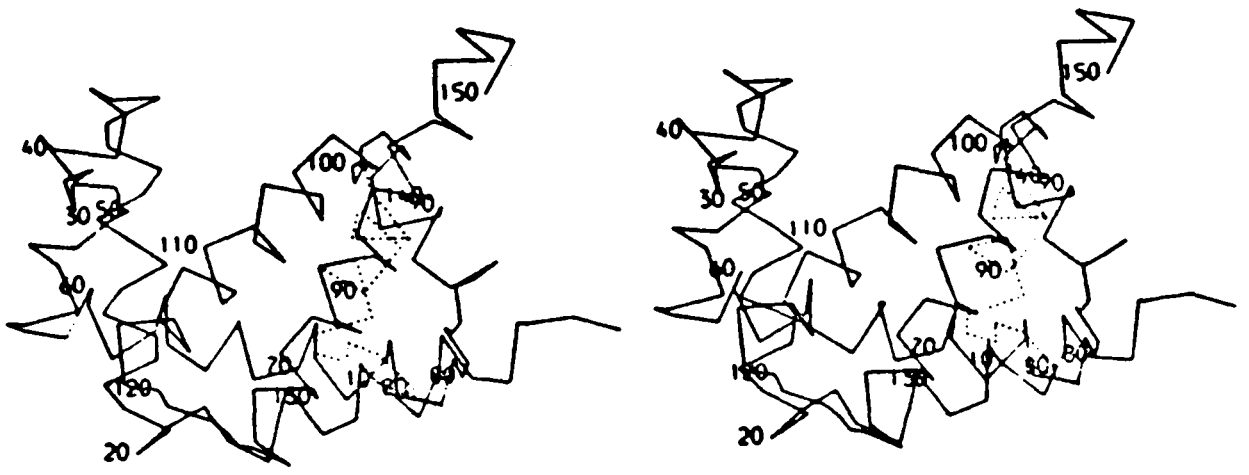
[b] The space filling criteria was applied only to structures which survived the distance constraint.

D I.97 satisfies the distance constraint.

S I.97 satisfies the space filling constraint.

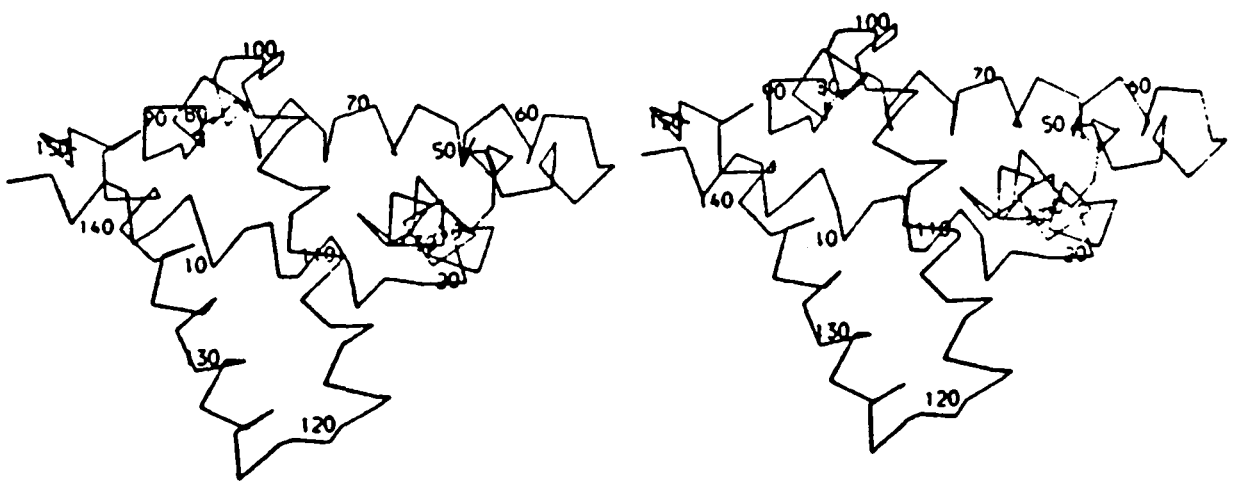
FIGURE 3.7

A Stereo Diagram of I:97 and I:88



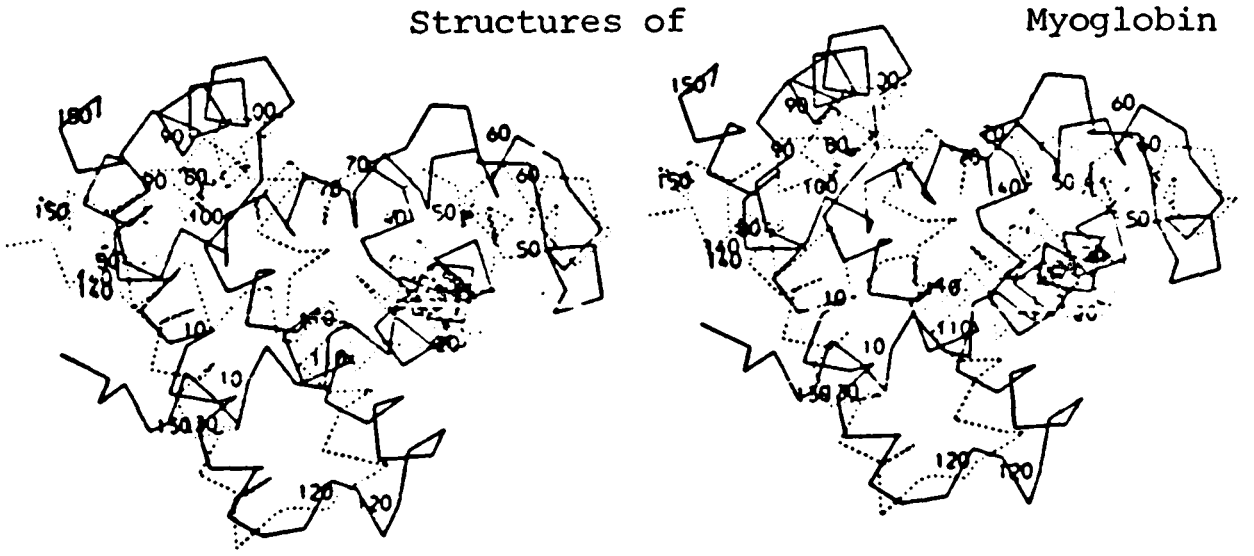
A stereo diagram of the two structures remaining as a result of the heme constraint. An alpha carbon representation of I:97 in solid lines with the differences of I:88 overlaid in dotted lines.

FIGURE 3.8. A Stereo Diagram of Myoglobin



An alpha carbon representation of I:97 rotated into the reference frame of the myoglobin coordinates (Bernstein et al., 1977).

FIGURE 3.9. A Stereo Diagram of the Predicted and Crystal Structures of Myoglobin



The super-imposition of the predicted myoglobin structure I:97 seen here in solid lines on the crystallographically determined alpha carbon coordinates connected by dotted lines.

these structures. The fact the I:97 is a better approximation of the crystal structure than I:88 is only clear a posteriori. Richmond & Richards (1978) have shown that E68 is a site used by the heme ring in reducing its accessible surface area upon binding to apomyoglobin. Thus, the E68-F90 interaction in I:88 would preclude the heme ring from adopting its preferred orientation. It is interesting to note that crystal structure of Hemoglobin M Hyde Park (Greer, 1971) which has a 70% occupancy for the heme ring in the β chain shows disorder in the C-terminal half of the F helix as well as the FG corner.

A stereo diagram of I:97 rotated into the coordinate system of myoglobin so as to minimise the sum of the squares of the distances between equivalenced C^{α} atoms in the predicted and crystal structure is shown in Figure 3.8. The superposition of this structure upon the C^{α} atoms of the crystal structure (shown with dotted lines) is presented in Figure 3.9. It is evident that the fold of I:97, the correct relative orientation of the individual units of secondary structure, approximates the myoglobin fold. Moreover, the phasing of residues along the helices, which is dictated by the location of the central residues, shows good agreement between the predicted and native structures. A comparison between the idealised parameters used in the construction of I:97 and the values found in the crystal structure is presented in Table 3.8. The errors in the approximation as measured by the r.m.s. deviation are surprisingly small except for the GH interaction (see Table 3.9).

Still, it is clear that there are problems with this hypothetical apomyoglobin. The C and D helices cannot be placed exactly since there are no docking sites above the cut-off on these short helices and we can only offer that they lie near B35 and E58. The docking angle for the GH pair, +19°, deviates significantly from -19° from the X-ray structure. No statement can be made about the coiled regions joining the α-helices. In spite of these problems, the observed and predicted structure and the relative

TABLE 3.8

Comparision of Actual Myoglobin Structure with Predicted Structure I:97

Helix Pairs	Length of Contact Normal (Å)		Acute Helix Axis Angle[a]		Full Helix Axis Angle[b]	
	Mb	I:97	Mb	I:97	Mb	I:97
BE	7.3	7.5	-76°	-80°	+104°	+100°
AH	8.5	8.5	-85°	-60°	+ 85°	+120°
BG	8.6	8.5	-55°	-60°	+125°	+120°
FH	8.7	8.5	-62°	-60°	- 62°	- 60°
GH	10.3	10.5	-19°	+19°	+161°	-161° (+199)

[a] Smallest torsional angle between helix axes

[b] Torsion angle between positive (N → C) directions of helix axes.

TABLE 3.9

The R.M.S. deviation (Å) between helix-helix pairings in the predicted and the crystal structures

I88 \ I97	A	B	E	F	G	H
A	-	2.7	3.1	6.5	3.7	<u>4.2</u>
B	2.7	-	<u>2.0</u>	2.5	<u>1.5</u>	3.1
E	3.1	<u>2.0</u>	-	3.1	2.6	2.2
F	13.7	4.2	<u>4.4</u>	-	3.7	<u>2.9</u>
G	3.7	<u>1.5</u>	2.6	.7	-	<u>12.0</u>
H	4.2	3.1	2.2	.8	<u>12.1</u>	-

<u>Legend</u>: For each possible helix pairing in I97 and I88, the R.M.S. deviation between the predicted and the crystal co-ordinates is reported. The helix-pairings used in the construction of I97 and I88 are underlined. Note that the difference between I97 and I88 is in the location of the F helix. F is docked to H in the former and to E in the latter.

positions of the contact normals agree reasonably well.

Perhaps the most pleasing aspect of the predicted structure is that small perturbations, possibly supplied by a subsequent higher level program, can be imagined which would push this structure closer to the X-ray coordinates.

(1) A weakly predicted site on D at residue 55 found below the cut-off could pair with another weak site, B30, in a class III interaction which would place the D helix very near its observed location without violating the short connection between D57 and E58 (see Figure 3.10).

(2) The very tight connection of the F helix with E and G would relax if H were rotated closer to the observed docking angle.

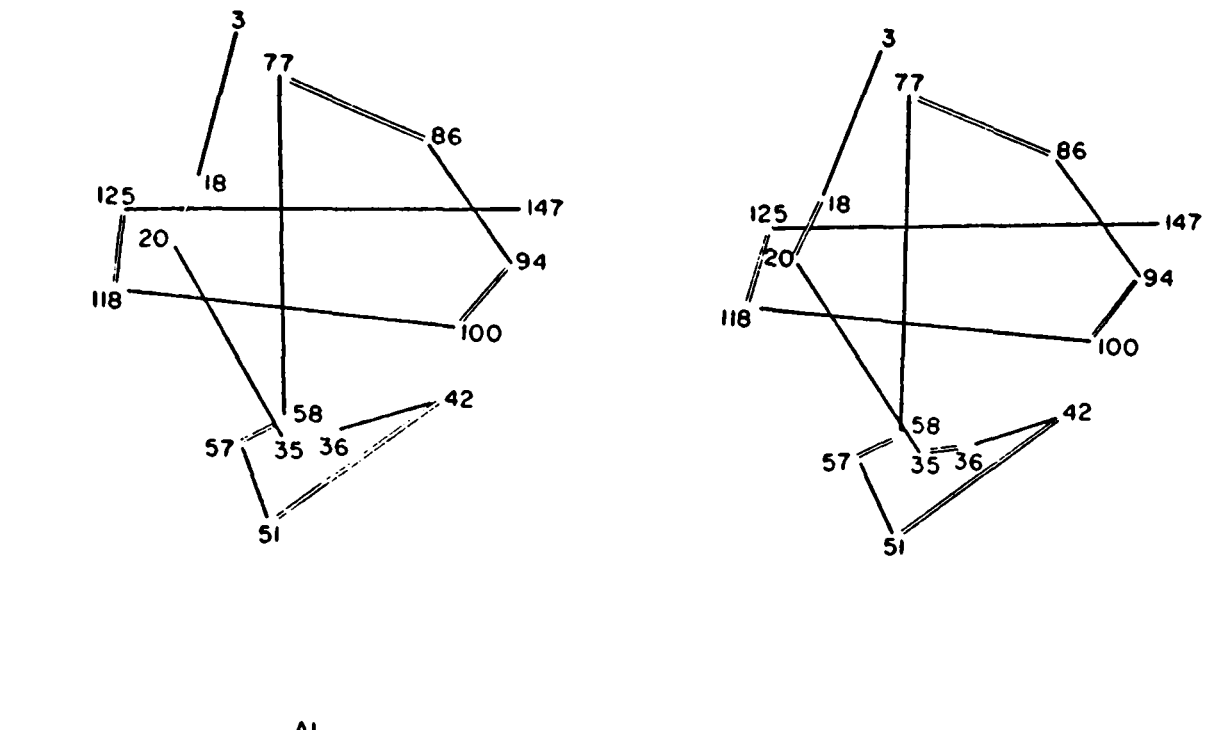
(3) The N terminus of the A helix is too close (6.93Å) to the N terminus of the F helix. Some flexibility in the docking angle of the AH pair would allow migration closer to the perpendicular, forcing A closer to its true position.

(4) Following the insertion of the heme group between the E and F helices, the C helix could easily be localised so as to surround the porphyrin ring.

Though an exact list of the N and C termini of the helices in myoglobin was used in this study, there is reason to believe that this is not necessary. The end points of the helices of I:97 were varied by one or two residues in either direction to see if any such manipulation would violate the end point restrictions. The only difficulties occurred when the A18-B20 connection was moved to A16-B18, A17-B18 or A18-B19 and when the G118-H125 connection was tried as G120-H123. For all cases, the maximum discrepancy in length was 1.7Å. Thus it might be possible for a conservative secondary structure prediction algorithm to provide the necessary data base for this program.

FIGURE 3.10

A Helical Axis Representation of Myoglobin with the D Helix Inserted



Stereo Diagram of the helical axes of the myoglobin structure with the inter-helical connecting links inserted and the structure I:97 with the D helix added according to the secondary site prediction. No predicted site is sufficiently strong (in the absence of the haem group) to allow the placement of the C helix. The latter would be constrained only by its endpoint connections to B and D.

1.9 Relative value of various constraints

Implicit in this approach to predict the structure of α-helical proteins is the sequential insertion of various restrictions on the allowed conformations. Success of any filter is obviously a function of the selectivity and ease of implementation. The effects of the constraints used in this study on myoglobin are presented in Figure 3.11. All structures to be discussed consider a C^{α} representation of the protein. Clearly, the number of distinct conformations available to an all atom representation of a protein is much larger than with the approximation used here.

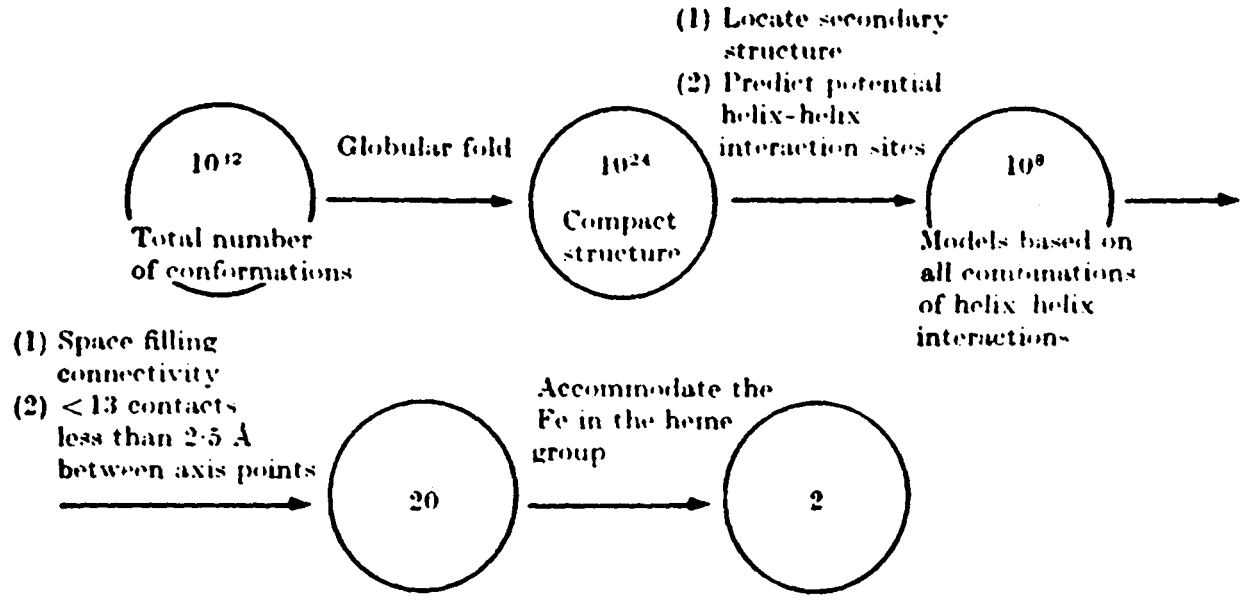
Work on self avoiding random walks of protein chains constrained to lie within a sphere has been used to quantify the number of compact globular folds of a polypeptide chain (Cohen & Sternberg, 1980b, Chapter 2). For myoglobin, approximately 10^{25} compact globular structures are possible. Levitt (1977) has shown that only 1 in 10^{8} structures generated by self avoiding random walks are compact. Thus, the total number of conformations is of the order of 10^{33}. Locating secondary structure and requiring helices to dock at predicted potential helix-helix interaction sites considerably limits the possible 10^{25} to a lattice of 10^{8} structures. Ensuring chain connectivity and that there are few close contacts limit the total number of reasonable structures to 20. Finally, requiring that these structures can accommodate the heme group lead to only two very similar structures.

1.10 Postlude - Myoglobin

One common strategy for the computational approach to protein folding employs a step-by-step procedure from the extended linear or random chain to the native compact structure(s). To minimise computational requirements, each step employs no more structural information than is necessary for the algorithm associated with that step. For overall efficiency each step should limit conformational space as far as possible. However, the

FIGURE 3.11

A Combinatorial Approach to the Prediction of the Structure of Myoglobin



Reduction of the folding problem by various restraints.

restrictions applied as a filtering mechanism must always be loose enough to include the real structure among the collection passed on to the next step. Regardless of the details of the algorithms actually employed, the usefulness of a given step will depend on the ratio of structures in to structures out. If this ratio cannot be kept high, the overall procedure may not converge fast enough to be useful in any practical sense. This problem is equivalent to the multiple minima difficulty of the alternate strategy of energy minimisation.

This approach assumes that the first step of the overall procedure, the prediction of elements of secondary structure, has already been carried out successfully. The algorithm attempts to produce a rough tertiary structure for the protein which would be suitable for further refinement, converging towards the true structure. This algorithm is restricted solely to helix-helix interactions and thus represents only a limited part of a full-scale approach to this second step. The problems encountered here, however, are reflected in the global procedures discussed in Chapters 4 & 5.

The first element of the procedure is the helix interaction site prediction suggested by Richmond (Richmond & Richards, 1978). This scheme is based solely on sequence data but has a stereochemical basis related to experimentally observed packing arrangements. The procedure is not all or none but ranks the possible sites roughly in order of the probable strength of the interaction. One can ensure inclusion of all actual sites by going far enough down the list, but the cost can be huge in increased computation later. A basic assumption in this study is that the approximate tertiary fold can be defined by considering only the strongest of all of the helix-helix interactions that may exist in the actual structure, and that these latter will be among the strongest that could be developed from any part of the given primary structure (see Ptitsyn & Rashin, 1975). How far down the ranked list one should to to get the proper balance of inclusiveness versus computation time will only be established by further experience with

other structures. The choice of cut-off made in this thesis was high enough to exclude any predicted sites on helices C and D. No statements about these helices can thus be made. Lowering the cut-off to include them would have included an unacceptably large number of sites on the other helices considering the present computational efficiency of the program.

A possible additional filter would be a compactness check. All small globular proteins and the domains of larger molecules have small axial ratios. The inertial ellipsoids calculated from X-ray data give maximum ratios of less than 3 to 1. From the known molecular weight of a given polypeptide, the maximum probable length of the folded structure can be calculated. This number would serve as a quick check on the overall shape of a growing structure and thus shorten the search by eliminating very extended conformations. If solution scattering data offered a measure of the radius of gyration of the protein, this could be used as a specific filter for eliminating misshapen structure.

If the output structures at this stage are to be useful, they must serve as suitable input for the next more detailed step of the overall folding program. The helix specifications would permit all atoms of the helical side chains to be positioned if a set of trial torsional angles, χ_i, were assumed. The connecting string regions are unspecified and variably restrained and would have to be built in by standard model building programs such as those of Diamond (1966) or Hermans & Ferro (1971). Energy minimisation might then be attempted. It seems unlikely that this large jump towards final refinement would be warranted at this stage. More structural information should be included, but some idealisation of the side chains perhaps as suggested by Levitt & Warshel (1976) would be more appropriate. General packing and hydrogen bond considerations would then be allowed to adjust the rough structure closer to a chemically realistic one. After that point full structural detail could be included.

Whether complete or idealised structures are used, packing adjust-

ment procedures generally run into difficulty when the objects are grossly misplaced. Errors can frequently be corrected if the objects move through each other, but there is usually no simple path by which this can be accomplished. Thus a tertiary prediction scheme should attempt to anticipate this difficulty where possible. A good example is provided by the G and H helices of myoglobin. Surveys of actual interactions where the two helices are nearly parallel show that the preferred helix axis angle is generally about +20°. However in myoglobin the angle is in fact -20°. To pass from one to the other would require the facing residues to collide as the axis angle went through 0°. In this case it would seem preferable to set the assumed helix angle to 0° even though no actual structure has such an angle. The next level program would recognise the overlap but would have the option of easily sampling both + and - angle shifts in the search for a more acceptable conformation. Other examples will no doubt turn up a more detailed study.

Havel et al. (1979) have shown that a small number of specific distance constraints applied to a polypeptide chain are not valuable in reducing the number of allowed conformations. This study on myoglobin demonstrates that given a small number of candidates for a protein structure, certain distance constraints can be very useful in limiting the number of reasonable folds. It should be noted that distance information accurate to within a few Ångstrom units was all that was necessary given the limits of our model.

Most approaches to the protein folding problem suggest that sequence information is all that is required to determine uniquely the structure of a protein. While an accurate consideration of all of the forces involved in the system would no doubt lead to a correct model, this goal has yet to be realised in current theoretical studies. It is clear from this work that biochemical information elucidating the distance between sequentially distant parts of the polypeptide chain can aid in the solution of this problem.

1.11 An Application of the α-helix Combinatorial Procedure to Tobacco Mosaic Virus Coat Protein

Several all helical proteins can be classified as a four-helix bundle with four helices lying approximately parallel to one another forming the long edges of a rectangular parallelipiped. These include hemerythrin (Hendrick et al., 1976), cytochrome b_{562} (Matthews et al., 1979) and the tobacco mosaic virus (TMV) coat protein (Bloomer et al., 1979; Stubbs et al., 1978). Following the site prediction rules of Richmond & Richards (1978) together with the steric arguments for site overlap (see section 3.3), a list of 8 possible helix-helix interaction sites were found (see Table 3.10). Although other lists could have been chosen (e.g. residue 124 as a type II site), no hydrophobic sites were ignored and so these 8 sites form a self-consistent set of all hydrophobic patches in the molecules (see Table 3.11).

These sites formed the input for the FOLD-BUILD-REALSPACE program sequence. 62 possible structures were generated, one of which resembles the TMV coat protein (see Figure 3.12). The r.m.s. deviation between these two structures is 4.45A. The problem encountered in the GH cross of myoglobin reappears in one of the TMV type III interactions. This is the major source of errors in the predicted coordinates.

Crippen (1979) has assembled a variety of experimental constraints on the TMV structure. We have not pursued this problem as a more general analysis of all helix-helix interactions seems necessary to explain the structures of other all-helical proteins. Some preliminary work toward this goal is discussed in the two sections which follow.

1.12 A General Survey of Helix-Helix Interactions

The site prediction algorithm developed by Richmond & Richards (1978) was based upon a subset of all known helix-helix interactions. Thus, the

TABLE 3.10

Secondary Structure Assignments and Potential Helix-Helix Interaction Sites in Tobacco Mosaic Virus Coat Protein.

Helices		Sites	
N-terminus	C-terminus	Type II	Type III
20	32	24,27	24,27
38	50	44	44
73	90	80,82,86	83
114	135	121,124, 125,129	121,124 125

TABLE 3.11

Predicted Sites and Potential Contact Area Changes for Helical Residues in Tobacco Mosaic Virus Coat Protein

Type II		Type III	
Cut-off = 99Å²		Cut-off = 158Å²	
Site	Å²	Site	Å²
24[c]	156	24[a]	231
27[c]	150	27[a]	202
44[c]	117	44[a]	158
80[a]	169	83[a]	225
82[d]	126	121[b]	222
86[b]	104	124[a]	164
121[c]	135	125[a]	224
124[c]	150		
125[c]	170		
129[a]	185		

[a] This site is used

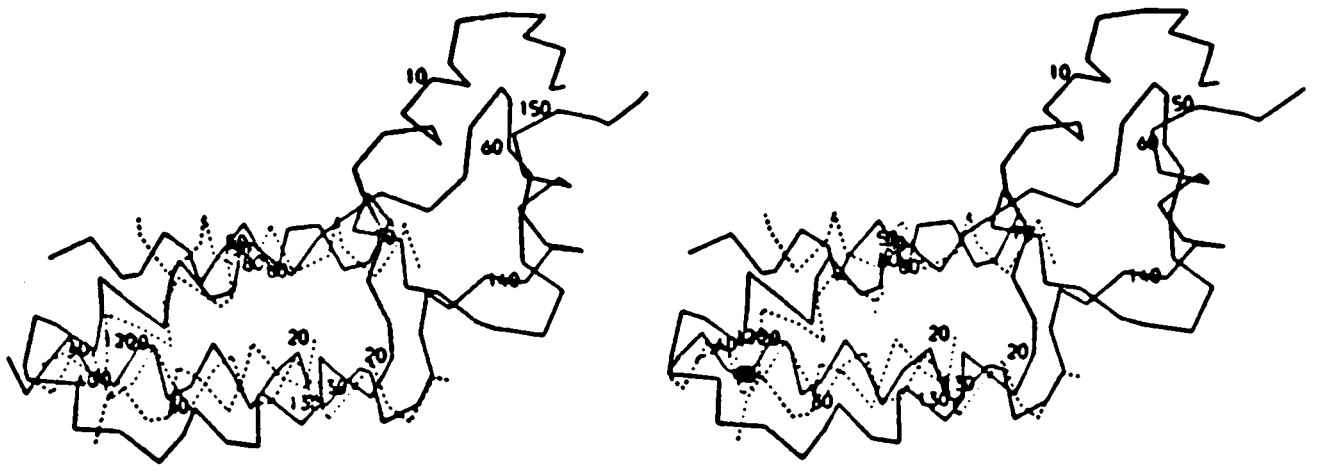
[b] i+3 or i+4 from a Type II site - not used

[c] Used as a Type III site

[d] i+4 from a blocked Type II site.

FIGURE 3.12

Tobacco Mosaic Virus Coat Protein - Predicted and Crystal Structures



This stereo diagram shows the agreement between the predicted coordinates for the C^{α} atoms in the helical segments of the Tobacco Mosaic Virus Coat Protein overlaid on the crystallographically determined coordinates (Bloomer et al., 1978).

generality of this approach as part of a larger structure prediction algorithm was in doubt. To overcome this objection, 49 pairs of interacting helices from the 22 proteins listed in Table 3.12 were analysed. The analytical method followed is essentially that of Richmond & Richards (1978).

1.12.1 Data Base

α-Helical residues in the 22 proteins were selected by a combination of two methods. Sequential sets of four or more residues were found with backbone dihedral angles such that $-130° < \phi < -10°$ and $-90° < \psi < -10°$ with ϕ of the N-terminal residue and ψ of the C-terminal residue ignored. These tentative assignments were compared to the crystallographers' assignments. Although minor discrepancies existed at the termini, a set of consistent assignments was established.

From three consecutive C_α atoms, a point that would lie on the axis of an ideal α-helix was located. The least squares line through these points defines the helix axis. If the deviation of a given axis point from the axis was greater than 1Å, that point was discarded and a new axis was computed. Helix axis points were then redefined as the projection of the C_β atom (C_α for glycine) onto the helix axis.

The contact normal was computed as the shortest distance between two helical axes. If this length exceeded 12Å, the pair was discarded. The non-polar accessible contact area (NPACA) was computed for each pair of helices and for then for each pair in isolation. If the difference between these two values was less than 100Å²*, the pair was discarded. Thus the 49 pairs which remained were involved in "strong" helix-helix interactions.

Central residues were defined as those residues which were near the contact normal and had a small skew angle. Let I be the point of inter-

* The accessible area data quoted in this section is qualitatively correct, but the values are approximately 5% less than their true value. This was due to an error in the cubing algorithm of the accessibility program supplied by Dr. T.J. Richmond. Due to the enormous number of calculations required, this data hs yet to be recomputed.

TABLE 3.12

Proteins included in a General Survey of Helix-Helix Interactions

Protein	Reference
Adenyl kinase	Schulz et al. (1974a)
Alcohol dehydrogenase	Eklund et al. (1976)
Cytochrome b_5	Mathews et al. (1971)
Carbonic anhydrase B	Liljas et al. (1972)
Carboxypeptidase A	Hartsuck & Lipscomb (1971)
Parvalbumin	Kretsinger & Knockolds (1973)
Cytochrome C (oxidised)	Dickerson et al. (1971)
Flavodoxin	Burnett et al. (1974)
Glyceraldehyde 3-phosphate dehydrogenase	Moras et al. (1975)
Lactate dehydrogenase	Holbrook et al. (1975)
Hemoglobin (lamprey)	Hendrickson et al. (1973)
Lysozyme (hen egg white)	Imoto et al. (1972)
Lysozyme (T4-phage)	Matthews & Remington (1974)
Myoglobin	Watson (1969)
Hemoglobin α-chain	Perutz et al. (1968)
Hemoglobin β-chain	Perutz et al. (1968)
Papain	Drenth et al. (1971)
Ribonuclease S	Richards & Wyckoff (1971)
Subtilisin BPN'	Kraut et al. (1971)
Staphlococyl nuclease	Arnone et al. (1971)
Triose phosphate isomerase	Banner et al. (1975)
Thermolysin	Colman et al. (1972)

section of a contact normal with a helix axis and J a helix axis point. Then if P is the distance between I and J, θ is the helix axis angle, and ϕ is the skew angle (see Figure 3.13), then the central residue is often that residue with a NPACA change >10Å² which minimises the value of S in Equation (3.1):

$$S = \alpha^2 P \sin\theta + |\phi| \csc\theta \qquad (3.1)$$

where $\alpha = (\pi/2)/(5.6)$*.

1.12.2 Analysis of Helix-Helix Interactions

A plot of the helix axis angles as a function of interaxial separation reveals two interesting features: there are four classes of interaction angles; and the distribution of Type III crosses is bimodal (see Figure 3.14). Class II was split in IIa and IIb because the group which clustered around -60° displayed i±3 packing of ridges into grooves while IIb had i±4 packing. This was evident when the per residue NPACA changes were examined in detail.

The skew angles used by Cohen et al. (1979) overestimated the average Type III skew angle and underestimated the average Type I skew angle. This angle is approximately ±25° in all helix-helix interactions (see Table 3.13). The variation in contact normal length is larger than had been expected. Fortunately, if the contact normal length is broken into the contributions of the two central residues, the total can be approximated by the average contribution of each residue type. This half contact normal length is roughly proportional to the volume of the central residues (see Figure 3.15).

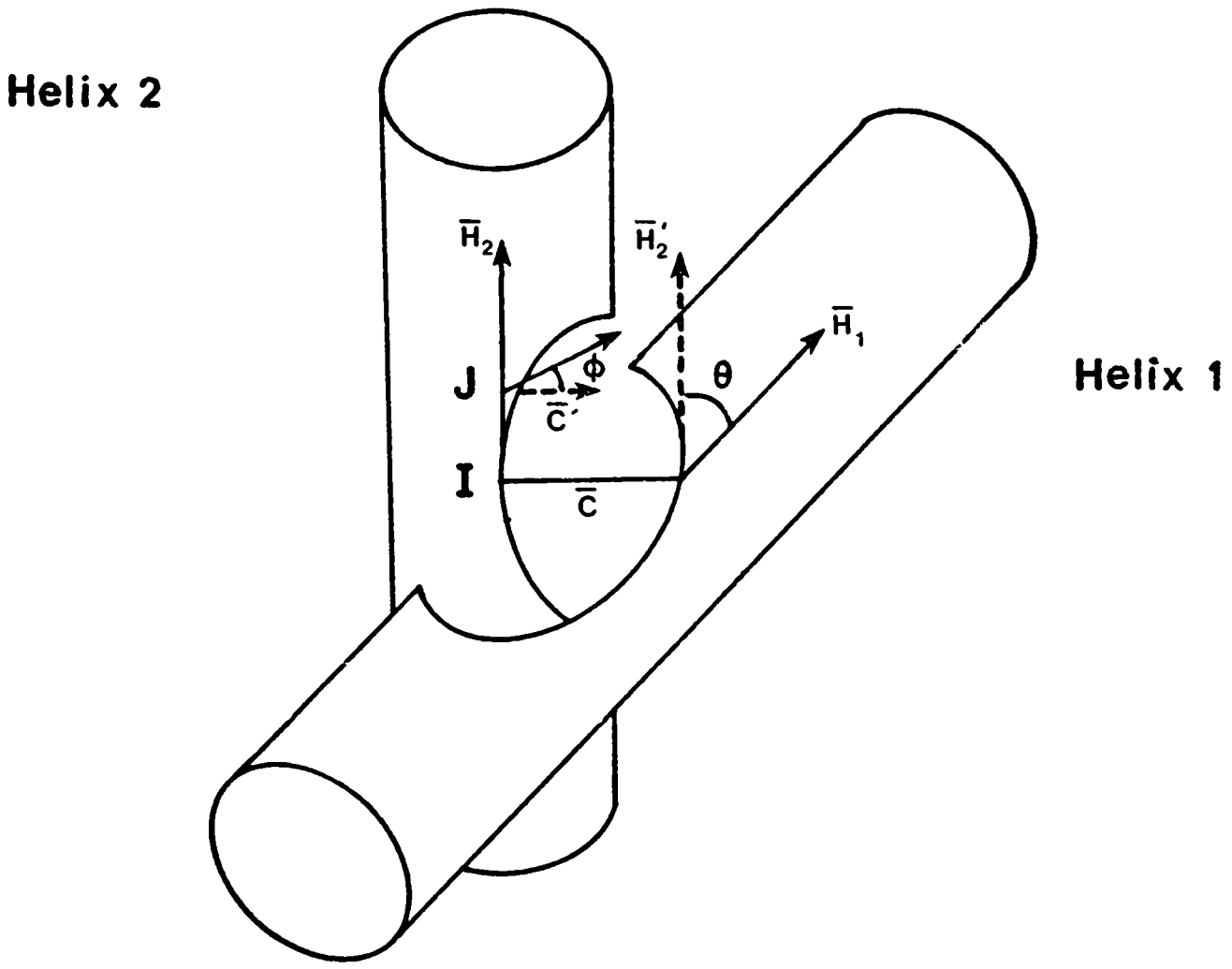
1.12.3 The Prediction of Helix-Helix Interaction Sites

The goal of this analysis was to aid in the development of a general

* 5.6Å is the radius of the cylinder of equal penetration for a pair of interacting helices (see Richmond & Richards, 1978).

FIGURE 3.13

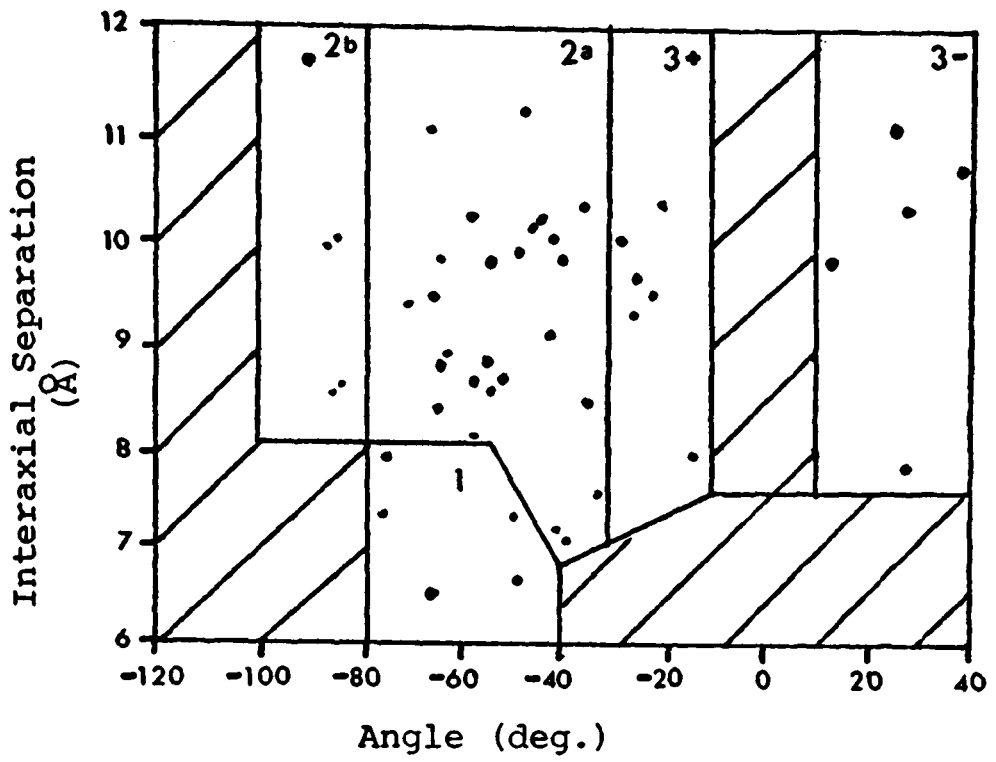
Schematic Helix-Helix Interaction



This is a schematic cross-section of a helix-helix interaction. θ is the dihedral interaxial angle defined by the angle between the translation of the helix axis vector H_2' and the helix axis vector H_1, ϕ is the skew angle for the central residue J on helix 2 defined by the vector joining the carbon of residue J to the axis and the translation of the contact normal C' and C, the line joining the two helices mutually perpendicular to H_1 and H_2 is the contact normal. The distance between I and J is P and θ, ϕ and P determine S in equation 3.1.

FIGURE 3.14

A Survey of Helix-Helix Interactions



The helix axis angle is plotted against interaxial separation for 50 helix-helix interactions. These interactions are grouped into 5 categories and the disallowed regions of the plot are marked. Note that no interactions are seen between -10° and +10° or between 40° and 60° (-120° to -140°).

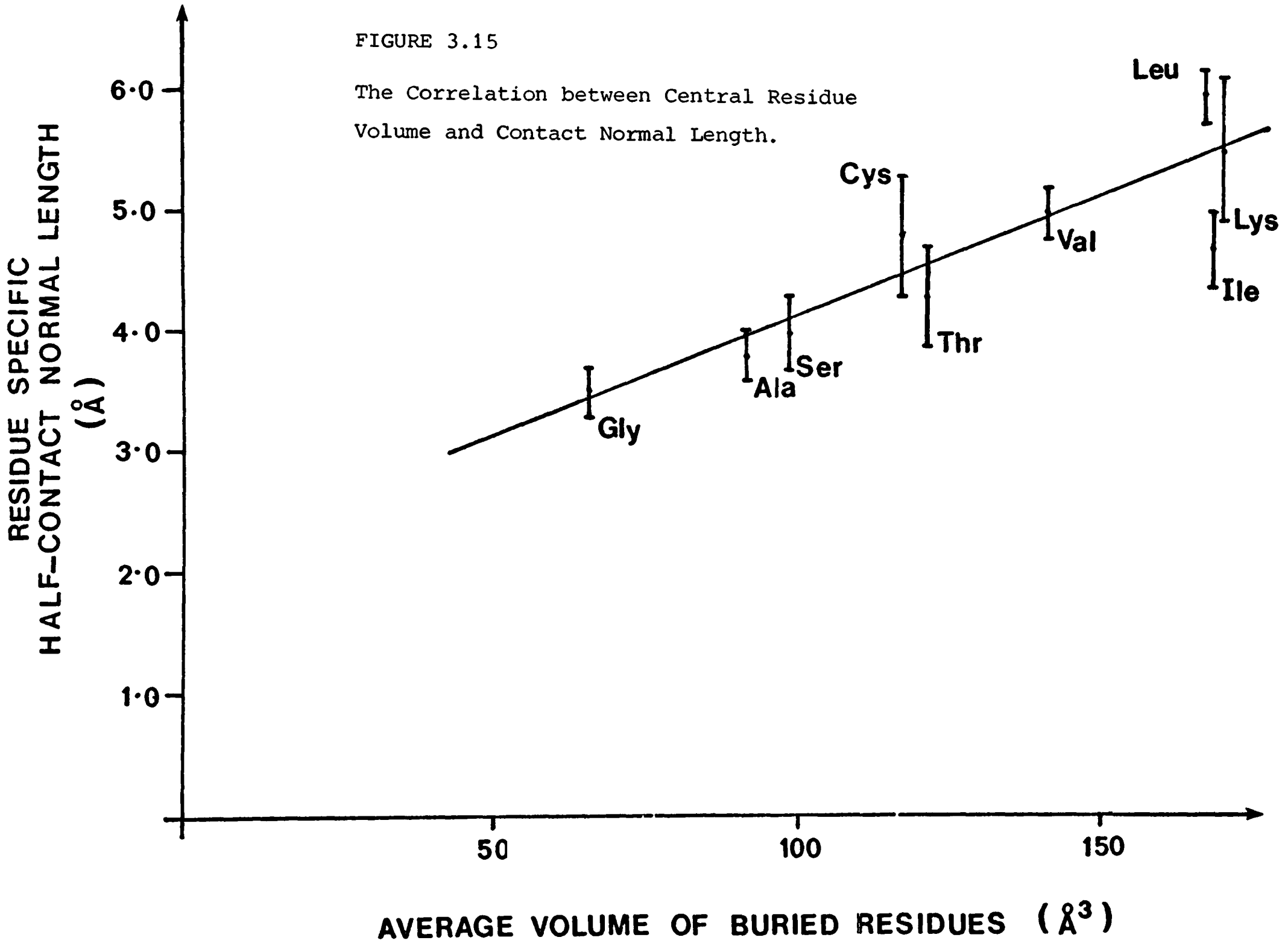
TABLE 3.13

Mean Geometry of Helix-Helix Classes

Class	Number	Interhelical Angle Ω°	Skew Angle $\lvert\theta°\rvert$	Contact Normal Length (Å)
I	6	-63±11	34±9	7.0 ± 0.6
IIa	26	-52±11	20±12	9.1 ± 1.2
IIb	5	-91± 5	33±20	9.8 ± 1.2
III	12	+2±26	22±25	9.8 ± 1.1
III^+	6	26± 5	25±18	10.1 ± 1.1
III^-	6	-23± 8	19±10	9.4 ± 0.8

FIGURE 3.15

The Correlation between Central Residue Volume and Contact Normal Length.



scheme for predicting potential helix-helix interaction sites. The over-prediction of sites is tolerable, but if sites that are necessary to construct a native-like "predicted" structure are overlooked, the algorithm will be ineffective.

As with the scheme of Richmond & Richards (1978), a residue will be considered a potential helix-helix interaction site for a particular interaction class if it is an allowed central residue and the constellation it is central to has a potential NPACA change above a certain cut-off. This figure is the sum of the potential NPACA contributions of the neighbouring residues in a constellation of residues appropriate for the interaction class (see Table 3.14). These constellations and allowed central residues are minor modifications to the Richmond & Richards (1978) algorithm.

Because these necessary additions brought a three-fold increase in the number of sites predicted, exclusion rules were defined which would limit the number of possible pairings FOLD and BUILD would have to sample. Three rules were developed to eliminate the pairing of certain central residues based on a set of possible pairings which were not observed and whose absence could be explained on a stereochemical basis.

(1) In Type I interactions, Ala-Ala pairings are disallowed. This follows from volume restrictions for this tight cross.

(2) Ser-Ser, Ser-Thr and Thr-Thr are disallowed. Since the region between two packed α-helices is largely hydrophobic, two hydrophilic residues cannot be tolerated.

(3) Leu-Leu, Leu-Ile, Leu-Lys, Ile-Ile, Ile-Lys and Lys-Lys are disallowed. This is due to the large volume of these residues and the lack of space between a pair of interacting helices to accommodate them simultaneously.

Obviously, each restriction reduces the problem of predicting the structure of all helical proteins further and further. The development of additional exclusion rules is in progress.

TABLE 3.14

Constellations of Residues Involved in Helix-Helix Interactions

Class	Constellation	Allowed Central Residues
I	-4 -3 0 +1 +3 +4	Gly, Ala
	-4 -3 -1 0 +3 +4	
IIa	-4 -3 0 +1 +4	Ala, Ser, Thr, Pro, Ile, Cys, Val, Leu, Met + Lys at the ends
	-4 -1 0 +3 +4	
IIb	-3 0 +1 +3 +4	
	-4 -3 -1 0 +3	
III	-7 -4 -3 0 +3 +4 +7 +8	Gly, Ala, Ser, Thr, Pro, Ile, Cys, Val, Leu, Met, + Lys at the ends
	-7 -4 -1 0 +3 +4 +7 +8	
	-8 -7 -4 -3 0 +1 +4 +7	
	-8 -7 -4 -3 0 +3 +4 +7	

A table of the constellations used in computing the potential area change for a hydrophobic patch for the four different classes, together with the allowed central residues. The NPACA contribution for the amino acids are Ala (18), Arg (10), Asn (2), Asp (2), Cys (24), Gln (5), Glu (6), Gly (11), His (18), Ile (33), Leu (33), Lys (18), Met (34), Phe (37), Pro (10), Ser (3), Thr (12), Trp (49), Tyr (27), Val (33).

1.13 Studies of Other All-Helical Proteins

Modifications to the algorithm of Richmond & Richards (1978) were suggested by this more general survey of helix-helix interactions. Could this altered algorithm be of some help in predicting the structure of other all-helical proteins. The salient results of this survey are:

(1) Four helix-helix interaction classes - Class II is divided into pairs with i±3, i±3 packing and i±4, i±4 packing. This produces angles of -60° and -82° respectively.

(2) The skew angles for all classes are ±25°.

(3) The Type III interaction angle can be ±20° but will be fixed at 0° for construction purposes - the distribution of Type III angles is bimodal with a trough about 0°.

(4) The contact normal length is taken as the sum of two residue-dependent half distances (see Figure 3.15).

These ideas were incorporated into BUILD with the following two filters:

(1) The distance between the helix axis points at the carbonyl end of one helix and the amino end of the next helix must be less than (3m+3)Å where m residues connect the two helices; and

(2) There are less than 50 contacts of <6.5Å between helix axis points on different helices. The value of 6.5Å was chosen as it is the closest interaxial separation observed in known protein structures (the B25-E65 separation in hemoglobin-β).

Since the number of potential interaction sites produced by the modified site prediction algorithm was significantly larger than that produced by the original algorithm, a complete search of all possible structures was not possible. Clearly, some method for determining which of the predicted sites cannot pair will be necessary to reduce the computer time requirement. This problem awaits further analysis. An easier, though equally vital question is: can other all-helical proteins be built from a list of the correct helix-helix interaction sites subject to the idealised

geometry previously described? One must demonstrate the existence of an answer to the general combinatorial problem before questions of this answer's uniqueness can be raised.

With the list of site pairs shown in Table 3.15, it was possible to construct C_α representations of the helical regions of thermolysin, myoglobin, T4-phase lysozyme, lamprey hemoglobin, and the α and β chains of hemoglobin. The r.m.s. deviations of these best "predicted" structures from the corresponding crystal structure averaged 3.3Å. This represents a slight improvement over the 4.25Å deviation seen in original myoglobin work. Pertinent statistics for these structures are in Table 3.16 and stereo diagrams of some of the "predicted" structures superimposed on the crystal structure can be seen in Figure 3.16.

Clearly, splitting Type II interactions into IIa and IIb and setting the ideal Type III angle to 0° improved the quality of the "predicted" structures. Moreover, the modified site prediction algorithm together with this simplified model of helix-helix interactions is accurate enough to allow the construction of five other all-helical proteins or protein domains. Although this approach is far from complete, it remains quite promising.

1.14 Detailed Energy Calculations on specific Helix-Helix Interactions

In an attempt to understand why specific helices pair, detailed energy calculations were performed (see section I.6.1) on the B-E helix interaction in myoglobin. This is a Type I cross with a pair of glycines as the central residues. The extreme interdigitation of residues in this region of myoglobin made this the most suitable candidate to study.

In this section, the energy of a polypeptide chain was calculated, in the gas phase, from the equation:

TABLE 3.15

Secondary Structure Assignments and Predicted Central Residues used in the Construction of Six Proteins

	Helices		Sites	
	N terminus	C terminus	Type	Residue Number
Thermolysin	67	87	I	141, 173
(α-domain)	137	151	IIa	240, 270, 291, 309
	160	180	IIb	79, 176
	233	246	III	175, 264, 292, 306
	260	274		
	281	296		
	301	313		
Myoglobin	3	18	I	25, 65
	20	35	IIa	28, 90, 110, 142
	58	77	IIb	10, 134
	86	94	III	108, 135
	100	118		
	123	147		
T4 phage lysozyme	2	11	IIa	7, 71, 74, 103
	61	78	III	98, 130, 149, 150
	93	105		
	129	134		
	143	155		
Lamprey Hemoglobin	12	28	I	34, 74
	30	44	II	101, 123, 137, 147
	68	88	III	19, 37, 85, 121
	92	106		
	112	127		
	132	148		

Cont.

TABLE 3.15, cont.

	Helices		Sites	
	N terminus	C terminus	Type	Residue Number
Hemoglobin α	3	18	I	25, 59
	20	35	IIa	28, 84, 104, 136
	52	71	IIb	10, 124
	80	88	III	102, 129
	93	105		
	118	135		
Hemoglobin β	4	18	I	24, 64
	20	34	IIa	27, 89, 109, 141
	57	76	IIb	11, 129
	85	93	III	107, 134
	99	117		
	123	144		

TABLE 3.16

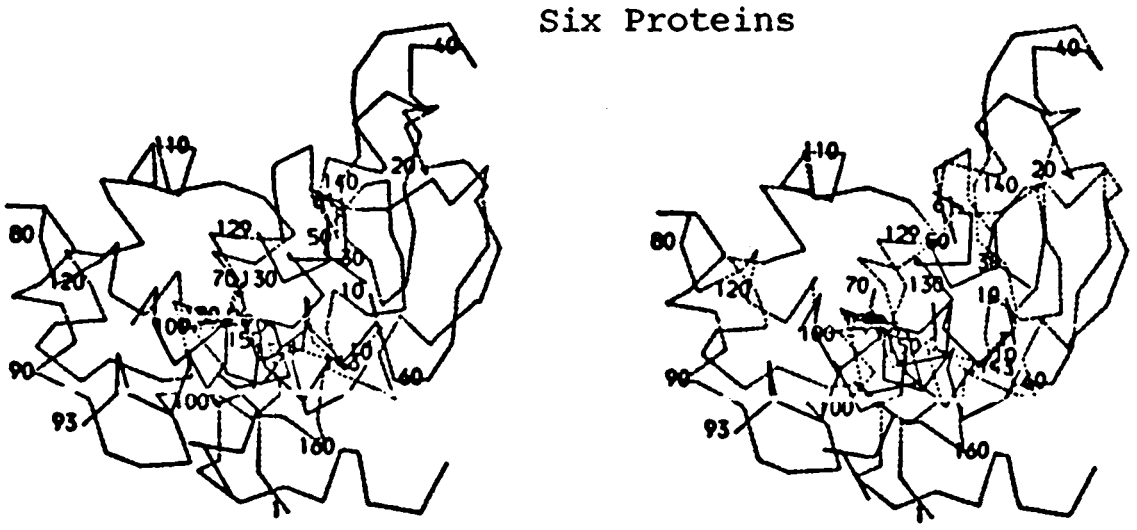
Statistics on the Best Predicted Structures for Six Proteins

	Number of Helices Placed	Number of Helical Residues	R.m.s. deviation from native structure	Bad contacts (a)	Close contacts (b)
Thermolysin	7	115	3.85	21	2
Myoglobin	6	105	2.99	34	7
T4-phage lysozyme	5	70	3.46	44	6
Lamprey hemoglobin	6	101	3.84	25	1
Hemoglobin β-chain	6	101	2.87	45	8
Hemoglobin α-chain	6	102	2.82	39	7

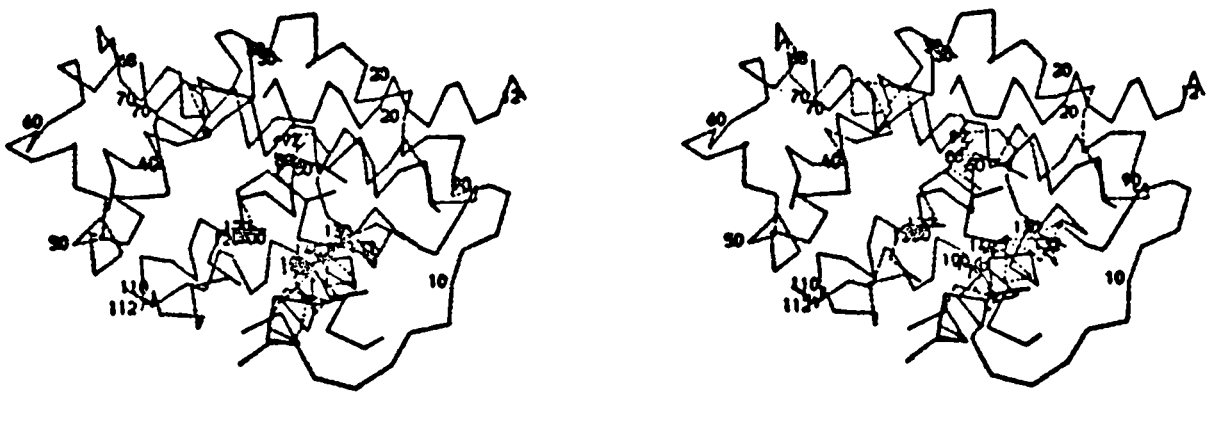
(a) "Bad contacts" are scored when the distance between helix axis point, the projection of the C^{α} atoms onto the helical axis, is <6.5Å.

(b) "Close contacts" are scored when C^{α} atoms on adjacent helices are within 3.6Å, 90% of the sum of their van der Waals radii, 4Å.

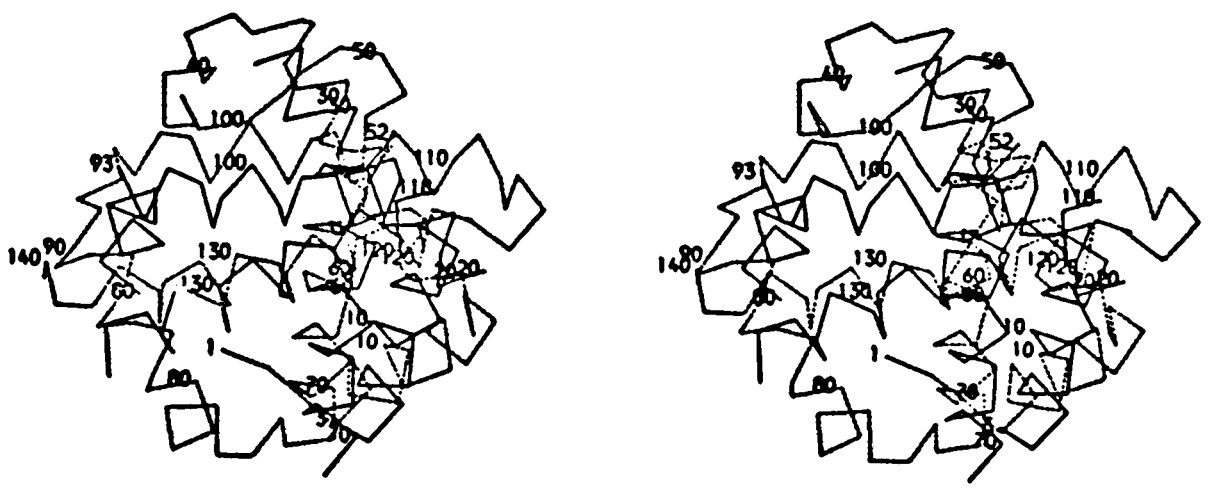
FIGURE 3.16. Stereo Diagrams of the "Predicted" Structures for Six Proteins



T4-phage lysozyme

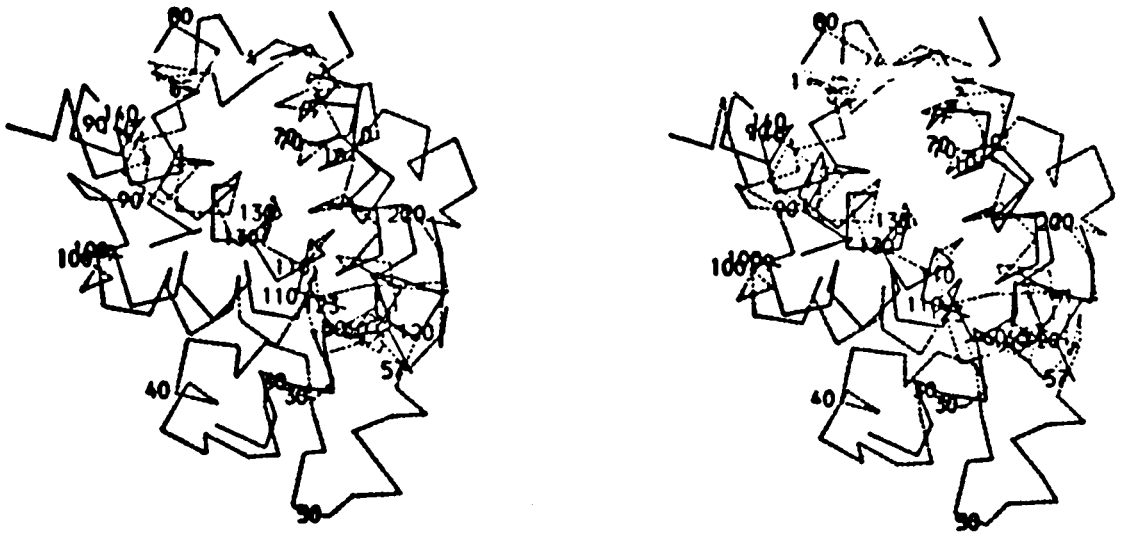


Lamprey Hemoglobin



Hemoglobin α-chain

FIGURE 3.16, continued.



Hemoglobin β-chain

Myoglobin

Thermolysin

$$E(\bar{X}_1,\bar{X}_2.....\bar{X}_m) = \sum_i K_i^{len}(\ell_i - \bar{\ell}_i)^2 \quad \text{(bond length)}$$

$$+ \sum_i K_i^{ang}(\theta_i - \bar{\theta}_i)^2 \quad \text{(bond angle)}$$

$$+ \sum_i K_i^{tor}(1 - \cos(3\tau_i)) \quad \text{(dihedral angle)}$$

$$+ \sum_i \sum_j -A_{ij}/r_{ij}^9 + B_{ij}/r_{ij}^{12} \quad \text{(van der Waals)}$$

$$+ \sum_i \sum_j g_i g_j/\varepsilon r_{ij} \quad \text{(electrostatic)} \quad (2)$$

Vicinal (1-4) van der Waals and electrostatic energies are reduced to reflect the conclusions of spectroscopic studies (for a general discussion of potential functions see Hagler et al., 1979). The parameters in this equation are: $\bar{X}_1,\bar{X}_2,.....\bar{X}_m$, the atomic coordinates; K_i^{len}, K_i^{ang} and K_i^{tor}, bond length, bond angle and torsional force constants; $\bar{\ell}_i$ and $\bar{\theta}_i$, optimal bond lengths and angles; ℓ_i, θ_i and τ_i, bond lengths, bond angles and torsion angles in the model; A_{ij} and B_{ij}, van der Waals attraction and repulsion terms; g_i and g_j, partial electronic charges; and r_{ij}, the distance between atoms i and j. Specific values for these parameters can be obtained from Hagler et al. (1979).

The energy of this isolated pair of helices was calculated from the crystallographic assignments of atomic position. Hydrogen atoms were placed on proton donors involved in hydrogen bonds and near the putative proton acceptor. The energy of the isolated B-E helices was calculated to be 996kcal in the gas phase with no attempt to model the hydrophobic effect beyond the normal van der Waals dispersion terms. All atom positions in both helices were allowed to vary. Upon minimisation, this energy was reduced to 98.9kcal (see Table 3.17). The bond angle and torsional terms dominate the energy of the native structure while van der Waals and electrostatic terms dominate the energy of the minimised structure. This calculation took seven hours in a PDP 11/70.

The minimised structure was perturbed to study the shape of

TABLE 3.17

Gas Phase Energy of the B-E Helix-Helix Interaction

Energy Terms	CYCLE Initial	50th	Final (875th)
Bond length	25.8	11.0	9.68
Bond angle	695	15.3	15.0
Dihedral angle	746	57.1	54.2
Van der Waals	-167	-176	-214
Electrostatic	-930	-918	-951
H-bond*	0	0	0
1-4 van der Waals	113	53.2	53.2
1-4 Electrostatic	1180	1130	1130
TOTAL	996	175	98.9

* The effect of the hydrogen bond was not specifically modelled as it is accurately accounted for in van der Waals and electrostatic terms (see Hagler et al., 1979).

the energy surface about this minima. Two perturbations were applied:

(1) Increase the contact normal length by 2Å and rotate the helix axis angle by 5°.

(2) Increase the contact normal length by 5Å.

The resulting structures were then energy minimised. All structures were compared to the minimised native structure and the r.m.s. deviation for all atoms was evaluated. The results are presented schematically in Figure 3.17. Clearly, relaxations of the perturbed structure did not produce conformations significantly closer to the energy minimised native structure. This result is not surprising as Hagler et al. (1979) have shown that the best potential functions available can reproduce the crystal structures of small molecules only within 0.3Å. Thus, a system with 400 or more atoms is too large for current techniques.

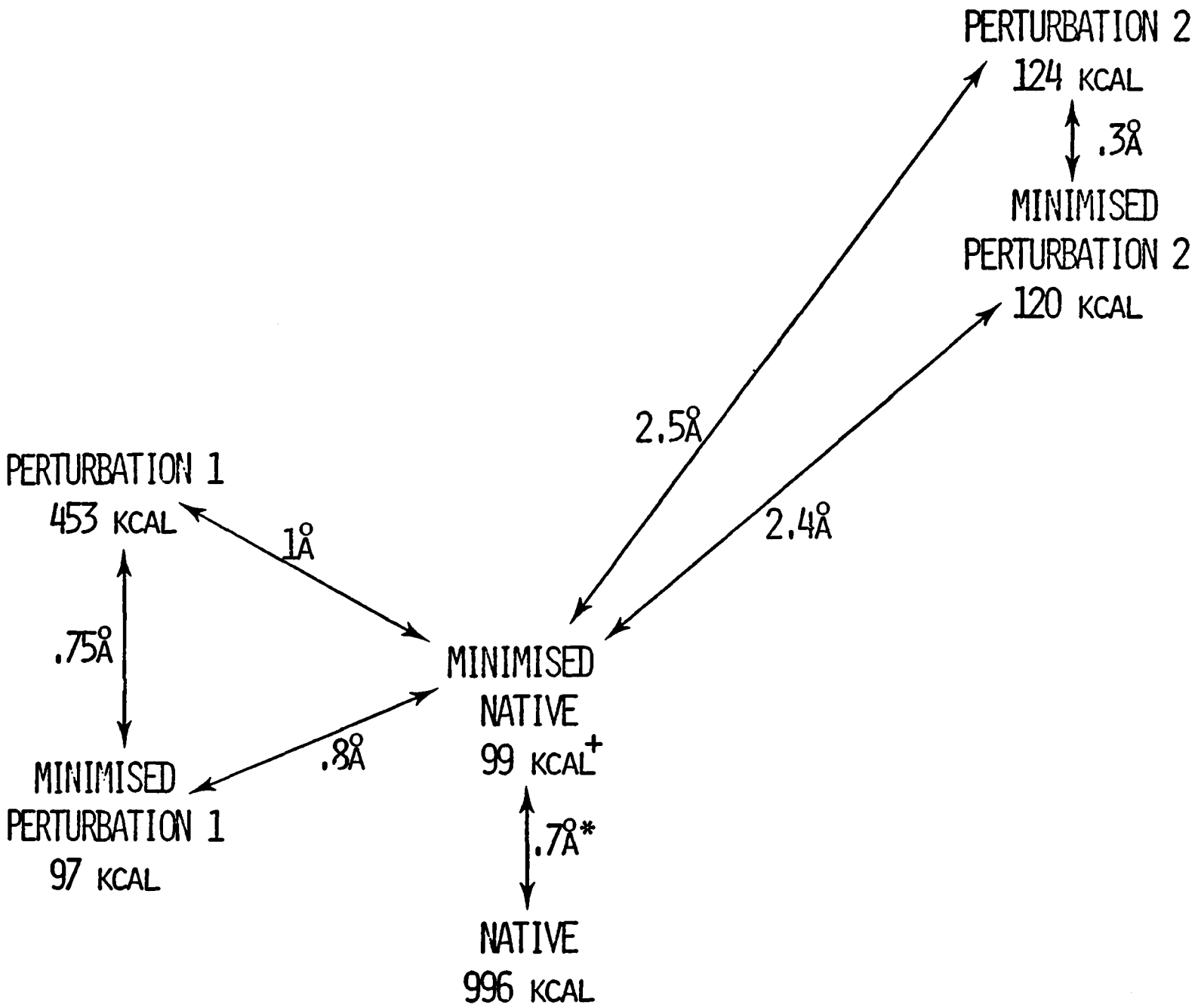
2. A Functional Role for Helix-Helix Interactions

In the preceding sections of this Chapter, a pair of interacting helices was classified as a rigid unit of protein substructure. This unit does, however, have three natural internal degrees of freedom, the inter-axial angle θ, and a skew angle for each helix ϕ_1 and ϕ_2. This section discusses a possible functional role for helix-helix interactions.

Proteins are not static units. Fischer's (1894) "lock and key" model for enzyme-substrate interactions led to the "induced fit" concept; an enzyme alters its conformation upon substrate binding to promote catalysis. This must be a low frequency mode of motion. Molecular dynamics calculations (e.g. Karplus & McCammon, 1979) or the analysis of crystallographic temperature factors (e.g. Sternberg et al., 1979) which have quantified many high frequency vibrations are at present unable to study gross conformational changes.

FIGURE 3.17.

ENERGY MINIMIZATION CALCULATIONS ON THE B-E CROSS IN MYOGLOBIN



PERTURBATION 1 - INCREASE CONTACT NORMAL LENGTH BY 2Å AND ROTATE HELIX AXIS ANGLE BY 5°

PERTURBATION 2 - INCREASE CONTACT NORMAL LENGTH BY 5Å

+ GAS PHASE ENERGY

* R.M.S. DEVIATION FOR ALL ATOMS

Experimental evidence for gross conformational changes following substrate binding is available. Solution studies of hexokinase (McDonald et al., 1978) have shown that the radius of gyration of the protein decreases by 1.95 ± 0.24Å when glucose and MgATP are bound. Native hexokinase and hexokinase + substrate crystallise in different forms. Following the determination of these two crystal structures, Bennett & Steitz (1978) showed that these two forms could be related by rotating one lobe towards the other about an instantaneous hinge. Anderson et al. (1979) have suggested that hinge-bending is a general feature of kinase activity.

Pickover et al. (1979) have shown that the structure of phosphoglycerate kinase (PGK) also changes when the substrates, MgATP and 3-phosphoglycerate, are bound. This results in a decrease in the radius of gyration for the ternary complex of 1.09 ± 0.34Å. They suggest that this change could be accounted for by a 9° to 12° rotation about the appropriate instantaneous hinge axis. Banks et al. (1978) have determined the structure of PGK but the ternary complex crystallises in a different form. The determination of this second crystal structure is in progress (D.Rice, personal communication).

PGK is a bilobal structure with MgATP bound to the C-terminal domain and two helices linking the lobes. Banks et al. (1973) have hypothesized that a cluster of charged residues (His 62, Arg 122, Glu 128, His 172) in the N-terminal domain are likely to constitute the rest of the active site. However, the distances between the β phosphate bound to PGK in the crystal structure and these target residues is approximately 20Å (see Table 3.18). If these residues are to bc part of a bilobal active site, a mechanism must exist to bring them together.

I propose that the contact normal joining the two helices which link the two lobes of the molecule forms an instantaneous axis about which the two lobes could hinge. Moreover, slight adjustments in the relative position of the two lobes could follow from changes in the dihedral helix

TABLE 3.18

Hinge Bending Calculations on Phosphoglycerate Kinase

Angular Motions θ= Change in inter-axial angle ϕ= Change in skew angle	Target Distances (Å)				Sum of Target Distances (Å)
	ATP β-phosphate to 62	ATP β-phosphate to 128	ATP β-phosphate to 172	ATP β-phosphate to Arg 122	
θ=0°, $\phi_1=\phi_2=0°$	20.35	22.32	20.40	18.69	81.77
θ=23°, $\phi_1=\phi_2=0°$	14.05	17.40	15.51	12.72	59.69
θ=29°, $\phi_1=\phi_2=0°$	12.40	16.29	14.29	11.28	54.28
θ=29°, $\phi_1=-11°$, $\phi_2=0°$	10.92	13.16	11.22	8.95	44.24
θ=29°, $\phi_1=-11°$, $\phi_2=11°$	11.01	10.85	9.80	8.00	36.66

axis angle and the two skew angles (see Figure 3.18). The goal of these rotations will be to bring the putative active site residues on the N-terminal lobe within 10Å of the ATP β phosphate.

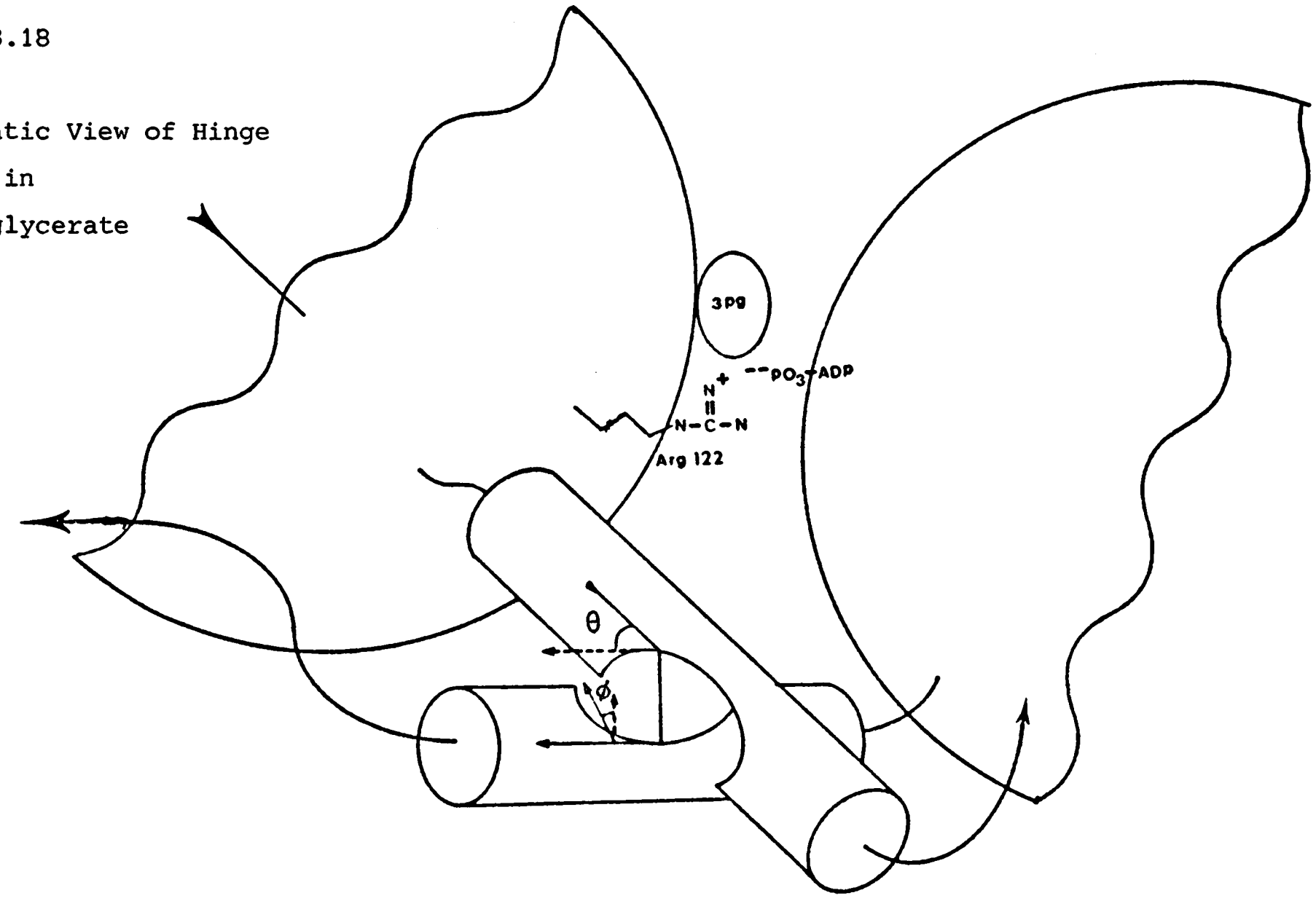
The helical axes of the two linking helices as well as the contact normal joining them were computed and rotation matrices for motions about the axes were produced. The chain was cut between the two lobes which then acted as independent rigid bodies. A crude line search minimiser with a step size of 0.1 radian was used to find the set of (θ,ϕ_1,ϕ_2) which minimised the sum of the target distances. These computations were performed by HINGE (see Appendix III). The final set of rotation angles was (29°, -11°, 11°). This brought Arg 122 to within 8Å of the β phosphate (see Table 3.18). These motions are consistent with the standard deviations for interaxial and skew angles observed in Type III interactions, θ ±26° and φ ±15° (see Table 3.13). The severed endpoints of the chain moved 12Å apart and 11 bad van der Waals contacts between C_α atoms were introduced. A procedure for regularising this structure to relieve these strains is being developed. The decrease in radius of gyration for our calculated structure was 10%. Scattering measurements suggest this decrease should be closer to 5% (Pickover et al., 1979). Perhaps a slight expansion of sections of the molecule will occur when the steric strain, which was concentrated on interactions between residues 164-170 and 388-392, is relieved.

The 8Å separation between the $N^{\eta}2$ of Arg 122 and the β phosphate is consistent with this residue mediating phosphate transfer. If $N^{\eta}2$ interacts with the γ phosphate of ATP, a 7.75Å separation between $N^{\eta}2$ and P^{β} would be expected. This is the sum of lengths of a hydrogen bond, 2.8Å, and 3 phosphorus oxygen bonds, 3 x 1.65Å = 4.95Å (see Figure 3.19).

Questions about the precise nature of this "hinged" conformation must await rebuilding of the current PGK model (D.Rice, personal communication) and the subsequent optimisation of the peptide geometry and van der Waals interactions which have been altered in this calculated conformational

FIGURE 3.18

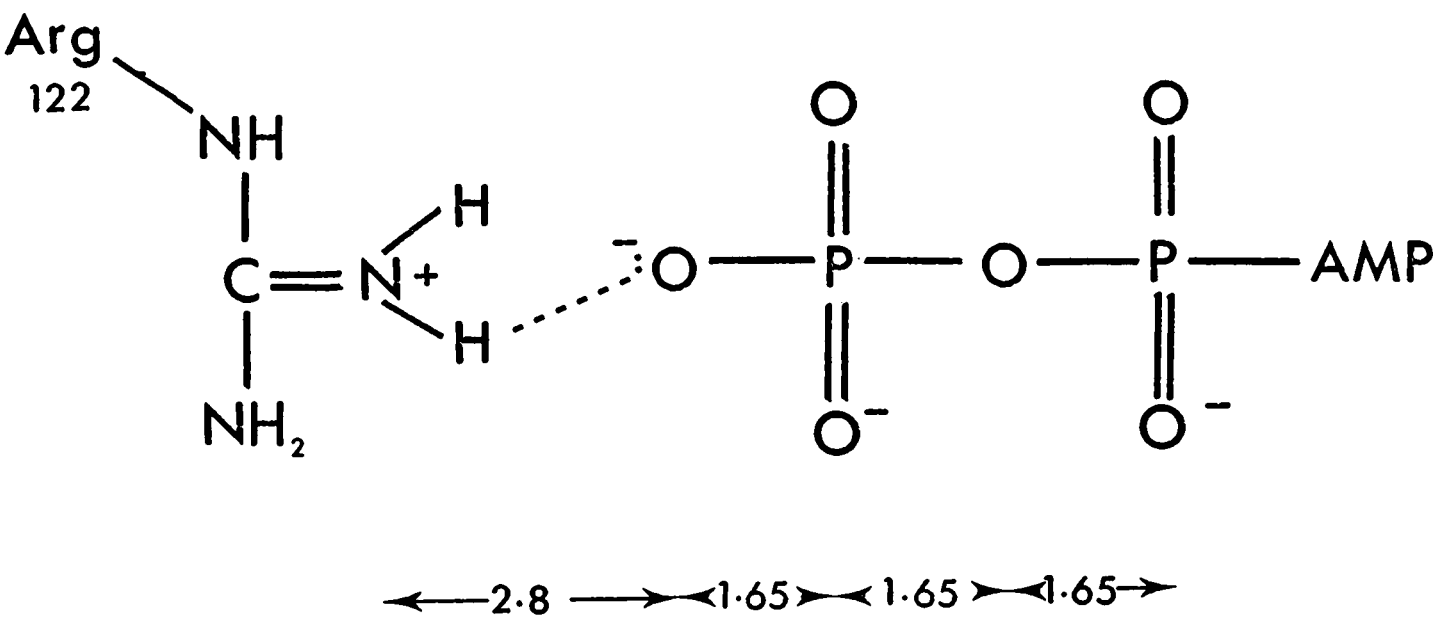
A Schematic View of Hinge Bending in Phosphoglycerate Kinase



The N-terminal lobe with Arg 122 is shown joined to the C-terminal lobe of phosphoglycerate kinase where ATP binding is seen. By changing ϕ, the skew angles, and θ, the helix axis angle, the two lobes, and more importantly Arg 122 and the ATP β phosphate are brought together. In this hinged position, catalysis of 3-phosphoglycerate could proceed.

FIGURE 3.19

The Geometry of an Arginine-ATP Complex.



An upper bound on the distance between the N_ε of Arginine 122 and the β phosphate of ATP is constructed from an average hydrogen bonding length, 2.8Å, and three phosphorus oxygen bonds, 1.65Å. The total distance, 7.75Å, is consistent with the values obtained from hinge bending calculations.

change. Perhaps this conformational change will produce favourable inter-digitations distinct from those seen in the ATP-PGK complex similar to those seen in the αβ contacts of oxy and deoxyhaemoglobin (Baldwin & Chothia, 1979).

3. A Model for the Kinetics of Protein Folding

3.1 Theory of Diffusion-Collision-Association

There are two distinct experimental and theoretical problems of protein folding, the thermodynamic issue of characterising the initial and final states and the kinetic question of the path that joins then (for a review, see Creighton, 1978). Some of the recent experimental studies have been summarised in Chapter I (see section I.5). This section considers theoretical models for the kinetic question. A very simple first order approximation is presented.

Baldwin and co-workers have concluded that folding is under intermediate, not nucleation, control (Baldwin, 1980; Schmid & Baldwin, 1979; Labhardt & Baldwin, 1979a,b). This suggests that there need not be an obligatory intermediate state and that folding may proceed along many paths.

Karplus & Weaver (1976, 1979) have proposed a model for investigating the diffusion-collision behaviour of α-helices. The time required for the interaction of two helices is:

$$\tau = \frac{1}{\beta} \frac{\ell \Delta V}{DA} \qquad (3.3)$$

where ℓ is the characteristic length, ΔV is the volume of finite diffusion space, D is a diffusion coefficient, A is the area of the target surface and β is the fraction of time that both segments are helical when they collide. I propose a diffusion-collision-adhesion model for the folding of myoglobin. (For experimental studies of myoglobin re-naturation, see

Harrison & Blout, 1965; Shen & Hermans, 1972.) In this formalism, unfolded apomyoglobin is considered as six fluctuating helices (ABEFGH). The interaction of two helices which actually pack in the protein (AH, BE, BG, GH and FH; Richmond & Richards, 1978) solidifies the two helix unit. Condensation continues toward the final structure with a fluctuating helix packing on a solid cluster or a pair of solid clusters interacting. This is seen schematically in Figure 3.20. This model is similar to that of Ptitsyn & Rashin (1975); intermediates are assemblies of α-helices. The major difference is that our model postulates the existence of many paths linking the unfolded and native states.

To calculate rate constants for the condensation of two helices, we assume that ΔV is the space between two centric spheres limiting the separation of complementary helix-helix interaction sites. The outer radius is determined by:

$$r_{ij}^{ext} = 1.5H + 3.3C \qquad (3.4)$$

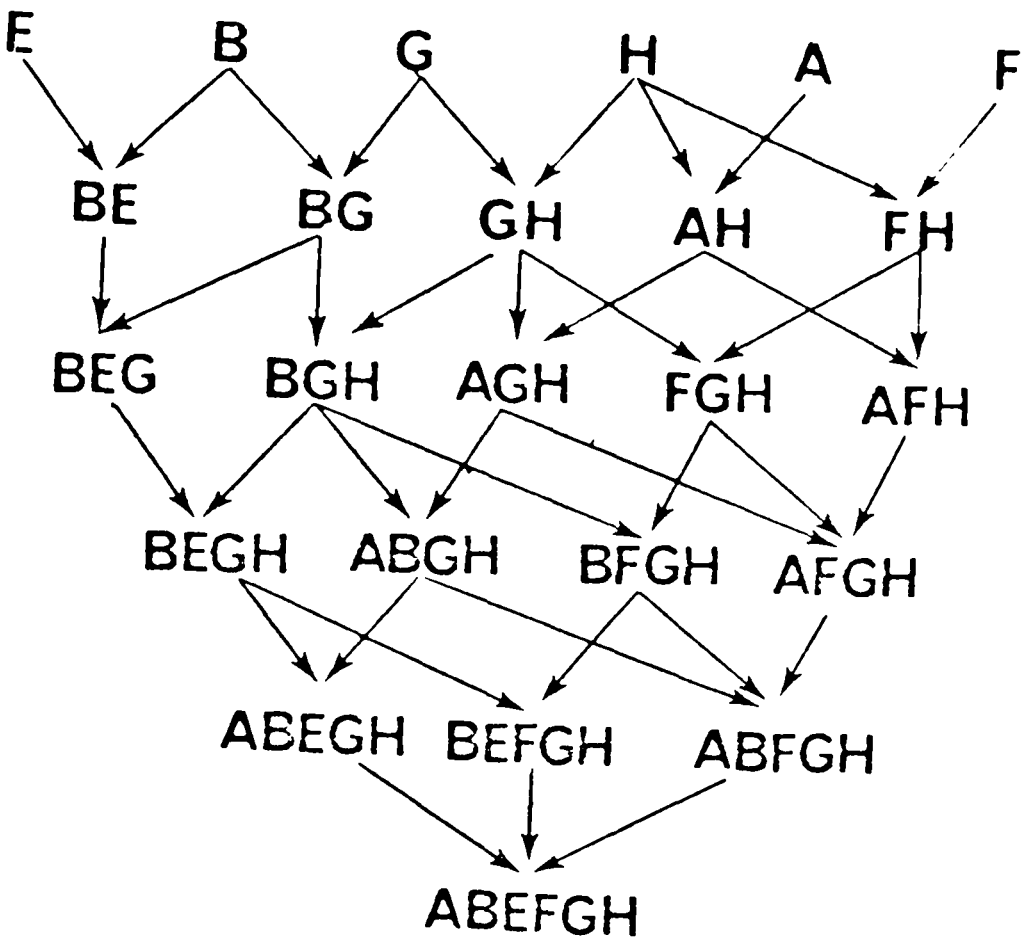
where i and j are the residues central to the helix-helix interaction, 1.5Å and 3.3Å are the pitch of an α-helix and an extended chain, and H and C are the number of residues between i and j in helical and coiled conformations. The inner radius, r_{ij}^{pack}, is taken as the length of the segment which is mutually perpendicular to the axes of the interacting helices (Richmond & Richards, 1978; Cohen et al., 1979). If the target area for the interaction is the average of the potential contact areas lost when the two helices pack (A_i and A_j), the equation (3.3) can be recast as:

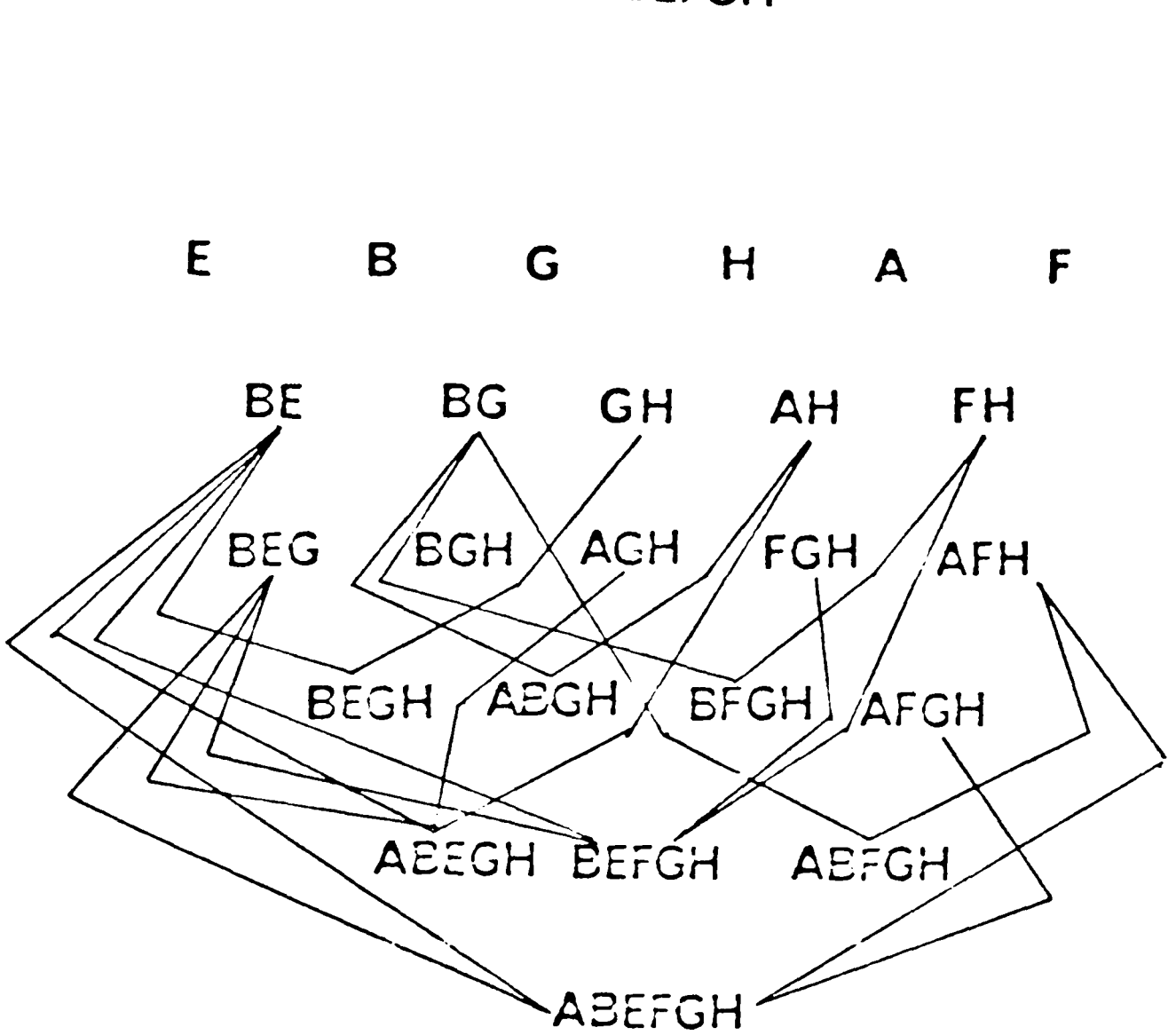
$$\tau_{ij} = \frac{\frac{1}{2}(r_{ij}^{ext} + r_{ij}^{pack}) \frac{4}{3}\pi\{(r_{ij}^{ext})^3 - (r_{ij}^{pack})^3\}}{\beta(10^{10}\text{Å s}^{-1}) \cdot \frac{1}{2}(A_i + A_j)} \qquad (3.5)$$

with the association constant $K_{ij}^{A} = 1/\tau_{ij}$. Clearly, r_{ij}^{ext} is a function of the degree of condensation of the molecule as well as where the fluctuating helices are in relation to what has been placed. For example, the volume of

FIGURE 3.20

Possible Pathways for Myoglobin Folding.





Each of the 24 states is a helix or collection of helices and the lines linking states are possible folding pathways. The upper panel shows pathways where the addition of a single helix to a cluster is allowed. The lower panel shows all cluster-cluster interaction pathways. If the two panels were superimposed on one another, all possible pathways for myoglobin refolding involving native-like intermediates would be included.

diffusion space available for the FH interaction is decreased if GH has been formed. The values for the parameters used to compute K_{ij}^{A} are listed in Table 3.19.

Dissociation constants were computed from a transition state model:

$$k_{cat} = \frac{kt}{h} e^{-\Delta G^{\ddagger}/RT} \qquad (3.6)$$

If a typical second order rate constant for helix-helix association is $10^{6} M^{-1} s^{-1}$ (Baldwin, 1980), then $\Delta G^{\ddagger}$ for assocation is 9.5kcal/mole. One estimate of the stability of an interacting pair of helices over an isolated pair of helices is the free energy loss calculated from accessible surface area considerations, ∿15kcal/mole (Richmond & Richards, 1978). Then $\Delta G^{\ddagger}$ for dissociation is approximately 15 + 9.5 = 24.5 kcal/mole (see Figure 3.21). This implies a second order rate constant for dissociation of $1.4 \times 10^{-5} M^{-1} s^{-1}$ which is negligible relative to the association rate constants. Thus this model is one of diffusion, collision, and adhesion. It is difficult to imagine that dissociation is always negligible as proteins do denature. However, in conditions which stabilise the folded state, helix-helix dissociation might indeed be small.

The only adjustable constant in this analysis is ρ, the probability that a potential helical segment will be helical when it interacts. The uniformity of ρ for all six helices stems from the Zimm & Bragg (1959) analysis of helix nucleation and elongation. Nucleation is the rate limiting step in helix formation and helix length is more a function of the helical tendencies of sequences of residues than of time. For a pair of fluctuating helices, $\beta = \rho^{2}$; for one fluctuating helix packing a solid cluster, $\beta = \rho$; and for two solid clusters packing $\beta = 1$. In the two simulations that follow, $\rho = 0.002$. Smaller values of ρ increase the lag time, decrease the rate of the transition between the initial and final states and slow the time required for equilibrium to be achieved. Larger values of ρ have the opposite effect.

TABLE 3.19

Parameters for the Diffusion-Collision Equation

Interaction	Average potential contact area loss on association* $A=(A_i+A_j)/2$ (Å)^2	r_{ij}^{pack} (Å)	r_{ij}^{ext} (Å)
G111-H135	185	10.5	25.05 (17.55[1], 19.05[3], 17.55[4])
A10-H134	170	8.5	189.3 (95.3[2], 26.0[3], 118.2[4], 142.05[5], 15.2[3,5])
B25-E65	145	7.5	50.70 (43.20[3], 24.00[3,4,5])
F90-H142	125	8.5	66.90 (54.15[1], 12.90[5])
B28-G110	155	8.5	120.00 (45.45[2], 112.50[5], 12.9[1,5] 89.40[4,5])

* from Richmond and Richards (1978)

1 AH formed

2 BE formed

3 BG formed

4 FH formed

5 GH formed

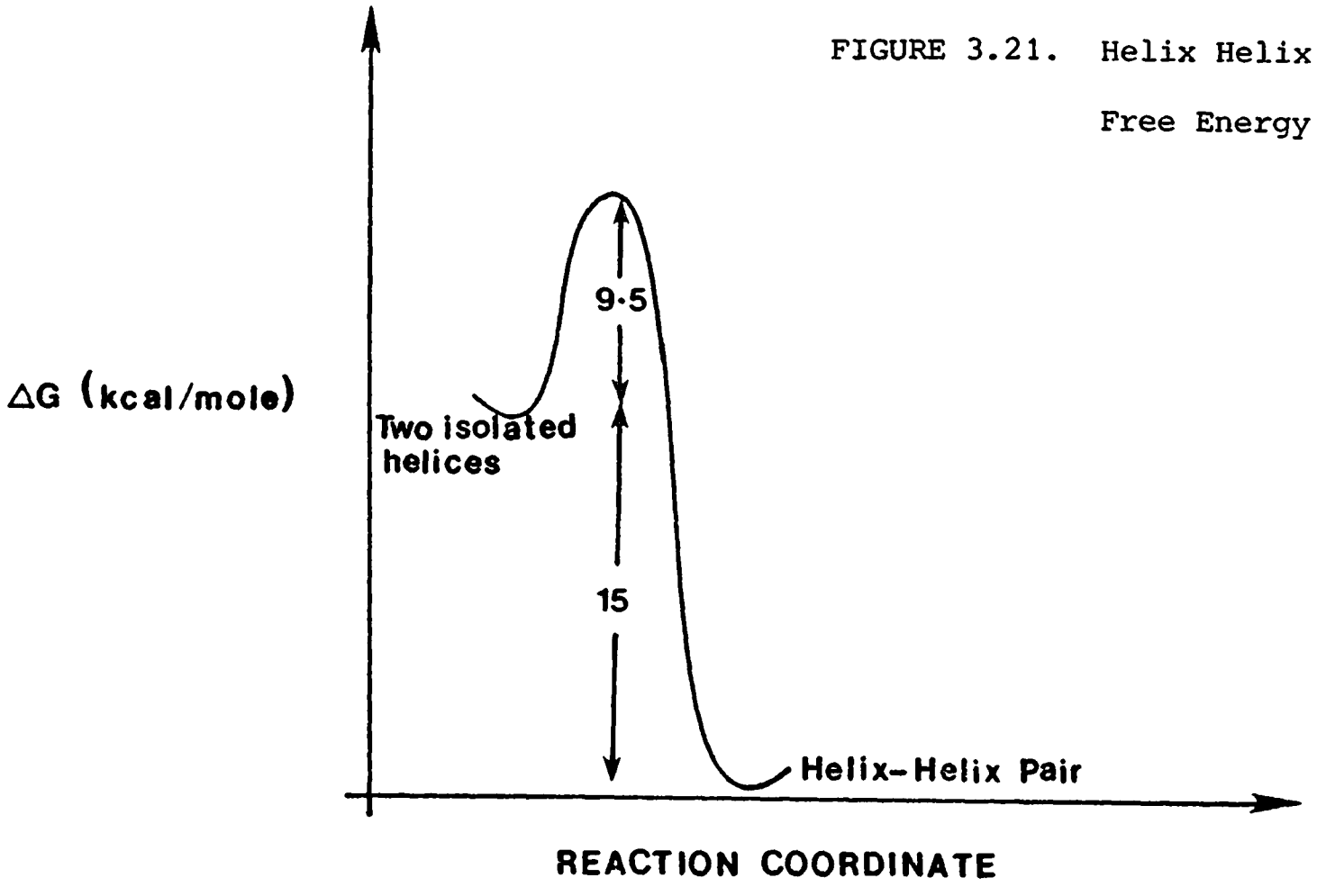


FIGURE 3.21. Helix Helix Interaction Free Energy Profile

Free energy profile of the difference in stability of an interacting pair of α-helices and two isolated helices. 9.5kcal/mole is the transition states approximation to the barrier height from a rate constant of 10^6 M s^{-1}. The 15 kcal/mole difference is taken from the non-polar accessible contact area measurements of Richmond & Richards (1978)

3.2 Renaturation of Apomyoglobin

In this model of apomyoglobin, there are 24 states (see Figure 3.20). There exists a matrix of 24x24 second order rate constants which define a system of simultaneous second order differential equations:

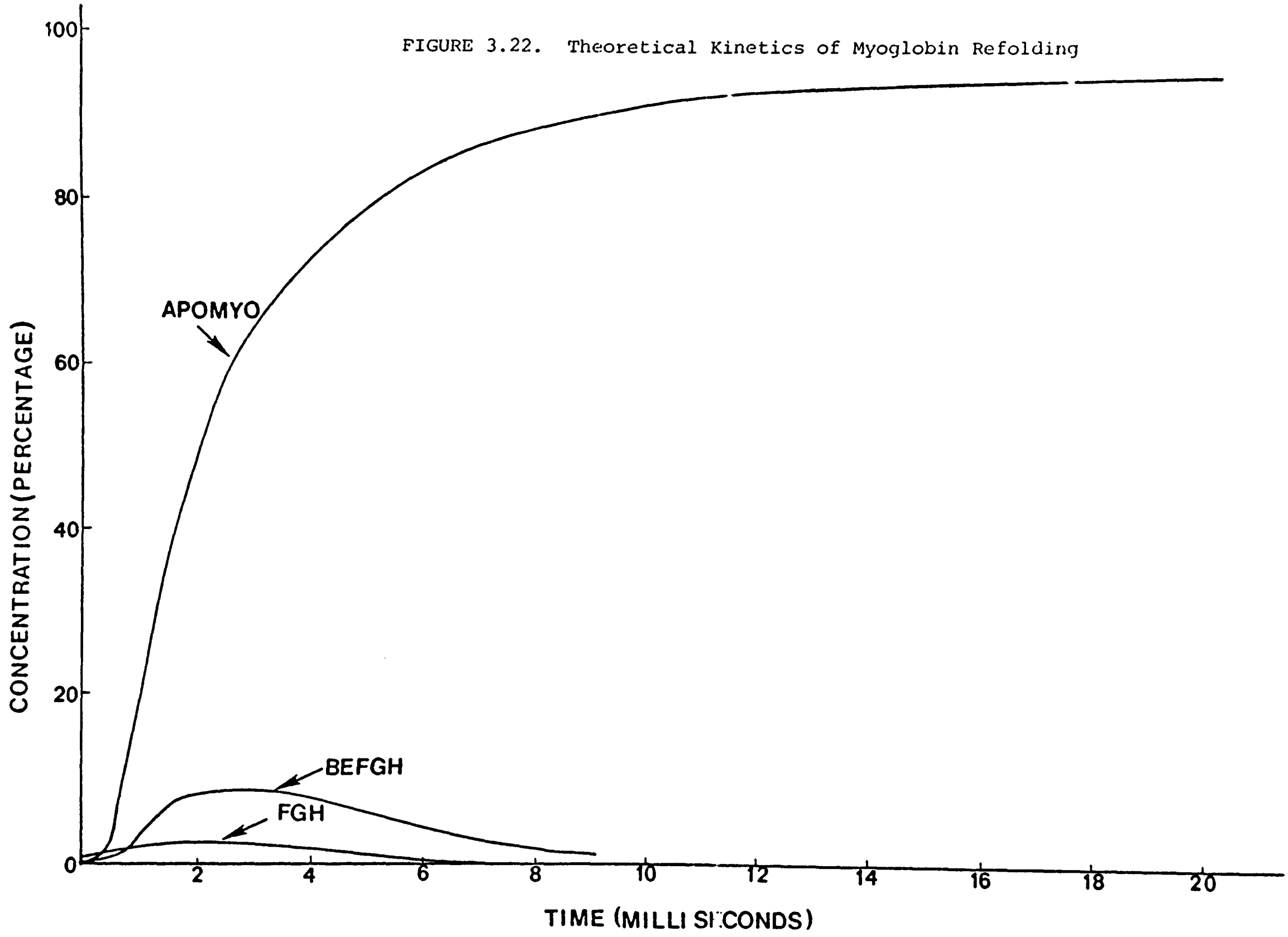
$$\frac{\partial [C_i]}{\partial t} = [C_i]^T [K_{ij}] [C_i] \qquad (3.7)$$

where $[C_i]$ is the row matrix of concentrations for each state i. As a closed form solution to this problem does not exist in general, the rate equations are solved iteratively with time steps of 25μsec. The system is constrained so that mass is conserved. Figure 3.22 shows the sigmoid behaviour of the final state as a function of time. The concentration begins to level off at 90% after 4ms and slowly continutes to completion in the next 16ms. Two intermediate states are significantly (>5%) populated: FGH and ABFGH. The lag phase, though noticeable, is small. This is undoubtedly because the initial steps in this simulation overwhelmingly favour products but that the rate constants place a lower bound on the time it takes the initial state to reach the final state. The computer program for this calculation, KINETIC, is in Appendix III.

3.3 Protein Biosynthesis and N-terminal Folding

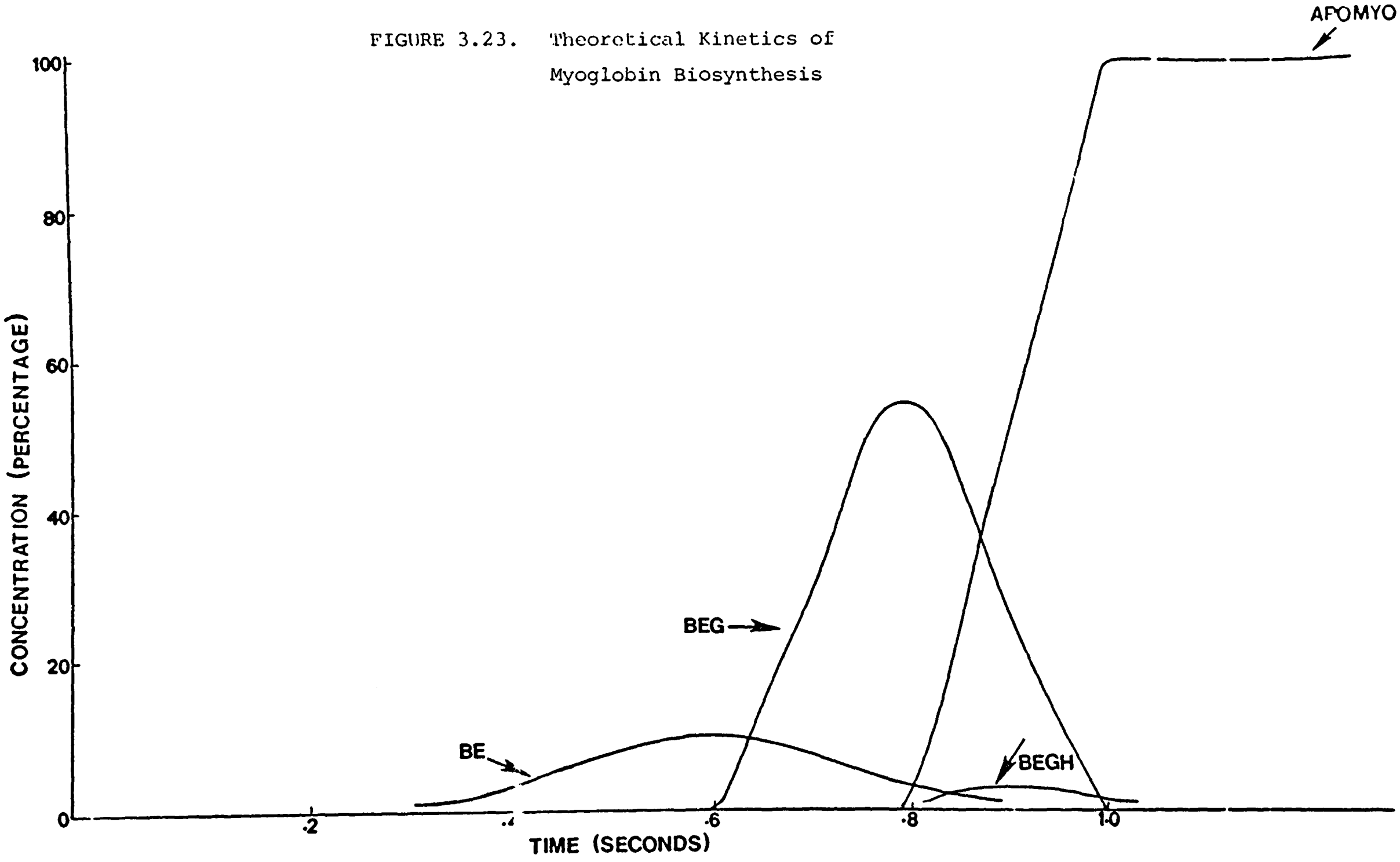
In an attempt to simulate biosynthesis, the initial concentrations of the isolated helical states were in sequential order iteratively increased from 0 to 100%. Thus the concentration of the H helix was 0.0 until the concentration of the G helix reached 1.0. The rate of synthesis was such that a myoglobin chain was made in 1.2 seconds. BE is the first intermediate to appear, followed by the production of BEG, BEGH and then the native state (see Figure 3.23). The computer program for this calculation, NTFOLD, is in Appendix III.

FIGURE 3.22. Theoretical Kinetics of Myoglobin Refolding



Simulated re-naturation kinetics of apomyoglobin as a function of time. The concentrations of intermediates which are significant are also plotted.

FIGURE 3.23. Theoretical Kinetics of Myoglobin Biosynthesis



Simulated folding kinetics of apomyoglobin during biosynthesis. The time taken to synthesise a helical segment is 0.2 seconds.

3.4 An Evaluation of the Results of a Diffusion-Collision-Adhesion Model for Protein Folding

In these two simulations, two distinct sets of intermediate states were significantly populated. Thus, folding could occur through many alternative pathways which ultimately converge to the native structure. Cooperativity results from the decrease in the volume of diffusion space as a function of molecular organisation. In principle, these intermediates could be trapped and identified in a stopped-flow experiment. Alternatively, the sites of helix-helix interaction could be chemically blocked to inhibit complete renaturation and the intermediates created could then be characterised. Studies by Zabin and co-workers suggest that intermediates in biosynthesis could be characterised by immunological probes (e.g. Hamlin & Zabin, 1972).

This model assumes that the interaction of mismatched pairs of hydrophobic patches on the surface of the α-helices are not stabilised. A more realistic assumption would be that mismatched helix-helix pairs could tautomerise to a correct pairing. These tautomerisations would have low activation barriers as the individual helical conformations have been stabilised and the volume of diffusion space is reduced. This tautomerisation mechanism might explain Creighton's (1978) finding that there is an obligatory mismatched disulphide bridge on the pathway joining the unfolded and native states of pancreatic trypsin inhibitor.

CHAPTER 4

ALL-β PROTEINS

In contrast to the highly cooperative local organisation of the α-helix, a β-sheet has a network of hydrogen bonds which link sequentially distant parts of the polypeptide chain. To satisfy the hydrophobic and hydrophilic tendencies of the individual amino acids in the chain, proteins sometimes adopt a "β-sheet sandwich" conformation. This common structural motif follows from the packing of two largely anti-parallel β-sheets. The result is an internal core for hydrophobic residues and the simultaneous accommodation of the hydrophilic preferences of neighbouring residues. A sheet sandwich typically has 100 amino acid residues and each sheet is constructed of three or four strands although concanavalin A has a total of 14 strands in two sheets. β-Sheet sandwiches are frequently observed as folding domains for large proteins (e.g. the immunoglobulins) and are often the nexus of intersubunit contacts in multimeric proteins.

1 Prelude

The detailed atomic structures of ten β-sheet sandwiches are analysed in this Chapter to provide insight into the organisation of this structural unit. The stability results from interstrand hydrogen bonding and the burial of hydrophobic residues both within and between the β-sheets. As Lifson & Sander (1980) had shown that there is little pairwise specificity for the aligned residues on neighbouring β-strands, concentration was placed on the hydrophobic effect between sheets. The magnitude of this effect was estimated from the change in non-polar accessible contact area (NPACA) using the empirical relationship 1Å^2 = 80 cal/mole (Richmond & Richards, 1978; Chothia, 1974).

The changes in NPACA upon the formation of a β-sheet from a collection of isolated β-strands as well as the NPACA change following the packing of the two β-sheets into a β-sandwich were evaluated. Although all residues lost NPACA upon β-sheet formation, losses of NPACA upon β-sandwich formation were usually limited to the hydrophobic residues which clustered to form a parallelogram on the internal face of the β-sheets. Moreover, these parallelograms are anti-complementary; the major diagonal of the pattern on the top sheet proceeds from the upper left to the lower right corners while the major diagonal of the pattern on the bottom sheet runs from the lower left to the upper right corners. Other geometric and hydrogen bonding preferences were quantified.

A combinatorial algorithm which incorporated these observations as well as topological restrictions noted in other studies of β-structure (Sternberg & Thorton, 1977a; Richardson, 1977; Ptitsyn et al., 1979) was developed and applied to nine proteins. With only secondary structure and amino acid sequence as input, this algorithm reduces the number of possible structures from $\sim 10^8$ to between 6 and 3000 structures. Moreover, one of these is a good approximation to the native with r.m.s. deviations between 1.4 and 5.1Å. This procedure was also applied to two proteins of known sequence and unknown structure, $\beta_2$microglobulin and the histocompatibility factor ac-2.

2 Definitions

2.1 Input Data

The following amino acid sequences (Dayhoff, 1976) were used in all of the analysis and prediction work in this chapter:

1 - 4: the human immunoglobulin IgG(λ) new fragment for the variable (V) and constant (C) domains of the heavy (H) and light (L) chains - FAHV, FAHC, FALV and FALC.

5 - 6: Human immunoglobulin F_c fragment C_H2 and C_H3 domains - FCH2 and FCH3.

7: Human Prealbumin - PRE.

8: Cu,Zn superoxide dismutase - SDM.

9: Jack Bean concanavalin A - CONA.

10: Bence-Jones protein variable dimer - REI.

The secondary structure assignments used in this Chapter are listed in Table 4.1. These assignments were made on the basis of hydrogen bonding diagrams and main chain dihedral angles.

2.2 Structural Data

The crystallographically determined atomic coordinates used in this Chapter were taken from the protein data bank (CONA, SDM & REI, Bernstein et al., 1977) or were supplied directly by the author (FAHV, FAHC, FALV & FALC, Saul et al., 1978; FCH2 & FCH3, Huber et al., 1976; PRE, Blake et al., 1978).

2.3 Description of an Ideal β-Sheet and β-Sandwich

All β-strands constructed in this Chapter were built using the model building program of Scheraga and co-workers (Momany et al., 1975) with ϕ = -120° and ψ = +140°. A β-sheet was formed by docking the ideal strands so that the interstrand separation was 4.25Å and inter interstrand dihedral interaxial angle was -20° (Chothia, 1974). A β-sandwich was constructed by placing one ideal β-sheet on top of another so that the intersheet separation was 10.0Å and the intersheet dihedral interaxial angle or twist was -30° (see Figure 4.1). This idealised construction will be used for comparing NPACA data from crystallographic coordinates to a polycysteine model as well as in the construction of the predicted structures at the end of this Chapter.

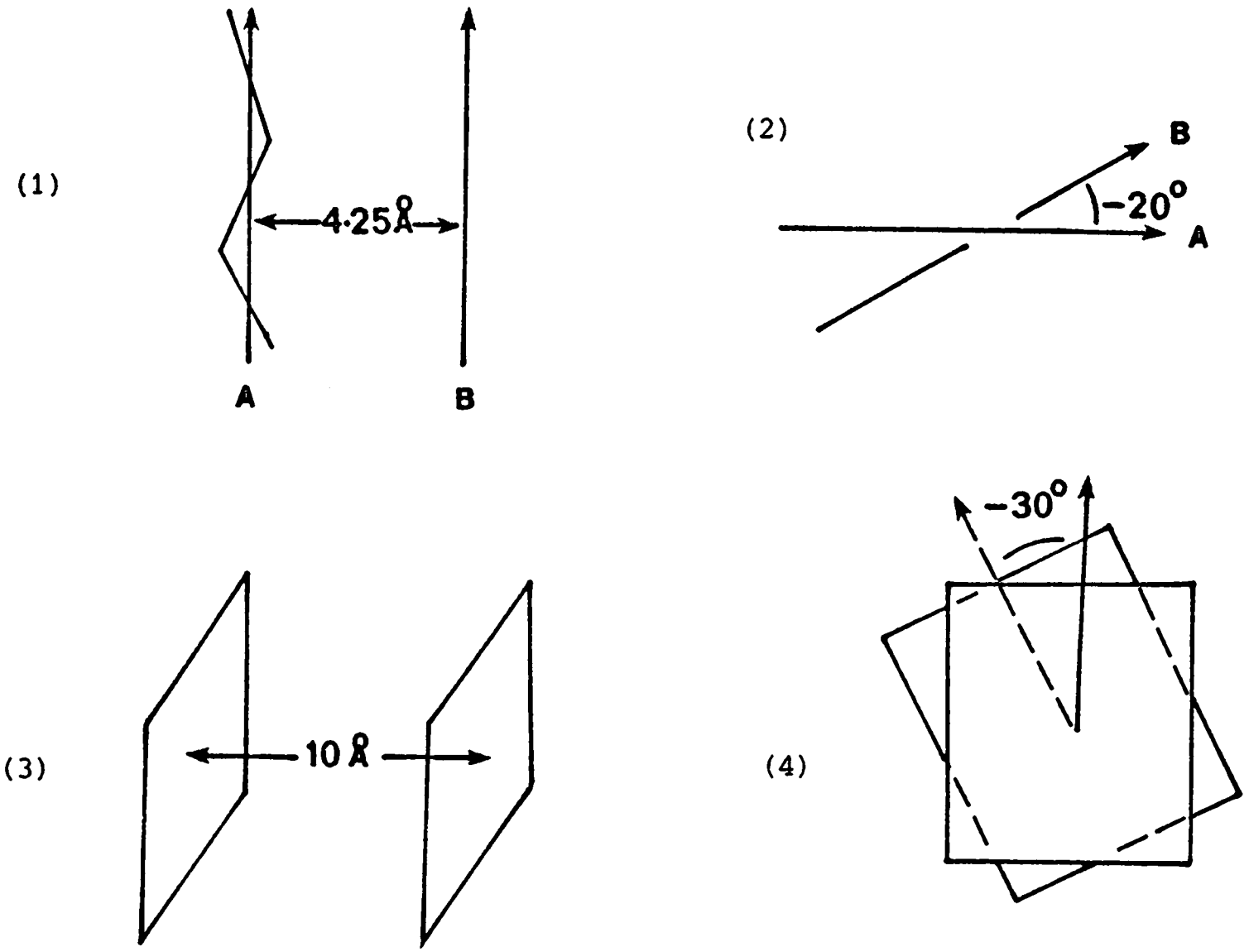
3 Analysis of β-Sandwiches

The stability of a β-sandwich results from interstrand hydrogen bond-

TABLE 4.1. Strand Assignments for 10 β-Sandwich Proteins

Protein	Strand	N-terminus	C-terminus
Pre-albumin	A	11	19
	B	28	36
	C	40	49
	D	53	55
	E	67	74
	F	91	97
	G	104	112
	H	115	123
Con-cana-valin A	A	3	11
	B	24	29
	C	36	40
	D	48	55
	E	60	66
	F	73	79
	G	87	97
	H	103	116
	I	124	130
	J	139	143
	K	147	149
	L	153	156
	M	169	181
	N	188	200
	O	208	215
REI - Bence-Jones Fragment	A	2	7
	B	10	13
	C	17	26
	D	33	38
	E	45	47
	F	61	65
	G	69	76
	H	84	90
	I	97	107
FAHV Human IgG(λ) Heavy Chain Variable Domain Antibody Fragment	A	4	7
	B	18	25
	C	32	39
	D	45	52
	E	56	60
	F	67	70
	G	78	82
	H	91	99
	I	103	114
FAHC Human IgG(λ) Heavy chain constant domain Antibody Fragment	A	121	128
	B	142	150
	C	154	159
	D	166	175
	E	179	181
	F	198	204
	G	209	215
FALV Human IgG(λ) light chain variable domain	A	8	12
	B	17	24
	C	36	40
	D	47	50
	E	57	62
	F	64	70
	G	79	87
	H	90	102
FALC Human IgG(λ) Light chain Constant domain	A	110	115
	B	126	134
	C	140	145
	D	154	163
	E	167	177
	F	187	193
	G	196	202
FCH2 Human Immuno-globulin crystalline fragment heavy chain	A	238	243
	B	259	264
	C	274	278
	D	293	294
	E	299	304
	F	318	323
	G	332	336
FCH3 Human Immuno-globulin crystalline fragment heavy chain	A	344	351
	B	362	373
	C	378	382
	D	390	399
	E	403	412
	F	422	428
	G	437	441
SDM Super Oxide Dismutase	A	2	10
	B	14	23
	C	26	35
	D	39	46
	E	80	87
	F	90	99
	G	113	118
	H	143	148

FIGURE 4.1. A Schematic View of Four Geometric Parameters Commonly Used in the Description of β-sandwiches.



(1) Interstrand Separation - this value is typically 4.25Å although variations for parallel and anti-parallel sheets are expected.

(2) Interstrand Angle - this parameter quantifies the characteristic twist of a β-sheet.

(3) Intersheet Separation - this is the perpendicular distance between the sheet planes for two close packed β-sheets.

(4) Intersheet Angle - the packing between β-sheets exhibits a characteristic negative angle of ∿-30° between all β-sandwiches.

ing and the burial of hydrophobic residues both within and between the β-sheets. An attempt to quantify these two contributions was made by studying the overlap and mismatch of residues along adjacent strands in a sheet and by calculating the per residue change in NPACA on packing the two isolated β-sheets into the native β-sandwich.

3.1 The Condensation of Isolated β-strands into a β-sheet

When two isolated β-strands are brought together to form a β-sheet, the sheet is stabilised by interstrand hydrogen bonding and the partial burial of some non-polar atoms. The NPACA change for forming a β-sheet from the isolated β-strand was calculated for the nine β-sandwich proteins (SDM was excluded from further analysis because only alpha carbon coordinates are generally available). Table 4.2 and Figure 4.2 reveal that these changes are extensive and similar for all nine examples. For prealbumin, the change in the hydrophobic contribution to free energy is 74 kcal/mole (see Figure 4.3) as estimated from the empirical relationship $1\mathring{A}^2$ = 80 cal/mole (Richmond & Richards, 1978; Chothia, 1974). This results in a per residue hydrophobic free energy change of between 0.89 kcal/mole-residue for FALC and 1.22 kcal/mole-residue for FAHV. Although these changes are extensive, they are not residue-specific and are accurately modelled by the changes expected for a guest residue in an ideal polycysteine β-sheet (see Table 4.3). Both hydrophobic and hydrophilic residues are buried in an amount proportional to their volume (see Figure 4.4).

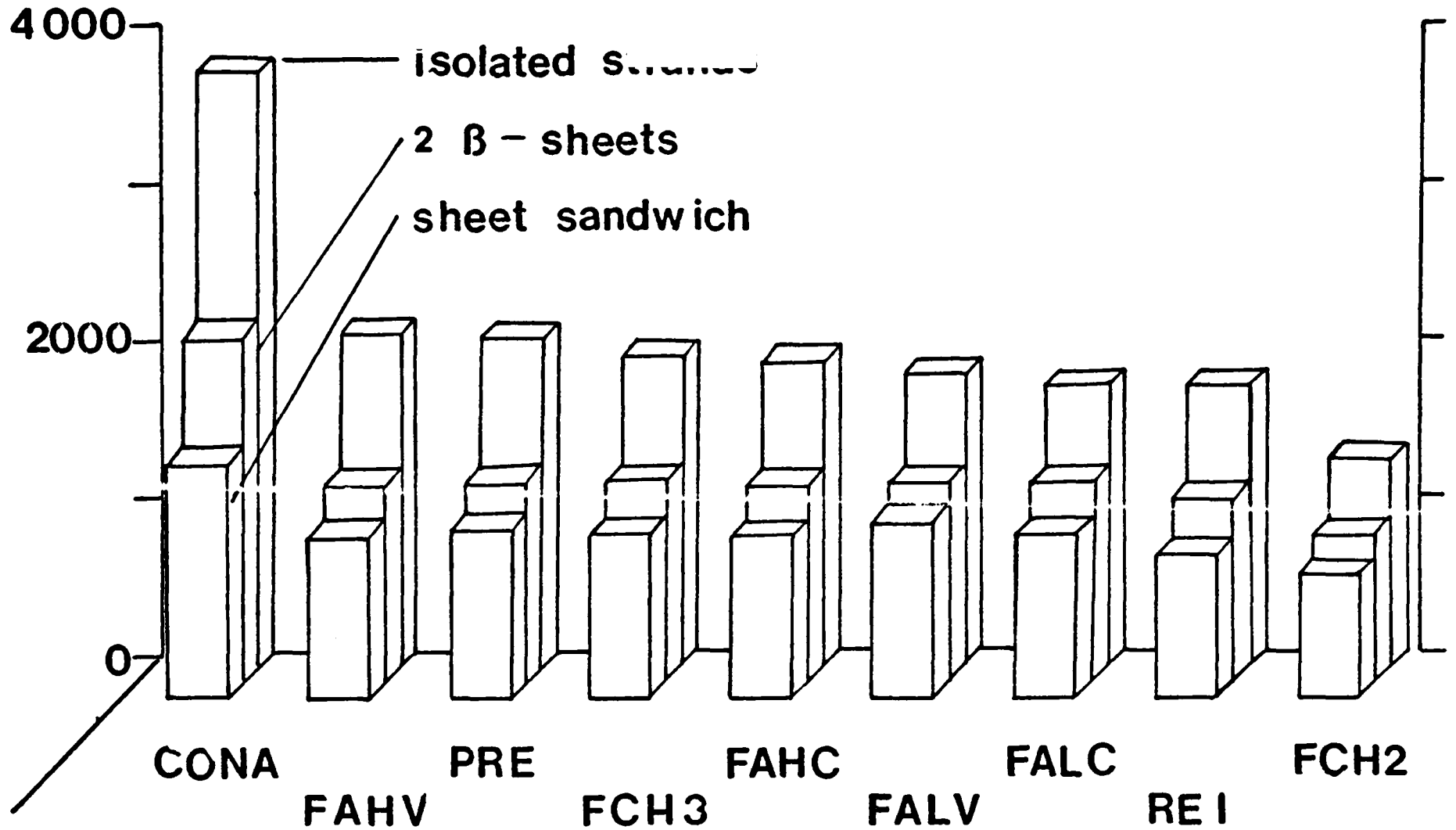
Thus β-sheet formation is stabilised by two contributions: a favourable hydrophobic effect following the burial of non-polar atoms; and a favourable electrostatic effect from the hydrogen bonds. These two concepts were unified through a related figure, Strand Overlap. Overlap is calculated as one half the sum over all residues in the sheet of the number of residues on neighouring strands which lie directly across from each residue, Thus, the hypothetical sheet:

TABLE 4.2

Average non-polar accessible contact area (NPACA) changes for the formation of β-sheets from isolated β-strands and β-sandwiches from 2 β-sheets.

Protein	Residues in β-sheet	NPACA Change Strands → Sheet			NPACA Change Sheet → Sandwich		
		Total (Å²)	Per residue (Å²)	(kcal/mole)	Total (Å²)	Per residue (Å²)	(kcal/mole)
Concana-valin A	118	1710	14.5	1.16	796	6.8	0.54
FALC	56	658	11.8	0.89	312	5.6	0.45
FALV	57	692	12.1	0.97	230	4.0	0.32
FAHV	63	963	15.3	1.22	370	5.9	0.47
FAHC	56	803	14.3	1.14	323	5.8	0.46
Pre-albumin	64	924	14.4	1.15	303	4.7	0.38
FCH2	36	492	13.7	1.09	243	6.7	0.54
FCH3	57	806	14.1	1.13	306	5.4	0.43
REI	60	743	12.4	0.99	336	5.6	0.45

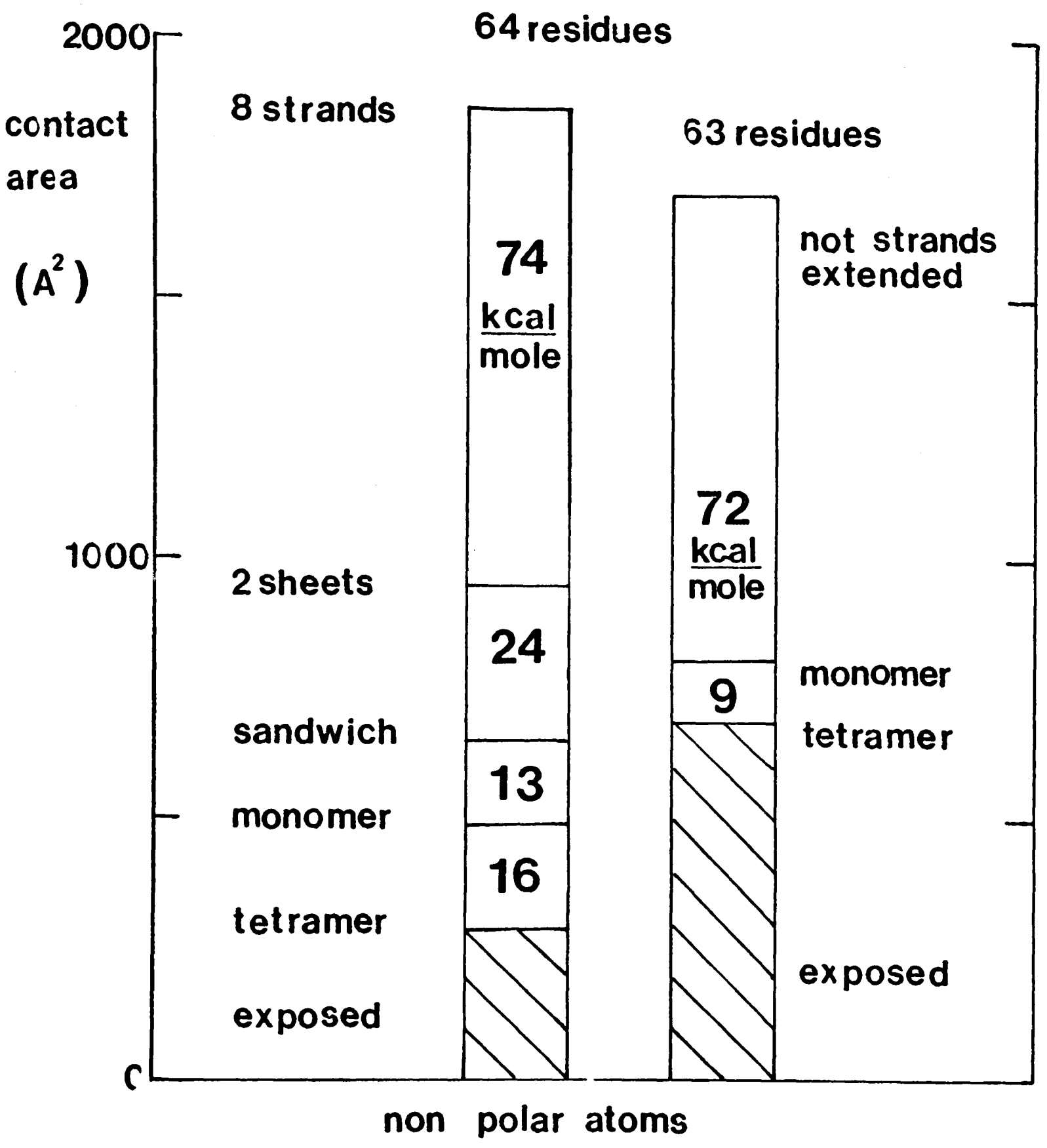
FIGURE 4.2. FORMATION OF β – SHEET SANDWICHES



For each sandwich, the Figure shows the solvent accessible, non-polar contact area for: all the strands when isolated; two β-sheets when isolated; and the β-sheet sandwich.

FIGURE 4.3

PREALBUMIN – SOLVENT CONTACT AREA



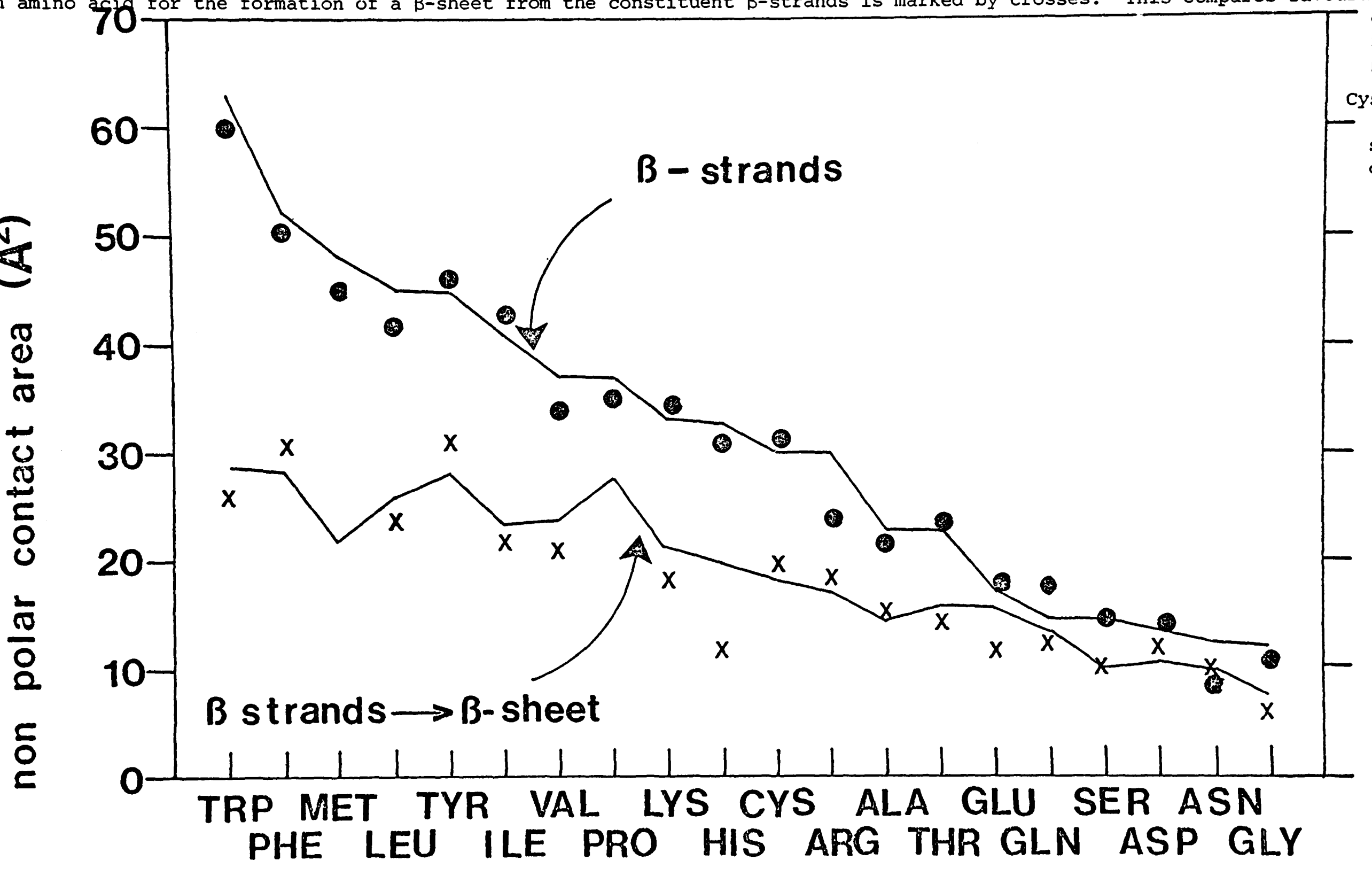
The difference in the hydrophobic effect between the various levels of structural organisation in prealbumin. This effect is quantified by accessible contact area calculations and translated in kcal/mole. Note that the major differences are seen for the residues in the secondary structure units shown on the left.

FIGURE 4.4. A Comparison of Observed Contact Area Changes in Nine Proteins and a Polycysteine Model for β-strands and β-sheet Formation

The average contact area for each amino acid in nine proteins when considered as isolated strands is marked by dots. This agrees well with the contact area of a residue in an ideal β-strand Cys_3XCys_3 shown by the connected points. The average change in contact area for each amino acid for the formation of a β-sheet from the constituent β-strands is marked by crosses. This compares favourably with the change expected for the ideal β-sheet:

$Cys_7:Cys_3XCys_3:Cys$

shown by the connected points



Glu Ala Val
Leu Leu
Lys Leu

would have overlap equal to four. A study of the ten β-sandwiches in the bubble diagrams (see Figure 4.5) indicates that all β-sheets have maximal or one less than maximal overlap. Thus the hypothetical sheet above would be reasonable, but:

Val Lys Leu
Gly Ala Val Leu
Leu Leu

would not. In the second example, the top strand could be shifted to the right by two residues and improve the overlap while maintaining the sidedness of the sheet. A shift of one residue would force the strand to reverse its sidedness as β-strands have a two-fold repeat.

3.2 The Formation of a β-sandwich from two β-sheets

The stability of a β-sandwich results from the ability of the constituent β-sheets to exclude water from interstitial spaces upon association. This solvent shielding effect was quantified as the per residue change in NPACA for packing two isolated β-sheets into a β-sandwich. For prealbumin, the change for all the residues in the β-sandwich is 24 kcal/mole and the per residue changes are between 0.32 kcal/mole-residue for FALV and 0.54 kcal/mole-residue for FCH3 (see Table 4.3). These changes are comparable for all nine examples (see Figure 4.2). However, the residues showing large changes are predominantly hydrophobic (see Table 4.3).

In a β-sandwich sequential side chains point alternatively toward and away from the other β-sheet producing two surfaces. However, β-sheets are far from planar. The average interstrand dihedral interaxial angle of -20° introduces distortions which force the two surfaces of the β-sheet to be asymmetric (Chothia, 1973). This may be responsible for the characteristic negative intersheet dihedral interaxial angle, $-30° \pm 10°$. This also causes

TABLE 4.3. Non-Polar Accessible Contact Areas (NPACA) in β-Sheets

Amino Acid	Mean & STD: NPACA for residues in isolated β-strands	Mean & STD: Change in NPACA on β-sheet formation	Mean & STD: NPACA for residues in isolated β-sheets	Mean & STD: Change in NPACA on β-sandwich formation	NPACA for a model poly-Cys. β-sheet with a guest res. $Cys_7:Cys_3$ X $Cys_3:Cys_7$ Single strand	Complete β-sheet	Change*
Alanine	21.5±3.4	15.2±3.6	9.8±5.4	8.0±5.0[a]	22.7	8.2	14.5
Arginine	23.8±2.5	18.4±2.4	11.7±6.0	2.3[b]	22.9	5.7	17.2
Asparagine	8.1±4.8	10.4	3.1±2.7	5.4	12.4	2.9	9.5
Aspartic Acid	13.4±3.2	12.8	7.5±4.6	3.7	13.4	2.8	10.6
Cysteine	30.9±2.0	19.8±3.7	14.0±8.4	11.2±3.7	29.6	11.2	18.4
Glutamine	17.1±4.1	12.9±2.9	10.3±6.4	7.1±3.0	14.5	0.6	13.9
Glutamic Acid	17.4±3.0	11.7±1.1	11.6±4.7	6.4	16.9	1.1	15.8
Glycine	11.3±2.1	6.2±2.7	6.7±4.2	4.5±2.0	11.9	4.2	7.5
Histidine	31.0±4.2	12.1±3.4	22.0±7.0	8.3±6.9	32.6	13.2	19.4
Isoleucine	43.1±3.7	22.7±4.1	24.9±10.2	16.0±7.1	41.8	18.2	23.6
Leucine	42.2 4.2	23.7 4.1	26.1 8.7	17.9 7.2	45.3	19.4	25.9
Lysine	32.3±4.8	18.5±6.4	23.4±9.9	16.0±1.5	33.3	12.4	20.9
Methionine	45.2±3.3	-	33.4±6.5	29.4±4.4	47.9	25.8	22.1
Phenyl-alanine	51.0±5.3	31.1±7.2	30.4±11.1	20.8±7.3	52.2	24.2	28.0
Proline	35.0±3.0	-	29.4±9.4	10.6±4.2	36.7	8.9	27.8
Serine	14.8±2.6	10.1±2.6	9.4±5.6	4.1±2.7	14.3	4.7	9.6
Threonine	23.3±3.3	14.7±3.8	16.0±7.7	9.9±4.7	22.5	6.7	15.8
Tryptophan	59.9±4.7	25.8±4.4	35.2±9.0	32.5±7.4	63.0	34.1	28.9
Tyrosine	44.5±2.0	31.3±5.7	22.8±11.5	12.6±4.5	44.9	16.7	28.2
Valine	34.4±3.9	20.9±4.3	21.3±7.7	14.3±5.7	37.2	13.4	23.8

All figures given in $\mathring{A}^2$.

[a] Only significant changes (>1$\mathring{A}^2$) included.

* Change on condensation.

[b] No standard deviation for one example.

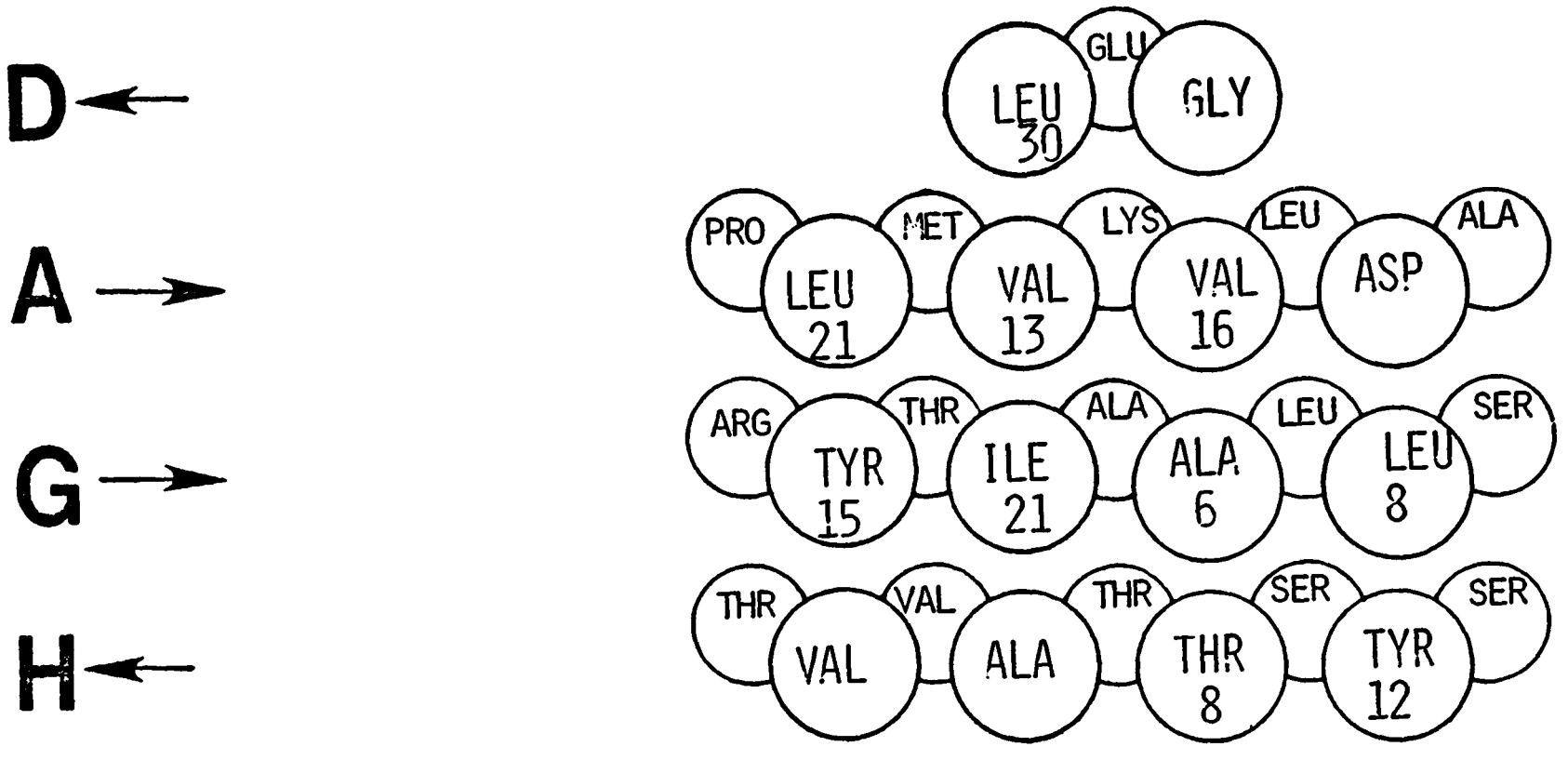
A per residue breakdown of the NPACA for isolated β-strands and the resultant change upon the formation of β-sheets as well as isolated β-sheets and the resultant change upon the formation of β-sandwiches. The values in columns 1, 2 and 3 fit the values for a model poly-cysteine β-sheet in columns 4, 5 and 6 respectively.

FIGURE 4.5

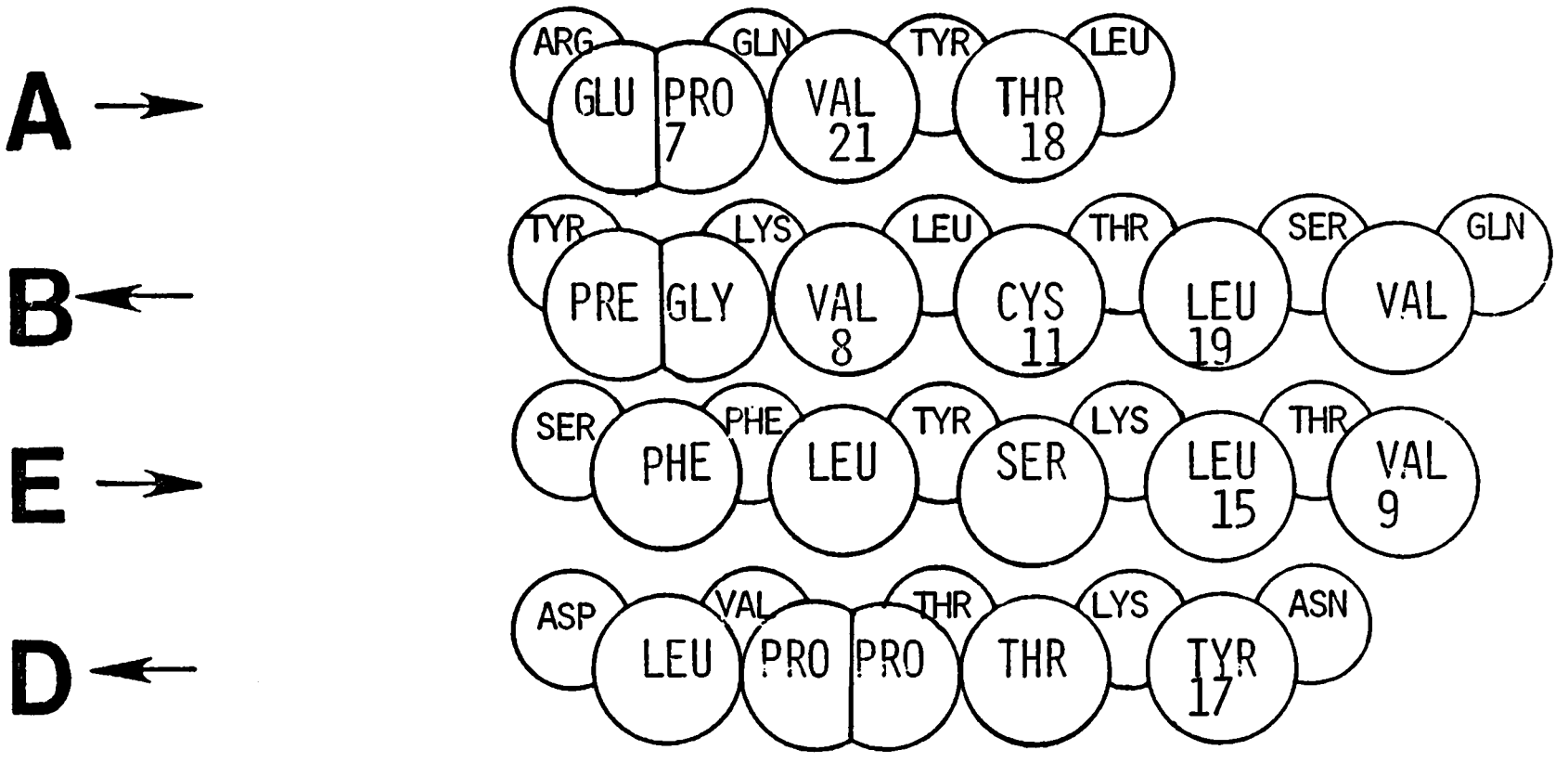
Non-Polar Accessible Contact Area Changes for β-sandwiches

The details of changes in accessible contact area for eight β-sandwiches for the transition between two isolated β-sheets and a β-sandwich. The amino acid sequence of the strands is included in these bubble diagrams with the large spheres on the internal faces of the β-sheets. Strand alignment derived from hydrogen bonding diagrams is shown. β-bulges are also depicted.

PREALBUMIN


D
A
G
H
GLU
LEU 30
GLY
PRO
LEU 21
MET
VAL 13
LYS
VAL 16
LEU
ASP
ALA
ARG
TYR 15
THR
ILE 21
ALA
ALA 6
LEU
LEU 8
SER
THR
VAL
VAL
ALA
THR
THR 8
SER
TYR 12
SER
C
B
E
F
TRP
THR
GLU
PRO
PHE
SER
ALA
GLY 6
LYS
THR 22
LYS
ALA
ARG
PHE
VAL 7
HIS
VAL 10
ALA
VAL 8
ILE
GLY
TYR
LYS
VAL 10
GLU
ILE 20
ASP
THR
ALA 8
PHE 18
VAL
VAL 18
GLU
ALA 18

FC CH3


A
ARG GLU PRO 7 GLN VAL 21 TYR THR 18 LEU
B
TYR PRE GLY LYS VAL 8 LEU CYS 11 THR LEU 19 SER VAL GLN
E
SER PHE PHE LEU TYR SER LYS LEU 15 THR VAL 9
D
ASP LEU VAL PRO PRO THR THR LYS TYR 17 ASN
G
GLN THR 10 LYS 18 SER LEU 26
F
MET VAL 16 SER CYS 11 SER PHE 20 VAL
C
ALA VAL 19 GLU 6 GLU TRP 39

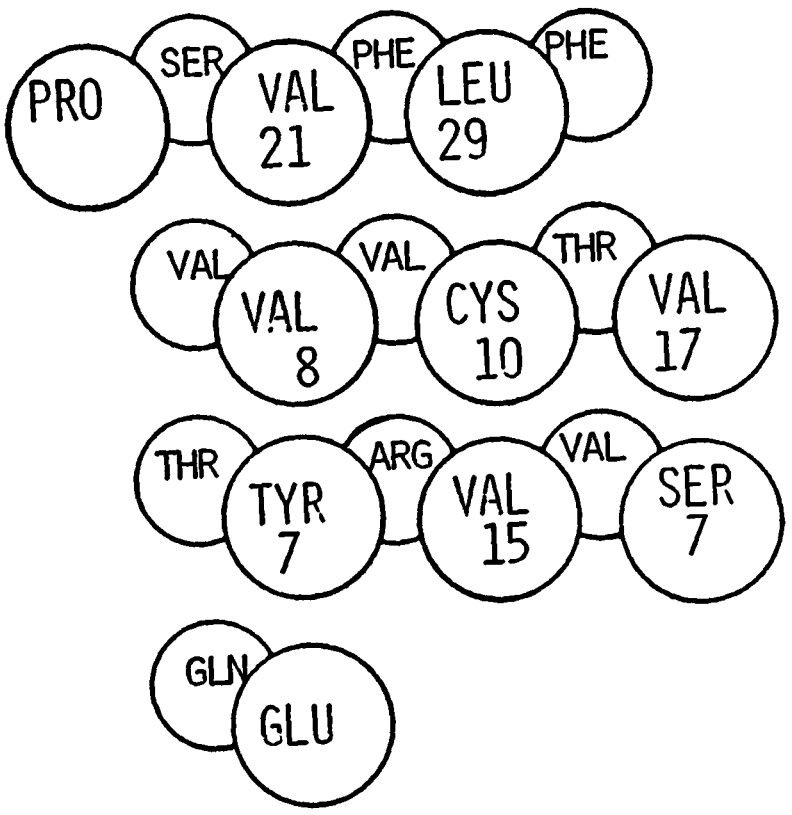
FC — CH2

A →

B ←

E →

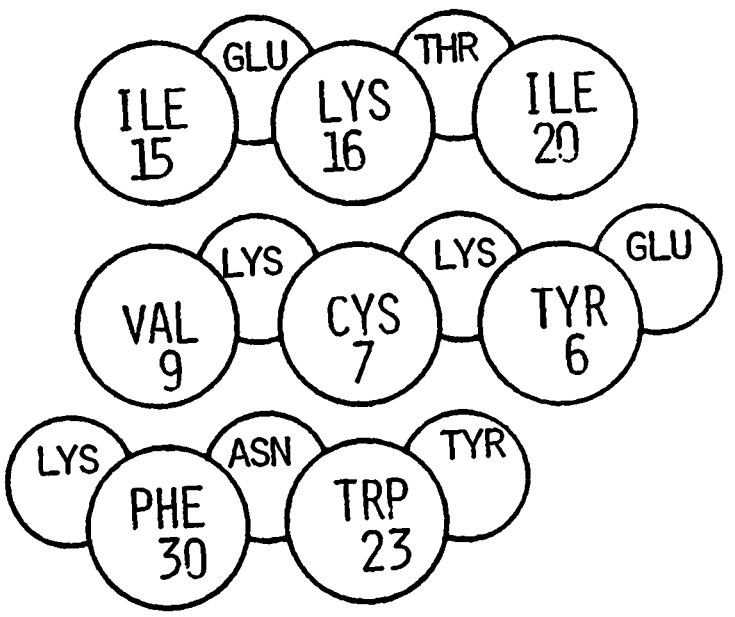
D ←



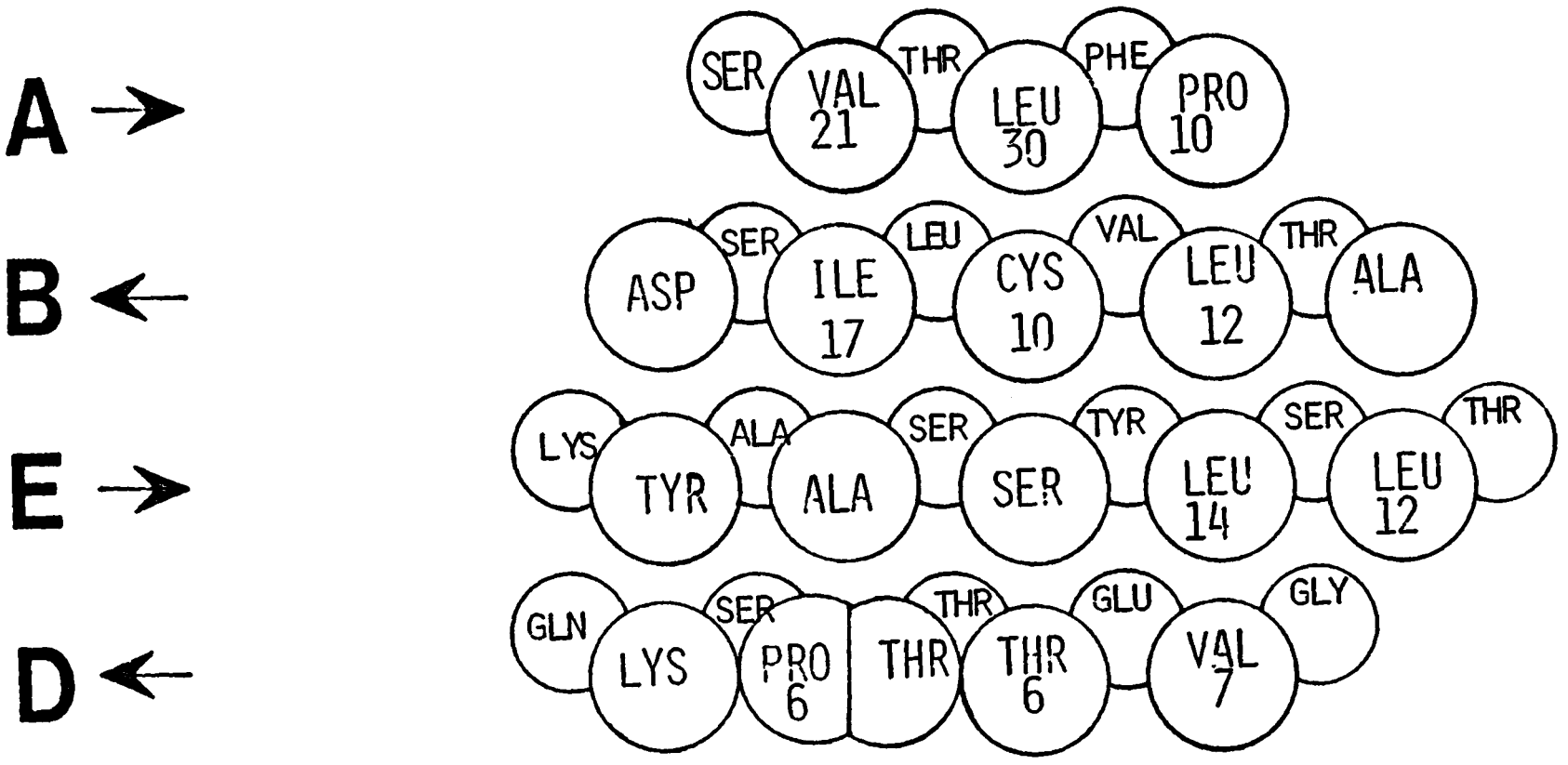
G →

F ←

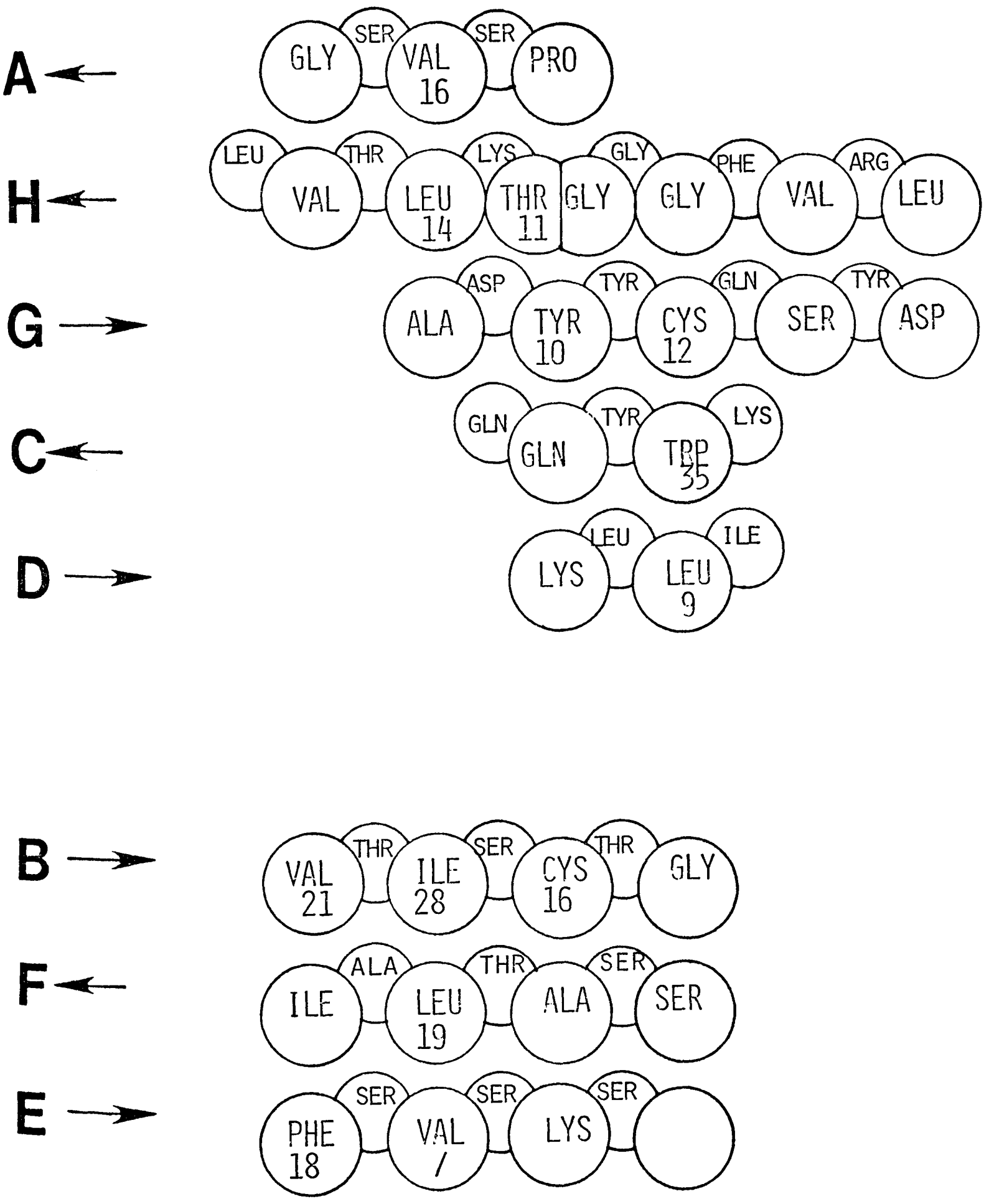
C →



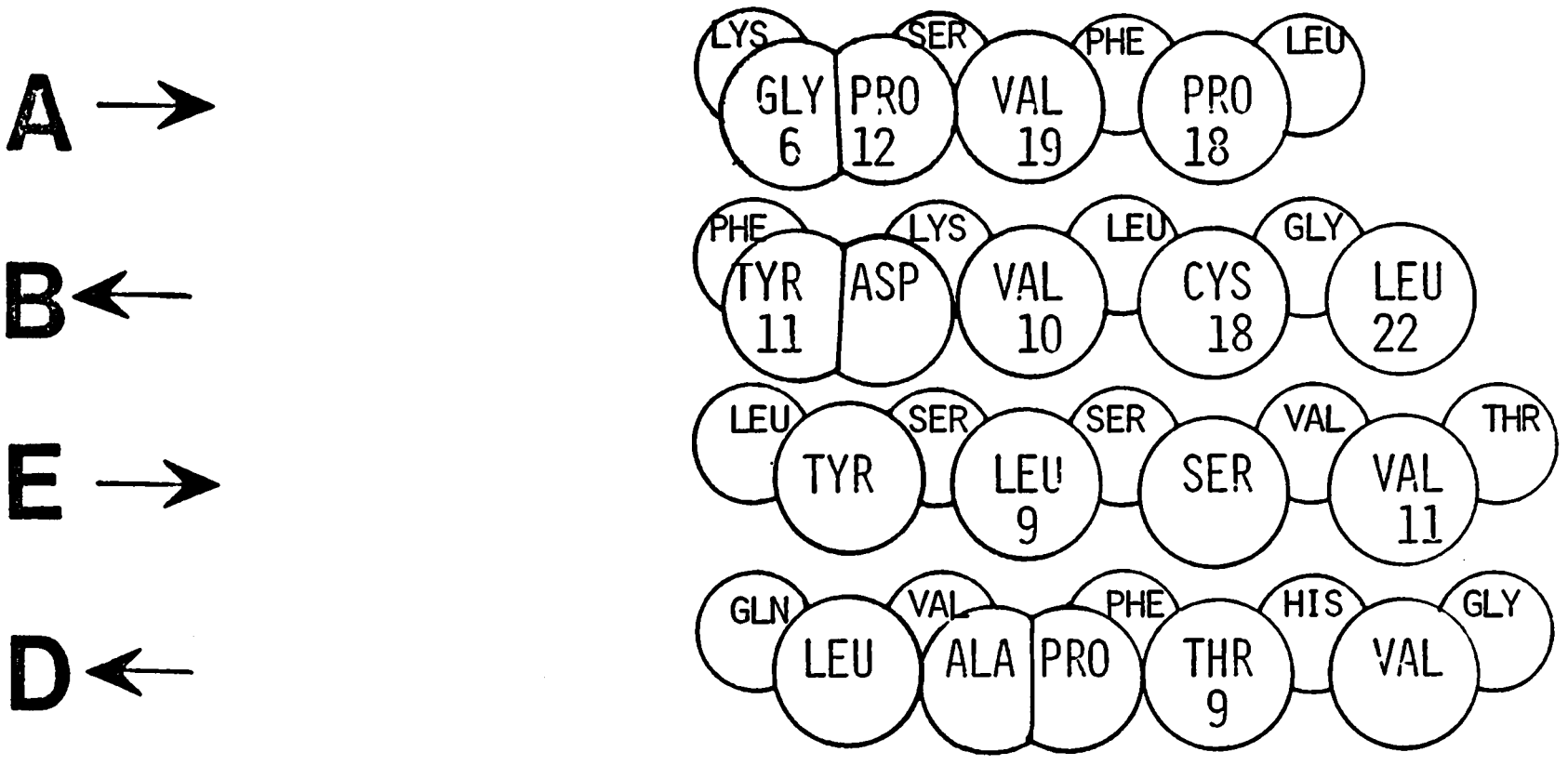
FAB LIGHT CONSTANT



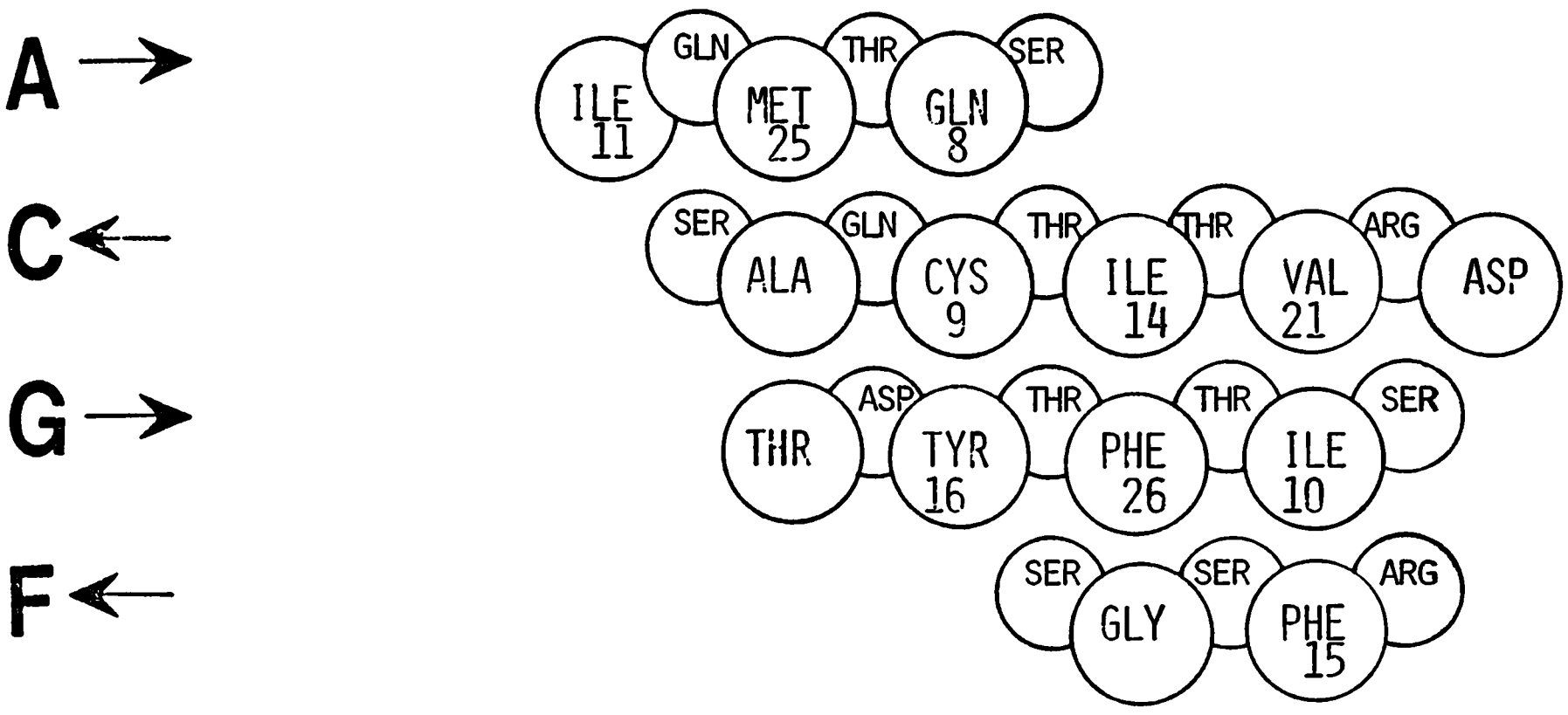
FAB LIGHT VARIABLE

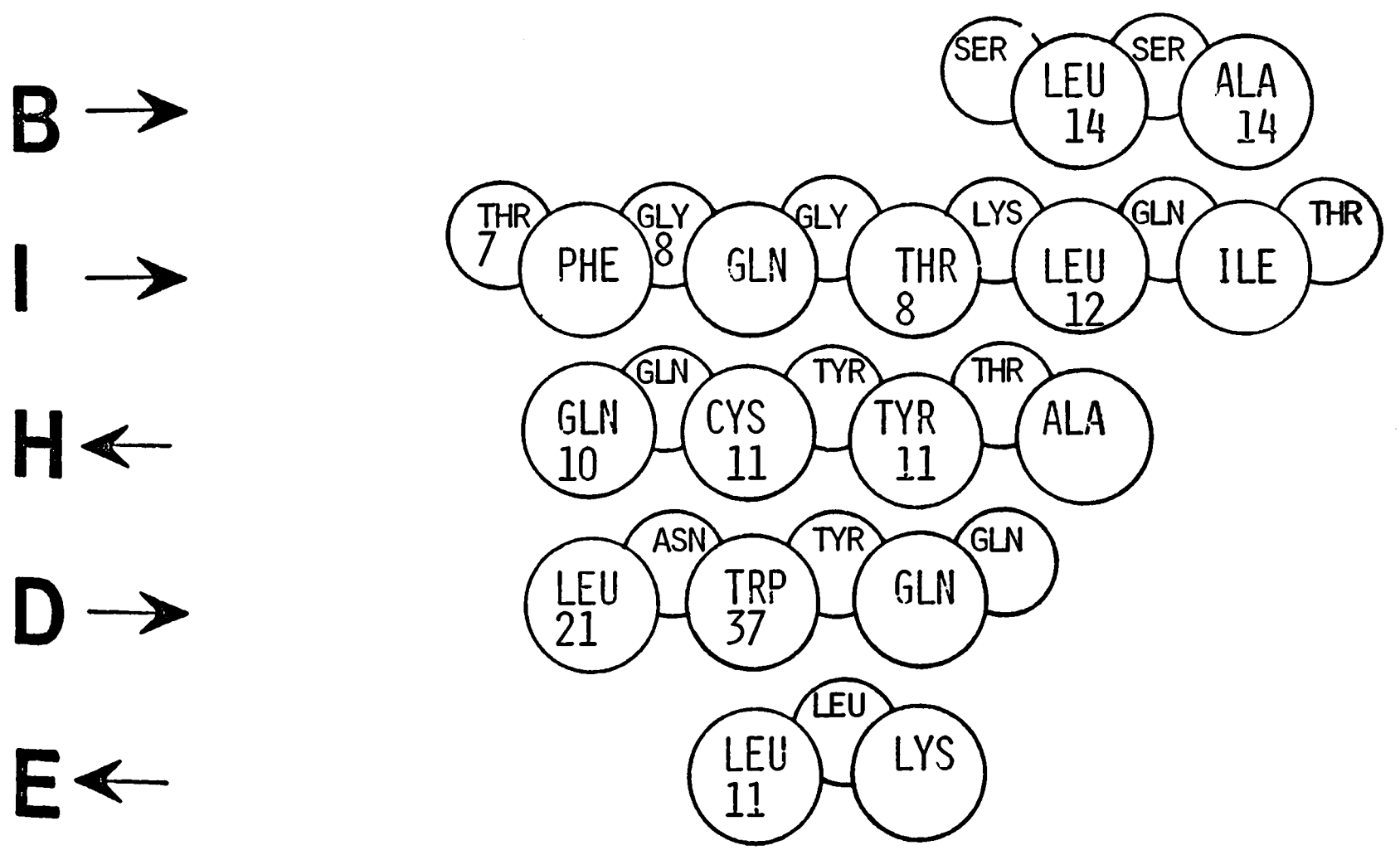


FAB HEAVY CONSTANT

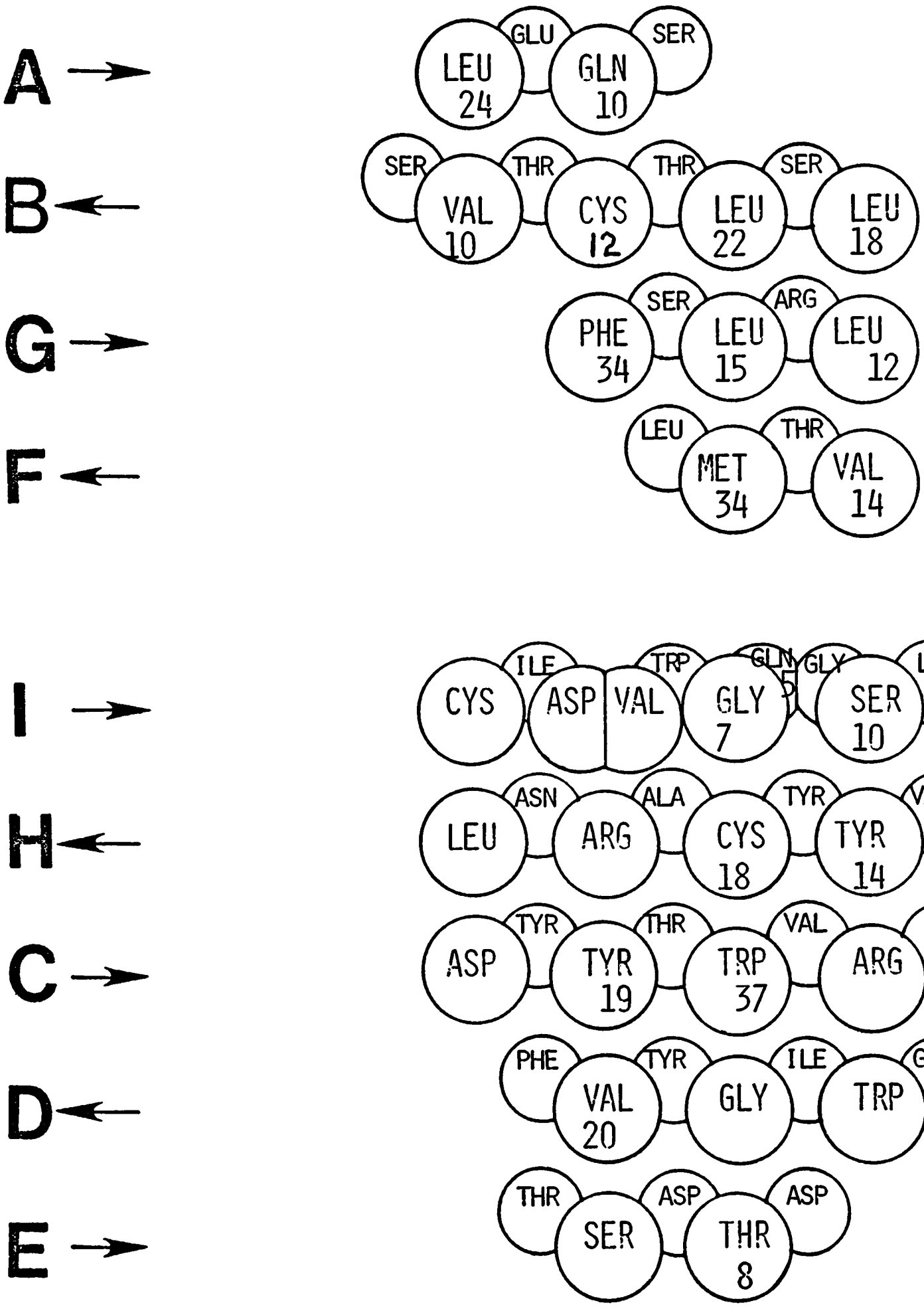


REI


A
GLN
THR
SER
ILE 11
MET 25
GLN 8
C
SER
GLN
THR
THR
ARG
ALA
CYS 9
ILE 14
VAL 21
ASP
G
ASP
THR
THR
SER
THR
TYR 16
PHE 26
ILE 10
F
SER
SER
ARG
GLY
PHE 15


B
SER
SER
LEU 14
ALA 14
I
THR 7
GLY 8
GLY
LYS
GLN
THR
PHE
GLN
THR 8
LEU 12
ILE
H
GLN
TYR
THR
GLN 10
CYS 11
TYR 11
ALA
D
ASN
TYR
GLN
LEU 21
TRP 37
GLN
E
LEU
LEU 11
LYS

FAB HEAVY VARIABLE

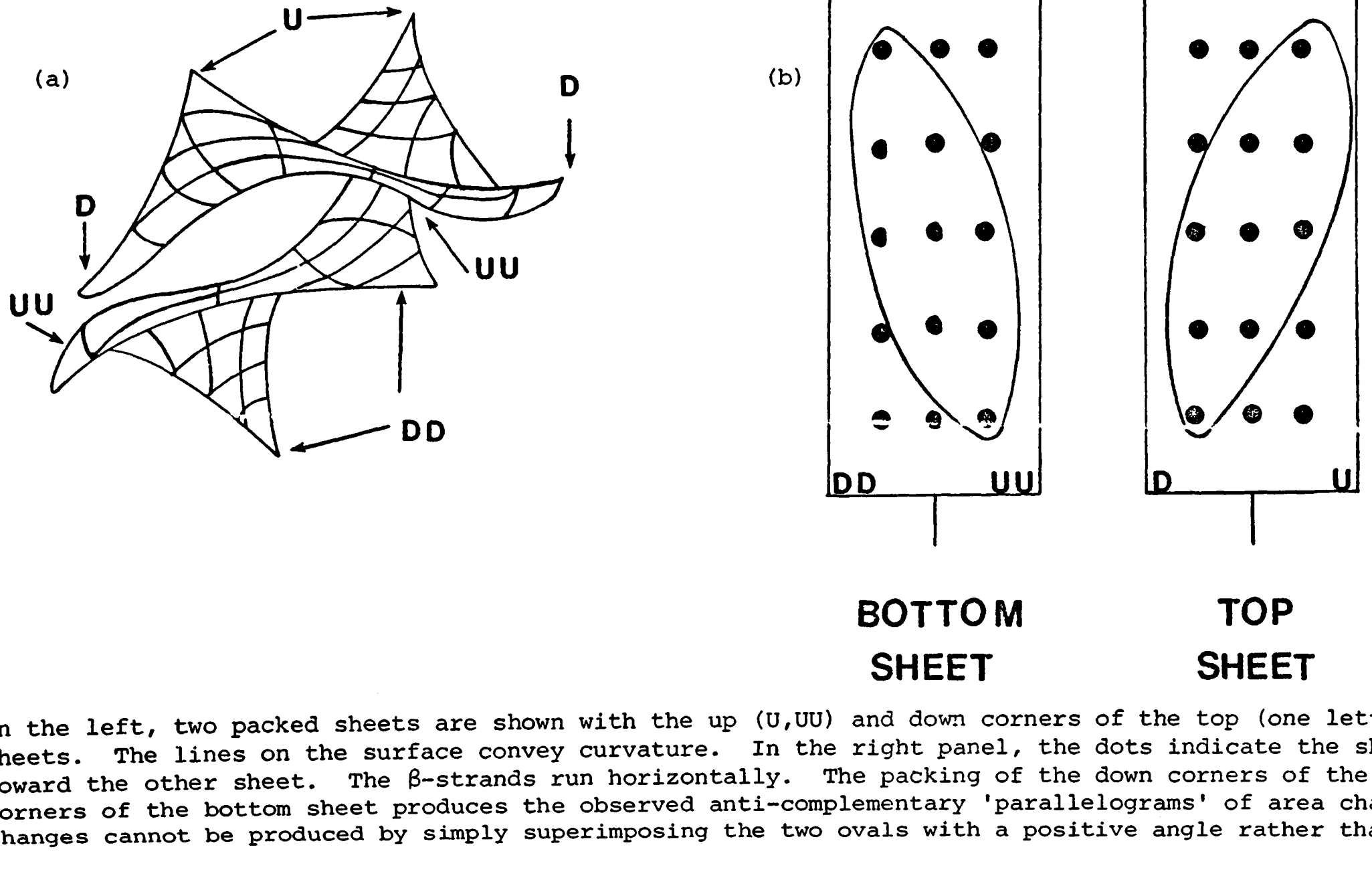


a β-sheet to have two up corners and two down corners (see Figure 4.6a). Thus the "down" residues on the bottom surface of the top sheet are likely to pack against the "up" residues on the top surface of the bottom sheet (see Figure 4.6b). This explains the observation that only 2/3 of the residues on the internal side of each sheet have substantial NPACA change ($> 5\text{Å}^2$). A survey of the nine β-sandwiches reveals that a pattern of anti-complementary parallelograms of hydrophobic residues on the interacting faces of β-sheets is common (see Figure 4.7). Moreover, residues with sizeable NPACA change ($> 5\text{Å}^2$) in the top sheet tend to progress from left to right, as one goes down the diagram, whereas in the bottom sheet they proceed from right to left.

A more detailed examination of the changes in NPACA upon the formation of a β-sandwich from two isolated β-sheets reveals that certain residues seem to be central to the sheet-sheet interaction. If three residues in a β-strand i-2, i and i+2 have significant area changes, then i, the central residue, is always Ala, Cys, Ile, Leu, Met, Phe, Pro, Trp, Tyr or Val and possibly Lys if this is an edge strand. If two or four residues on a β-strand (i-2, i) or (i-4, i-2, i, i+2) have significant area changes then either i or i-2 is a central residue restricted to the aforementioned set of amino acids. Moreover, the relative positions of the central residues in the nine β-sandwiches can be generalised into 32 allowed central residue phasings (see Table 4.4). The area changes for residues upon β-sandwich formation agree well with the values calculated from an ideal poly-cysteine β-sheet with a guest residue (see Figure 4.8).

The validity of this analysis of β-sandwiches rests on the significance of the hydrophobic effect in protein folding and the concentration of this effect in secondary structure. For prealbumin, 61% of the NPACA lost on going from isolated chain into the tetramer is concentrated in the β-sandwich residues although these residues are only half of the total number of residues (see Figure 4.9). The residues with β-structure also are responsible

FIGURE 4.6. A Schematic View of Two Twisted β-Sheets.



On the left, two packed sheets are shown with the up (U,UU) and down corners of the top (one letter) and bottom (two letters) sheets. The lines on the surface convey curvature. In the right panel, the dots indicate the sheet residues that point toward the other sheet. The β-strands run horizontally. The packing of the down corners of the top sheet against the up corners of the bottom sheet produces the observed anti-complementary 'parallelograms' of area changes. Note these area changes cannot be produced by simply superimposing the two ovals with a positive angle rather than the observed negative angle.

FIGURE 4.7. Strand-Alignment Diagrams for β-sheets

The relative positions of the strand residues are shown. β-bulges are indicated as in Pro Thr in strand D of FALC. Residues whose side chains points towards the other sheet are in capital letters. In the diagrams, observed area changes on sheet sandwiching are indicated by enclosing the residue in a rectangle if the area change is > 10Å² and by underlining if the area change is between 5 and 10Å². In the predicted diagrams, the central residues are in rectangles whilst the other residues that form the site are underlined. The relative phasings of the central residues are shown.

The potential Φ-area of a site is the sum of the available areas for the 2, 3 or 4 component non-polar residues. The area for a non-polar residue X is the available Φ-area of X in a model three-stranded twisted antiparallel sheet $(Cys)_7:(Cys_3\ X\ Cys_3):(Cys)_7$. This area agrees well with the observed change for X on sheet/sheet association. The values (Å²) we calculated are: Ala (8), Cys (11), Ile (18), Leu (19), Met (26), Phe (24), Pro (9), Trp (34), Tyr (17), Val (13) and Lys (12) if it occurs in an edge strand. Other amino acids were designated as polar and assigned zero area.

FIGURE 4.7.

FALC OBSERVED

A → ser VAL thr LEU phe PRO

B ← ASP ser ILE leu CYS val LEU thr ALA

E → lys TYR ala ALA ser SER tyr LEU ser LEU thr

D ← gln LYS ser PRO/THR thr THR glu VAL gly

G → SER thr VAL glu LYS thr VAL

F ← HIS thr VAL gln CYS ser TYR

C → VAL thr VAL ala TRP lys

FALV OBSERVED

B ← GLY thr CYS ser ILE thr VAL arg

F → SER ser ALA thr LEU ala ILE

E ← ser LYS ser VAL ser PHE

A → PRO ser VAL ser GLY

H → LEU arg VAL phe GLY gly THR/GLY lys LEU thr VAL leu

G ← ASP tyr SER gln CYS tyr TYR asp ALA

C → lys TRP tyr GLN gln

D ← ile LEU leu LYS

FAHC OBSERVED

A → lys GLY/PRO ser VAL phe PRO leu

B ← phe TYR/ASP lys VAL leu CYS gly LEU

E → leu TYR ser LEU ser SER val VAL thr

D ← gln LEU val ALA/PRO phe THR his VAL gly

G → THR lys VAL asp LYS lys VAL

F ← HIS asn VAL asn CYS ile TYR

C ← VAL thr VAL ser TRP asn

FAHV OBSERVED

A → LEU glu GLN ser

B ← ser VAL thr CYS thr LEU ser LEU

G → PHE ser LEU arg LEU

F ← leu MET thr VAL

I → CYS ile ASP/VAL trp GLY gln/gly SER leu VAL thr

H ← LEU asn ARG ala CYS tyr TYR val ALA

C → ASP tyr TYR thr TRP val ARG gln

D ← phe VAL tyr GLY ile TRP glu LEU

E → thr SER asp THR asp

TABLE 4.4

32 Allowed Central Residue Phasings

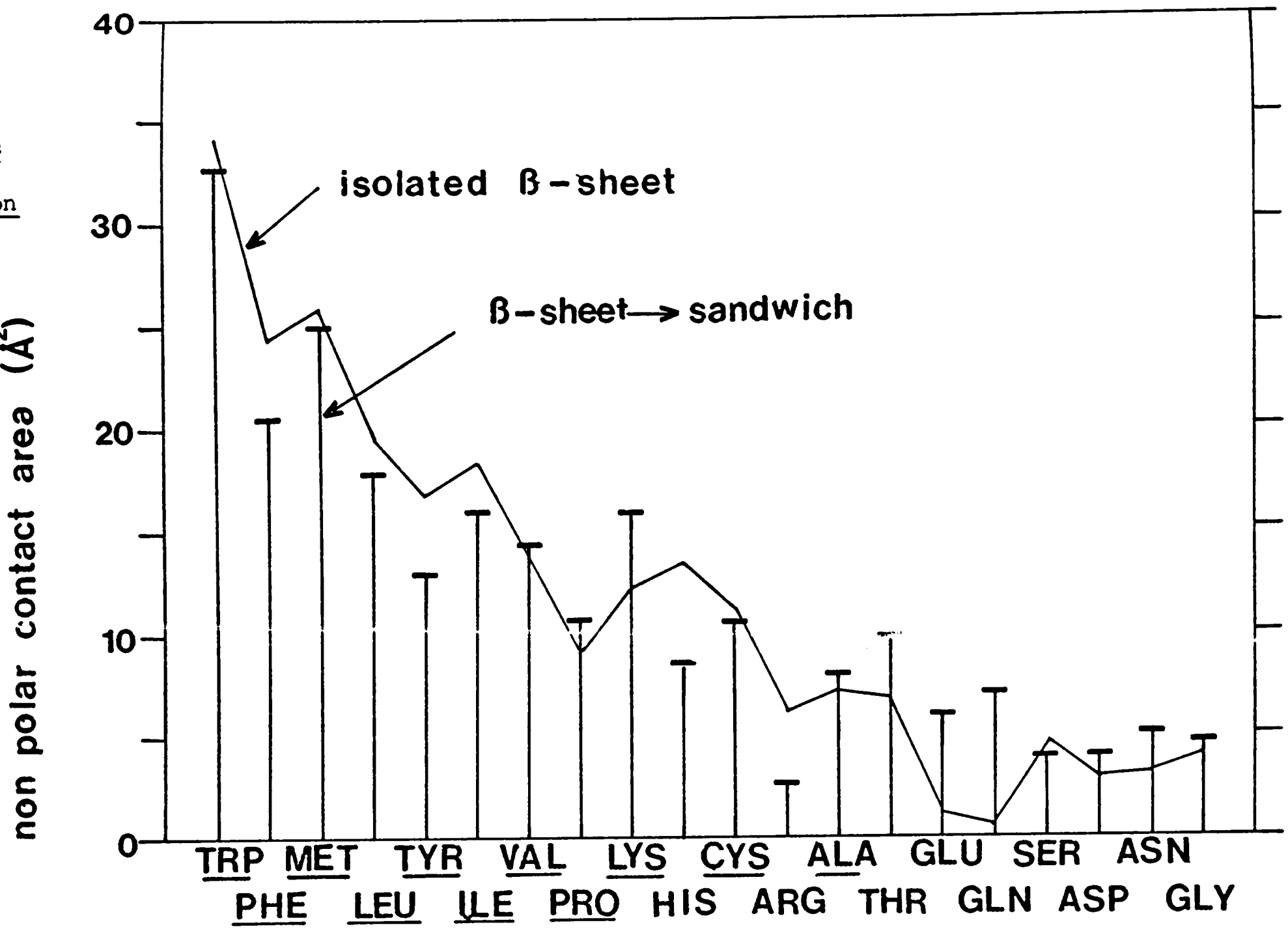
Top Sheet	Bottom Sheet	Top Sheet	Bottom Sheet
-2,0,+2	+2,0,-2*	0,0,+2	+2,0,-2
	+2,0,0		+2,0,0
	0,0,-2		0,0,-2
	0,0,0		0,0,0
	-2,0,0		-2,0,0
	0,0,+2		0,0,+2
-2,0,0	+2,0,-2	0,0,0	+2,0,-2
	+2,0,0		+2,0,0
	0,0,-2		0,0,-2
	0,0,0		0,0,0
	-2,0,0		-2,0,0
	0,0,+2		0,0,+2
0,0,-2	+2,0,-2	+2,0,0	+2,0,-2
	+2,0,0		+2,0,0
	0,0,-2		0,0,-2
	0,0,0		0,0,0

* Relative position of the central residues for the 3 strands in a sheet, e.g. +2,0,-2 is j+2, k, l-2 where j,k,l are aligned in the sheet.
-2,0,+2, -2,0,0, 0,0,+2, 0,0,0 have correct +ve shift for top sheet.
0,0,-2 and +2,0,0 have wrong shift.
+2,0,-2, +2,0,0, 0,0,-2, 0,0,0 have correct shift for bottom sheet.
-2,0,0 and +2,0,-2 have wrong shift.
Pairings: Top O.K. Bottom O.K.
Top not O.K. Bottom O.K.
Top O.K. Bottom not O.K.

FIGURE 4.8.

Contact Area Changes

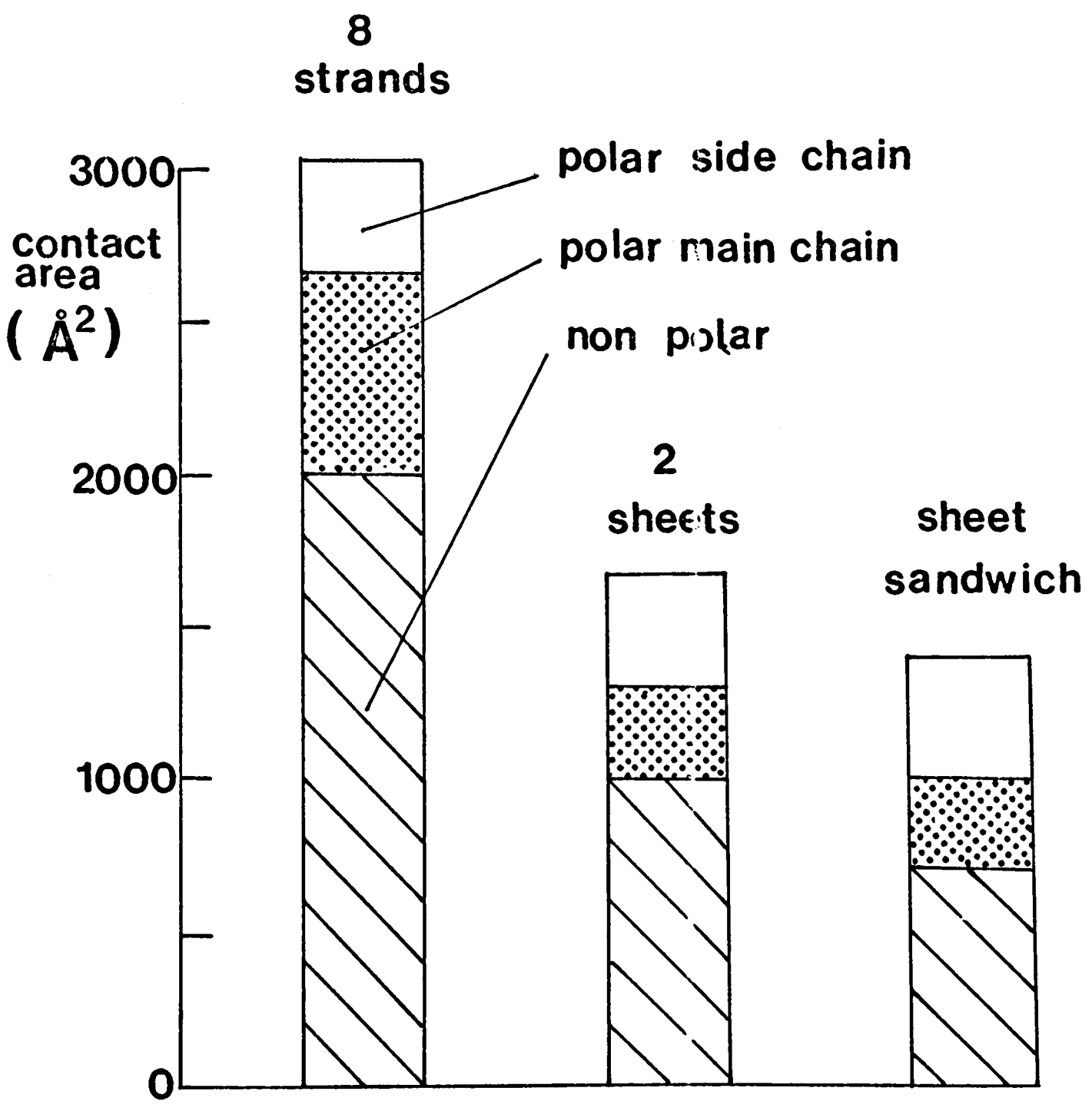
on Sandwich Formation



The non-polar area of a residue X in a β-sheet was calculated from a model system of a 3-stranded, twisted antiparallel β-sheet: $(Cys)_7$:(Cys_3 X Cys_3):$(Cys)_7$. These values are connected. For each residue, the mean observed change of non-polar contact area on mediating the formation of a sandwich from a β-sheet is plotted by a vertical line. Practically all the accessible contact area of the residue in a sheet is buried when two sheets pack. Residues underlined can contribute to the non-polar interaction site, Lys being allowed only if the β-strand is at the edge of the sheet.

FIGURE 4.9.

A detailed breakdown of changes in non-polar accessible contact area for prealbumin in going from isolated β-strands to 2 β-sheets to the β-sandwich.



for 61% of the NPACA lost on monomer formation.

If NPACA changes are important to determining β-structure, then these changes should show some agreement with the Chou & Fasman (1974) β propensity $\langle P_\beta \rangle$ of an amino acid. A plot of $\langle P_\beta \rangle$ against the change in NPACA upon the formation of a β-sheet from isolated strands suggests some correlation, but that Pro, Glu and Lys are exceptions (see Figure 4.10). The fact that proline cannot form one of two possible hydrogen bonds may explain why NPACA changes overestimate its β propensity. Similarly the charges on glutamic acid and lysine must complicate this simple one parameter model.

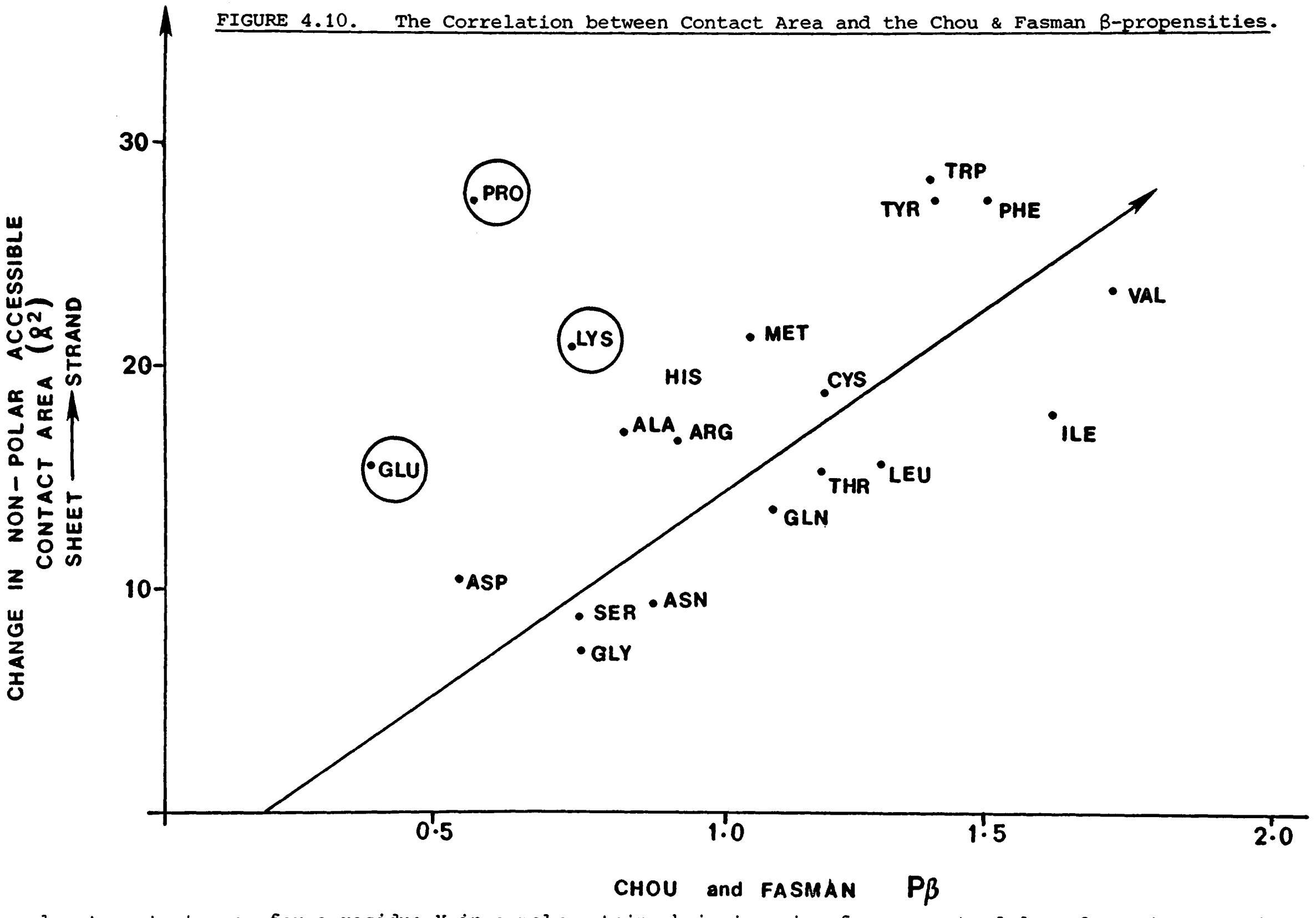
4 An Algorithm to predict the Structure of β-sandwiches

The goal of this analysis of β-sandwiches was to provide insight into the constraints which limit the myriad of possible spatial juxtapositions of the residues along the polypeptide chain. Equation 13 of Chapter 2 suggests that there are $\sim 3 \times 10^{17}$ compact structures for a chain with 100 residues. However, the distinct pattern of hydrophobic residues along the internal faces of the β-sheets in the β-sandwich suggests a geometric construction for structures which might adequately partition hydrophobic residues from the solvent. Of course, many of these geometrically reasonable structures might not satisfy other topological and steric constraints. This section describes an algorithm which endeavours to predict the structure of all-β proteins through a combinatorial sampling of a large region of conformation space. Filters are applied to reduce the number of acceptable structures. A flow diagram of this procedure is shown in Figure 4.11.

1) Locate Hydrophobic patches on the β-strands.

The input to this algorithm is the amino acid sequence and an observed (or postulated) β-strand assignment. The first step is to predict not only which surface of each β-strand will point towards the other sheet but also

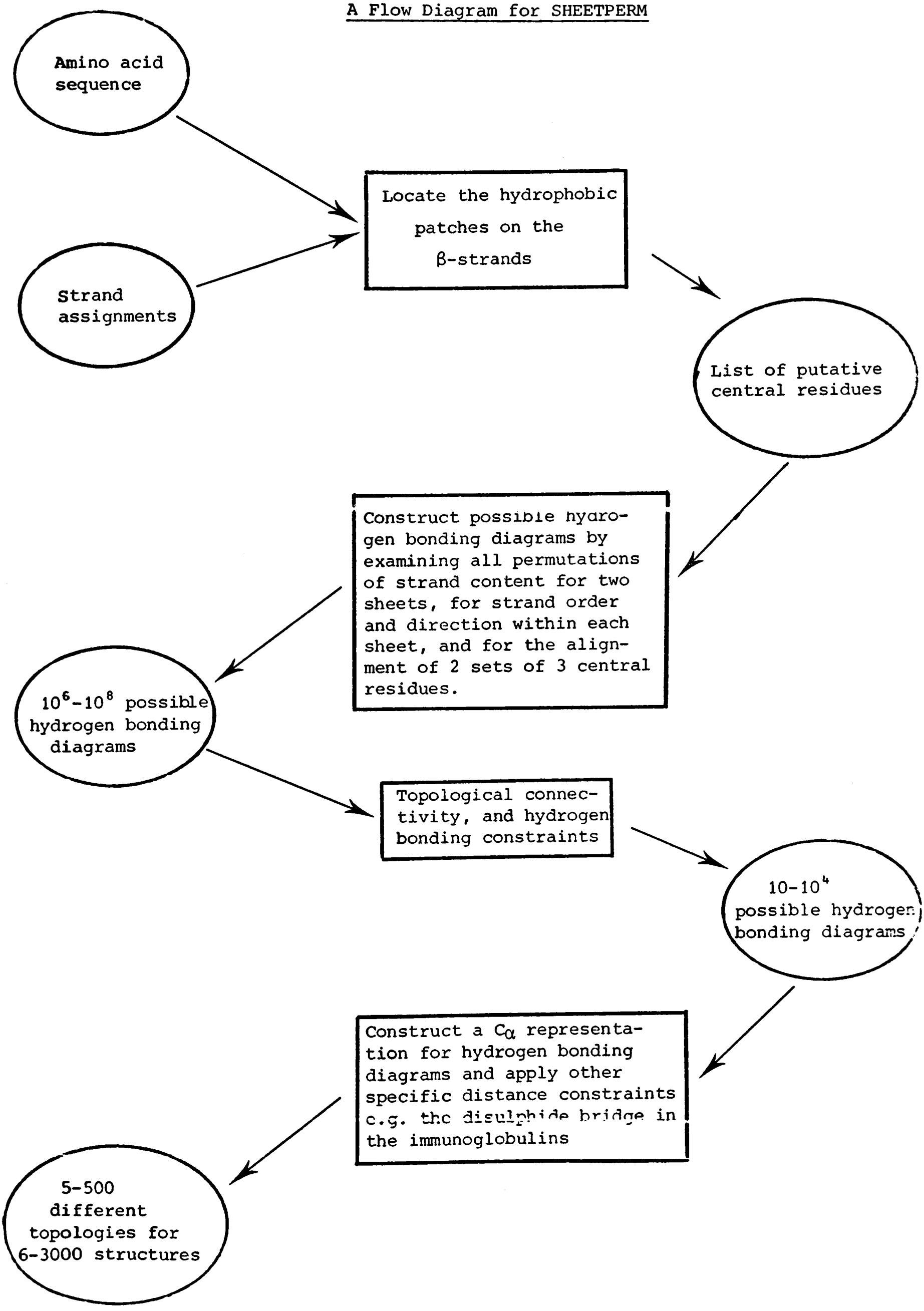
FIGURE 4.10. The Correlation between Contact Area and the Chou & Fasman β-propensities.



Change in solvent contact area for a residue X in a polycystein chain in going from an extended conformation to a 3-stranded anti-parallel β-sheet as a function of the Chou & Fasman parameter P_β. Three residues, Glu, Pro & Lys have much smaller P_β than this simple contact area model would suggest.

FIGURE 4.11

A Flow Diagram for SHEETPERM



the location along the strand of the few residues that mediate the hydrophobic interaction. This was achieved by scanning each strand for a continuous patch of non-polar residues on one of its sides, e.g. 3 residues at positions i-2, i, i+2. The location of this sheet/sheet interaction site is represented by a central residue at position i. To distinguish between alternate central residues we consider first the number of non-polar residues in the patch (4, 3 and 2) and then the magnitude of their possible hydrophobic contribution to sheet/sheet stacking in terms of the available NPACA. The potential NPACA change for a site is the sum of the available areas for the 2, 3 or 4 component non-polar residues. The area for a non-polar residue X is the available NPACA of X in a model, three-stranded twisted antiparallel sheet $(Cys)_7:(Cys_3\ X\ Cys_3):(Cys)_7$ as this area agrees well with the observed change for X on sheet/sheet association. The values ($Å^2$) are: Ala (8), Cys (11), Ile (18), Leu (19), Met (26), Phe (24), Pro (9), Trp (34), Tyr (17), Val (13) and Lys (12) if it occurs in an edge strand. Other residues were designated as polar and assigned zero area. This yields a set of one or two possible central residues for each β-strand. Two possible central residues result when there are an even number of non-polar residues in the patch or if Lys is involved (see Figure 4.12).

2) Construct possible β-sheet hydrogen-bonding diagrams.

The next step is to place the β-strands in the two sheets. As not all of the strands contribute markedly to the sheet-sheet interaction, the β-sandwich was modelled as a central hydrophobic core of two 3-stranded β-sheets. Additional strands are to be placed once the central core has been predicted. For a protein with 8 strands, there are $\frac{1}{2}\binom{8}{3}\binom{5}{3}$‡ possible ways of sorting the strands into two 3-stranded sheets. Within each sheet, there are 6 strand orders and 2^3 strand directions. Coupled with the 32 allowed central residue phasings that generalise the observed anticomplementary

‡ $\binom{n}{k} = n!/(n-k)!k!$

FIGURE 4.12

Predicted Strand Alignment Diagram

FALC PREDICTED

```
A → O                ser VAL thr [LEU] phe PRO
B ← O/+2         ASP ser ILE leu [CYS] val [LEU] thr ala
E → +2       lys TYR ala ALA ser SER tyr [LEU] ser LEU thr
D ←          gln LYS ser [PRO]/ALA thr THR glu VAL gly

G → O            SER thr VAL glu [LYS] thr VAL
F ← O            HIS thr VAL gln [CYS] ser TYR
C → -2           VAL thr [VAL] ala TRP lys
```

The predicted strand alignment for the light chain constant domain fragment of the Immunoglobulin IgG(λ) new. Residues on the internal faces of the β-sheets are shown in capital letters. Putative central residues on the sheet are boxed and ancillary hydrophobic residues in the patch are underlined. The relative phasing of the central residues is indicated next to the strand position and direction.

parallelograms of the hydrophobic patches (see Table 4.4), this means that at least $\frac{1}{2}\binom{8}{3}\binom{5}{3}.(6.2^3)^2.32$ or 2×10^7 possible hydrogen-bonding diagrams are considered. Figure 4.12 shows FALC with the set of central residue phasings which produced that hydrogen-bonding diagram when A,B and E were in the top sheet and G,F and C were in the bottom sheet with A+, B-, E+, G+, F- and C+*.

From each central core, a set of complete structures was constructed with the previously unplaced strands being located in every possible unoccupied position with every direction. No attempt was made to phase the unplaced strands with respect to the central core. Clearly, a different approach is required for the two 7-stranded sheets in CONA.

Although this may seem to be an unwieldy number of structures to consider, it represents a well-structured reduced subset of the $\sim 3 \times 10^{17}$ compact conformations. Moreover, these structures can be arranged in a tree hierarchy and generic branches of structures can be eliminated.

3) Constraints on allowed structures.

The list of generated structures is filtered by imposing the following restrictions that quantify observed topological (i, ii, iii), steric (iv) and hydrogen-bonding (v, vi) features of known β-sandwiches.

(i) The connection between two parallel β-strands in the same sheet is right handed (Sternberg & Thorton, 1977a; Richardson, 1976). Thus in REI (see Figure 4.13) strands B and D are in the bottom sheet, C is in the top sheet, and the bottom sheet is oriented so that B runs up the page, then since B is parallel to D, D must be on the right hand side of B.

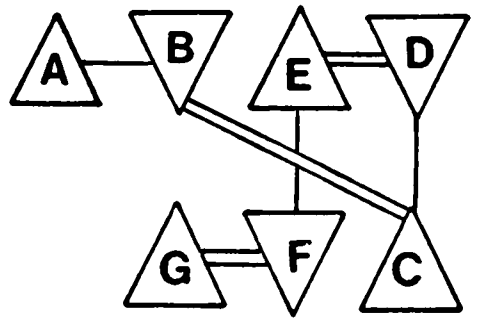
(ii) Connections between β-strands do not cross (Ptitsyn et al., 1976). Figure 4.13 shows that the doubled lines which represent loops on the near edge of the β-sandwich never cross each other nor are there

* + indicates that the strand runs from left to right across the page, and - from right to left.

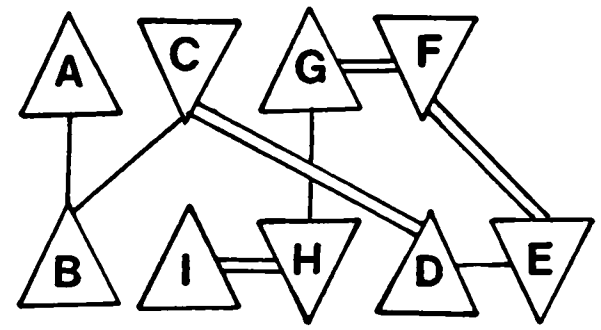
FIGURE 4.13.

Schematic Diagrams of β-sandwich Proteins

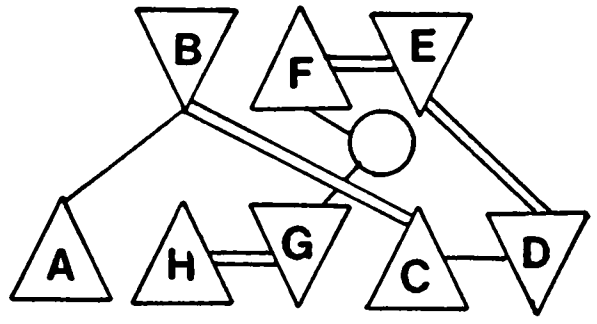
Schematic diagrams of all the β-sheet sandwiches considered. Each β-strand is viewed along its strand direction and is represented by a triangle whose apex points up or down according to whether the strand is viewed from the N- or C-terminus. A circle represents an α-helix. The connections are shown in double lines if they start at the end of the sheet close to the viewer.



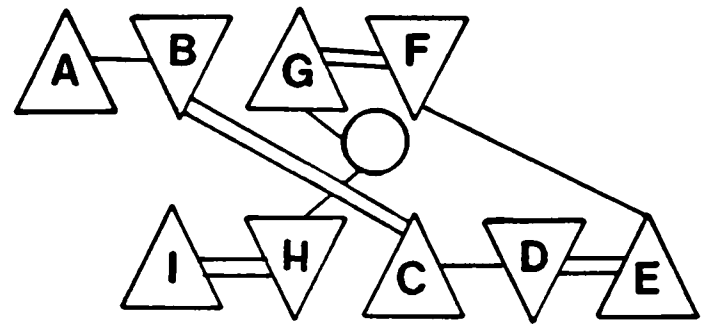
FALC, FAHC, FCH2, FCH3



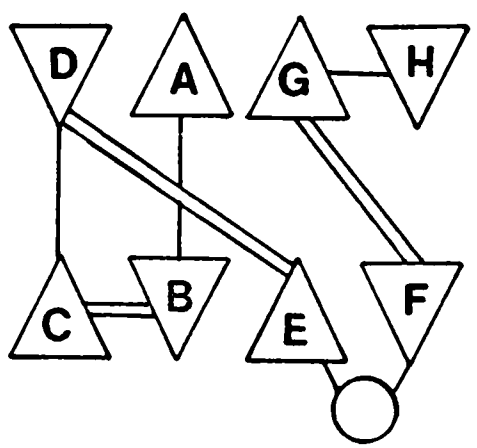
REI



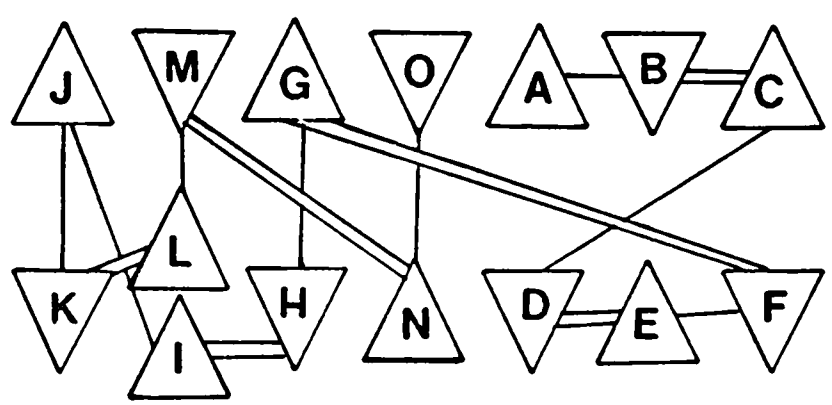
FALV



FAHV



PRE



CONA

crossovers for the loops on the far edge represented by single lines.

(iii) The topology of strands includes a generalised 'Greek key'. This pattern has been noted as a common structural motif in all-β proteins (Richardson, 1977; Ptitsyn et al., 1979). Each sandwich topology is cyclised into a barrel and all possible splices of the barrel into a planar representation are examined. At least one of these strand orders must be j... j+3... j+2, j+1 with anti-parallel connections between j and j+1, j+1 and j+2, and j+2 and j+3. For example, super oxide dismutase would have the planar representation ABCFEDGH. The Greek key would be formed by strands C,D,E and F.

(iv) There are sufficient residues between the strands to make the required connection length. A connection is not permitted to be parallel if there are more than 4 residues in each of two consecutive β-strands and less than 12 residues in the connecting loop. For anti-parallel connections, the minimum number of residues required to join the C and N termini of consecutive strands is equal to the relative stagger between these endpoints in the hydrogen bonding diagram plus two if this is an antiparallel connection between the two sheets and plus five if this is a parallel connection. These values are added to account for the steric problem encountered in intersheet and parallel connections.

(v) Given the strand composition of the sheets but not the strand positions, structures must contain no fewer than one less than the maximum number of hydrogen bonds that could possibly be formed. This lower bound is computed as the sum of the number of residues in the two shorter strands minus one.

(vi) The two β-sheets in the central core must have a high degree of strand overlap. This notion was quantified as the number of potential main-chain hydrogen bonds formed as a fraction of the number of main-chain nitrogen (or oxygen) atoms in the sheet. The significance of

strand overlap has been discussed previously. Structures allowed by filters (i) to (v) were rank ordered on strand overlap.

4) Construction of a C^{α} representation and Application of Specific Distance Constraints.

From a predicted sheet diagram for the central core, suitable C^{α} co-ordinates can be obtained. Each β-sheet was built by aligning idealised twisted β-strands based on the positions of the central residues. The two sheets were sandwiched by aligning the central residues of the middle strands and placing the sheets at their typical separation (10Å) and rotation (-30°). The coordinates were then examined to see if the C^{α} atoms of the two cysteine residues involved in the immunoglobulin intrachain disulphide bridge were within 15Å of each other. In addition, this disulphide bridge is not allowed between cystine residues in non-adjacent β-strands in the same β-sheet. This conclusion is based on a general analysis of disulphide bridges in proteins (J.M.Thorton, personal communication).

Two Fortran programs were written to perform the calculations described. SHEETPERM handles parts 1, 2 and 3 i, ii, iv-vi. DRAWIBM codes for the remaining steps (see Appendix IV). These programs were implemented on an IBM 360/195 and the total CPU time required for the nine examples considered was ∿20 hours.

5 Results of the Prediction of β-sandwiches

SHEETPERM and DRAWIBM were uniformly applied to the nine β-sandwiches: FCH2, FCH3, FALC, FAHC, PRE, SDM, FAHV, FALV & REI. From a starting set of over 10^7 structures, a reduced list of allowed β-sandwiches rank ordered on strand overlap is produced. The reduction is typically five or six orders of magnitude (see Table 4.5). For each of the nine β-sandwiches, one member of the reduced list has all or most of the relative residue posi-

Structure	No. of strands	No. of possible structures from all strands	No. of structures generated from strands with sites	Position in list of strand overlap of native structure	r.m.s.d. predicted/ crystal structure Å	No. of residues
(1)	(2)	(3)	(4)	(5)	(6)	
FCH2	7(6)	5×10^7	6×10^5	283 (137)	2.3	34
FCH3	7(7)	5×10^7	3×10^6	268 (55)	4.9	50
FALC	7(7)	5×10^7	3×10^6	10 (5)	1.4	46
FAHC	7(7)	5×10^7	3×10^6	6 (6)	3.2	48
PRE	8(7)	2×10^8	3×10^6	10 (8)	2.0	51
FALV	8(8)	2×10^8	1×10^7	2907 (438)	4.9	48
SDM	8(8)	2×10^8	1×10^7	795 (190)	5.1	51
FAHV	9(7)	6×10^8	3×10^6	481 (91)	3.5	46
REI	9(7)	6×10^8	3×10^6	937 (150)	3.6	48
B2-M	7(6)	5×10^7	6×10^5	35 (7)	-	53
AC-2	7(6)	5×10^7	6×10^5	93 (5)	-	48

Table 4.5 **Prediction of β-sheet sandwiches**

Legend to Table 4.5.

(1) See section 4.2 for abbreviations.

(2) The number of β-strands in the sandwich with the number of strands with sites in brackets.

(3) For n strands, the number of topologies for two 3-stranded sheets is $(^{n}C_3)(^{n-3}C_3)\ 3!3!2^3 2^3/4$, where $(^{n}C_r) = n!/(n-r)!r!$. For each topology we estimate that there are 625 (i.e. 5^4) alternative sheet structures that have a high degree of strand overlap.

(4) A lower estimate of the number of sheet structures generated from strands with sites placed with the 32 different phasings. In the top sheet the central residues can be phased in 6 ways: -2,0,+2; -2,0,0; 0,0,+2; 0,0,0; +2,0,0; 0,0,-2. In the bottom sheet the 6 phasings are: +2,0,-2; +2,0,0; 0,0,-2; 0,0,0; -2,0,0; 0,0,+2. Every pairing is allowed except for either of the last two phasings in the top sheet with the last two in the bottom sheet as both these phasings have the wrong direction.

(5) The number of strand-alignment diagrams (with the number of different topologies in brackets) that have the same or higher overlap than the chosen approximation to the native.

(6) The root mean square deviation between equivalenced C^{α} atoms in the native and predicted 6-stranded core. Thus accurate prediction of the strand-alignment diagram is sufficient to yield a C^{α} structure close enough to the native that energy minimization could be applied.

tions in the native sheet strand alignment diagram (e.g. see Figure 4.14). The deviations are generally caused by the presence of β-bulges (Richardson, 1978) that disrupt the register of some residues in a β-strand with respect to the central core. Apart from β-bulges, seven of the nine sandwiches are correctly built. The exceptions are in FCH2 where one strand is shifted by two residues, and in FAHV where one strand is one residue away from the correct alignment. A more complete discussion of the similarities and discrepancies between the best predicted and native structures is included in the legends to the stereo diagrams of the best predicted structure overlaid on the native (see Figure 4.15).

A strand alignment diagram provides a useful intermediate for an algorithm to relate sequence to structure. These diagrams contain sufficient information to construct a C^{α} representation for the protein that is generally close to the native. The r.m.s. deviation between equivalenced C^{α} atoms in the native and best predicted structures ranges from 1.4Å to 4.9Å with 34 to 51 residues being placed. A random prediction of a compact structure for pancreatic trypsin inhibitor with 57 residues would have a mean r.m.s. deviation from the native of 11.9 ± 1.5Å (Cohen & Sternberg, 1980b).

The general success of a combinatorial approach hinges on the quality and ease of implementation of various filters which eliminate certain structures from further consideration. Table 4.5 indicates the selectivity of the various filters by giving the number of central core structures that have no less than the strand overlap of the best predicted structure. This is an upper limit on the number of structures which would have to be sampled to locate a good approximation to the native structure. For four of the nine sandwiches, the algorithm is highly selective and <50 alternative structures (<20 different topologies) would need to be surveyed to locate the central core of the native structure. The algorithm is fairly selective (<3000 structures, <500 topologies) for the other sandwiches.

FIGURE 4.14

Strand Alignment Diagrams for FALC - Observed and Predicted

FALC OBSERVED

A → ser [VAL] thr [LEU] phe [PRO]

B ← ASP ser [ILE] leu [CYS] val [LEU] thr ALA

E → lys TYR ala ALA ser SER tyr [LEU] ser [LEU] thr

D ← gln LYS ser PRO/THR thr THR glu VAL gly

G → SER thr VAL glu [LYS] thr [VAL]

F ← HIS thr [VAL] gln [CYS] ser [TYR]

C → [VAL] thr [VAL] ala [TRP] lys

FALC PREDICTED

A → O ser VAL thr [LEU] phe PRO

B ← O/+2 ASP ser ILE leu [CYS] val [LEU] thr ala

E → +2 lys TYR ala ALA ser SER tyr [LEU] ser LEU thr

D ← gln LYS ser [PRO]/ALA thr THR glu VAL gly

G → O SER thr VAL glu [LYS] thr VAL

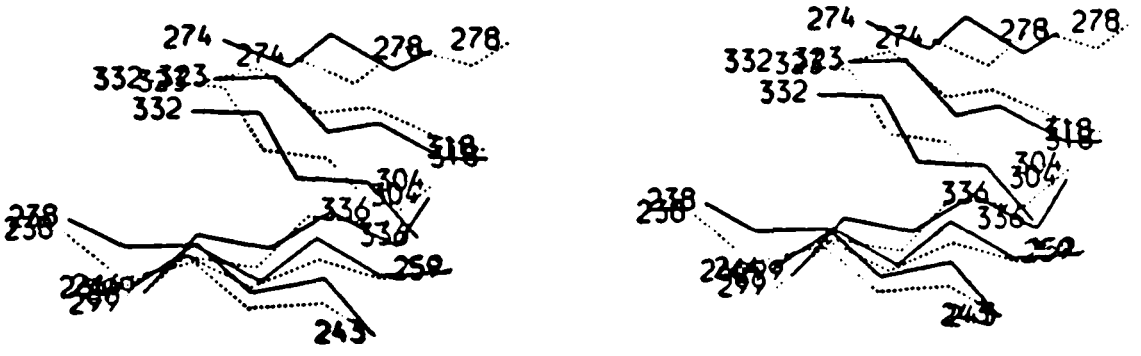
F ← O HIS thr VAL gln [CYS] ser TYR

C → -2 VAL thr [VAL] ala TRP lys

The relative positions of the residues are shown. β-bulges are indicated as in Pro Thr of strand D. Resiudes whose side chains point towards the other sheet are in capital letters. In the top panel, observed NPACA changes on sheet sandwiching are indicated by a box if the area change is >10Å² and by a line if the area change is >5Å² but <10Å². In the lower panel, predicted central residues are boxed and ancillary hydrophobic residues which contribute to the patch are underlined. The relative phasing of central residues used to construct the predicted strand alignment accompanies the sequential strand position and strand direction.

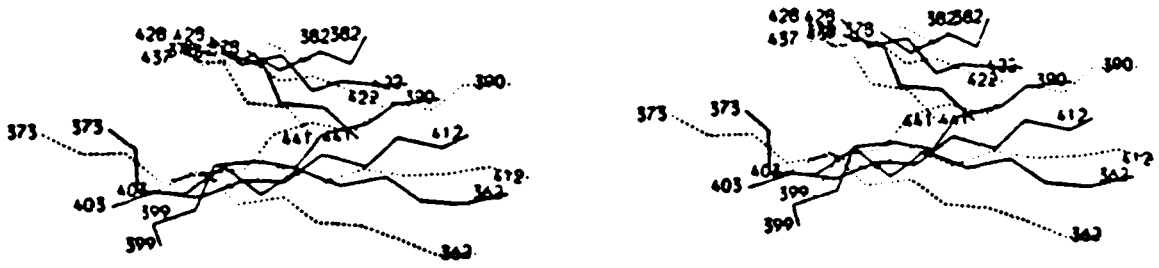
FIGURE 4.15. Predicted Structures for Nine β-Sandwiches

Stereo diagrams of the predicted structure (in dotted lines) which best approximates the crystal structure overlaid on the crystal structure (in solid lines) for the nine β-sheet sandwiches are shown.



Human Immunoglobulin F_c fragment C_H2 (FCH2)

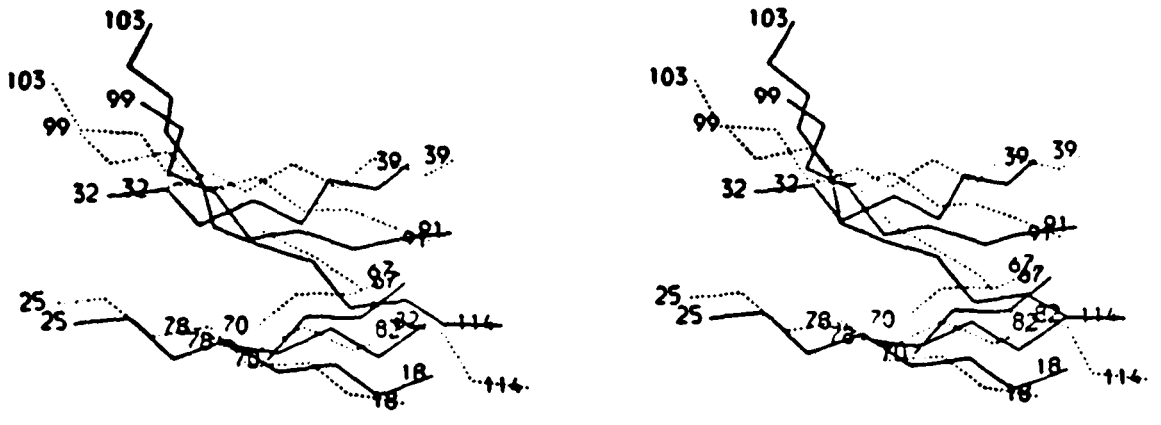
The r.m.s. deviation between these two structures is 2.3Å for the 34 residues constituting the six core β-strands. The major error in the predicted structure is that strand C (residues 274-278) is two residues out of phase. This structure survives as one of 283 alternatives.



Human Immunoglobulin F_c fragment C_H3 (FCH3)

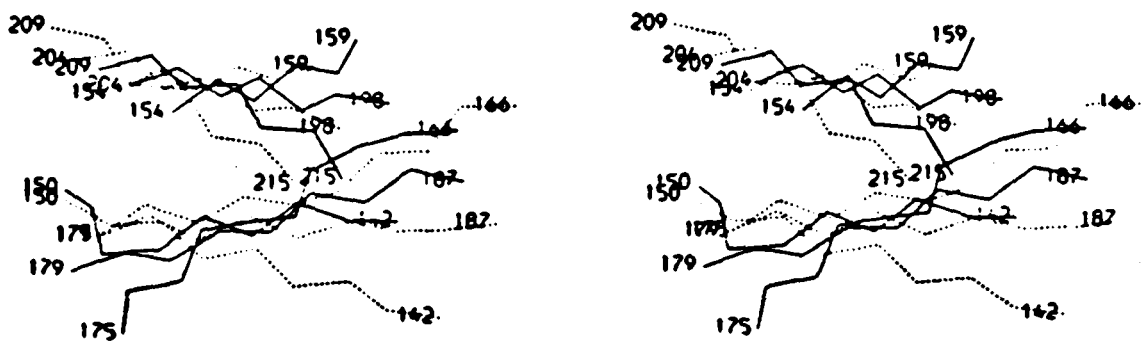
The r.m.s. deviation between these two structures is 4.9Å for the 50 residues constituting the six core β-strands. β-bulges in strands A (residues 390-399) and B (residues 403-412) are the major source of the errors in the predicted structure. This structure survives as one of 55 alternatives.

FIGURE 4.15, cont.



Human Immunoglobulin IgG(λ) new fragment of the Variable domain of the heavy chain (FAHV).

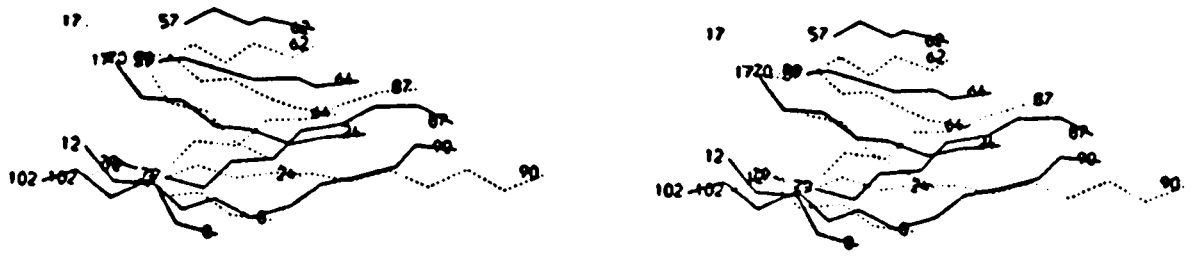
The r.m.s. deviation between these two structures is 3.5Å for the 46 residues constituting the six core β-strands. Strand H (residues 91-99) is one residue out of phase as the exterior face of this strand is more hydrophobic than the interior face. This structure remains as one of 481 alternatives.



Human Immunoglobulin IgG(λ) new Fragment of the Constant domain of the heavy chain (FAHC)

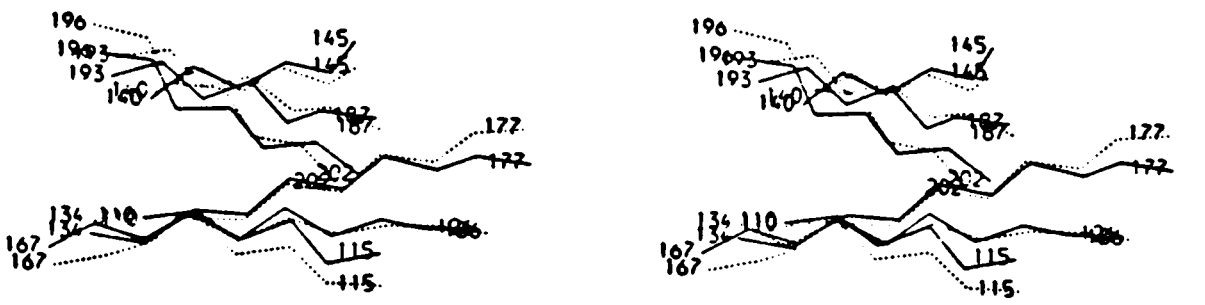
The r.m.s. deviation between these structures is 3.2Å for the 48 residues constituting the six core β-strands. The major source of error is the phasing of the two sheets relative to one another. Thus, the top sheet of the predicted structure appears to the left of the crystal structure while the bottom appears to the right. This structure remains as only one of 6 alternatives.

FIGURE 4.15, cont.



Human Immunoglobulin IgG(λ) new fragment of the variable domain of the light chain (FALV)

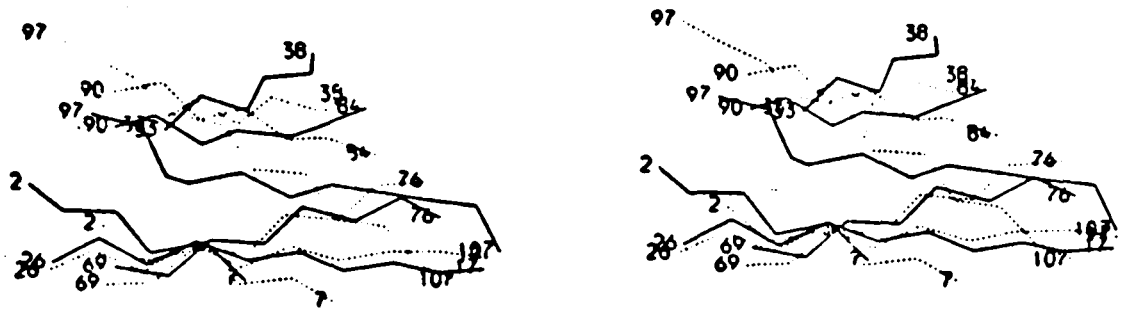
The r.m.s. deviation between these two structures is 4.9Å for the 48 residues constituting the six core β-strands. The major source of error in the predicted structure is the distortion in the crystal structure of strand H (residues 90-102) created by a β-bulge in the middle of this long strand. The algorithm is least selective in this example with 2907 alternatives comparable to the native-like alternative.



Human Immunoglobulin IgG(λ) new fragment of the constant domain of the light chain (FALC)

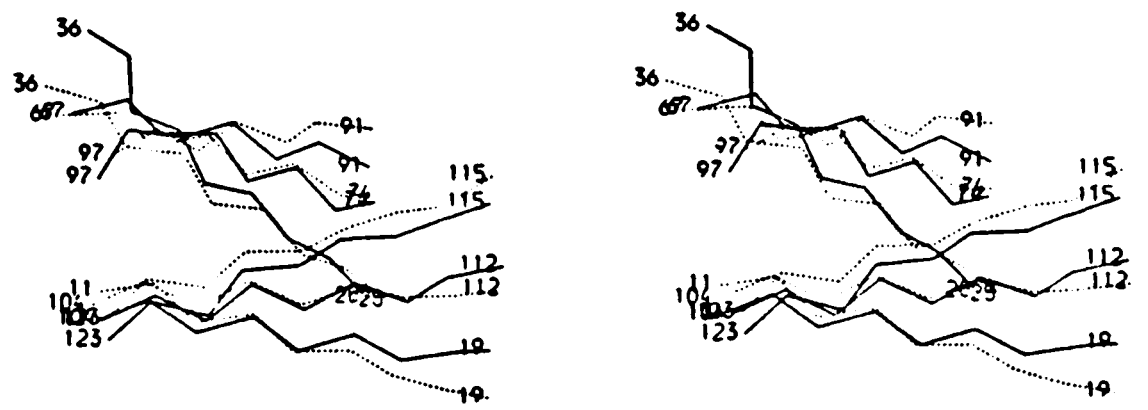
The r.m.s. deviation between these two structures is 1.4Å for the 46 residues constituting the six core β-strands. The agreement between these two structures only breaks down at the endpoints of the strands. This structure survives as one of 10 alternatives.

FIGURE 4.15, cont.



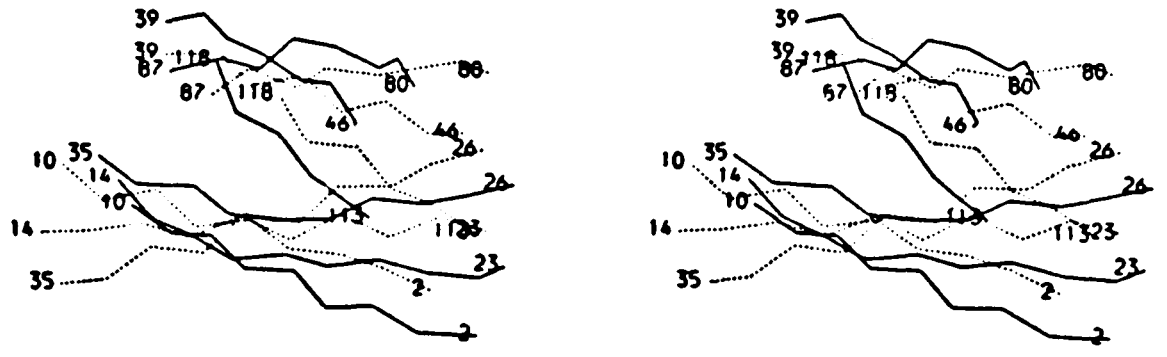
Bence-Jones Protein Variable Fragment (REI)

The r.m.s. deviation between these two structures is 3.6Å for 48 residues constituting the six core β-strands. The edge strand I (residues 97-107) is distorted by a β-bulge. This is not accurately modelled by an ideal β-sheet. This structure remains as one of 937 alternatives.



Human Prealbumin (PRE)

The r.m.s. deviation between these two structures is 2.0Å for 51 residues in the six core β-strands. Errors are limited to a slight fraying of residues at the endpoints of the strands. This structure remains as one of 10 alternatives.



Cu,Zn-Super Oxide Dismutase (SDM)

The r.m.s. deviation between these two structures is 5.1Å for the 51 residues in the six core β-strands. The β-barrel character of the crystal structure is not accurately modelled by two distinct idealised β-sheets. This structure remains as one of 795 alternatives.

One striking result of these calculations is that the role of the coiled regions linking β-strand have little if any sequence specific role in determining the structure of β-sandwiches. This is consistent with the hypervariability of coiled segments in the various immunoglobulin domains of known sequence (Beale & Feinstein, 1976). Of course, these residues have functional significance.

6 Application to the prediction of unknown structures: the histocompatibility antigens

The notions quantified in the algorithm provide, for the first time, the basis for making a secondary structure prediction and obtaining a tertiary fold for β-sheet sandwiches. Clearly further work on each stage of the overall scheme is required. Existing secondary structure prediction algorithms need to be improved to obtain a more accurate location of the β-strands. Given the predicted core, the few unplaced β-strands and the connections have to be located. Then one has to explore the use of simplified energy calculations (Levitt, 1976; Kuntz et al., 1976; Robson & Osguthorpe, 1979) first to select the correct fold from the reduced list and subsequently to refine the structure.

A demonstration of the current status of this approach was made by assessing the validity of predicted sheet diagrams for the light chain (β_2-m) and a segment of heavy chain (ac-2) of the HLA-B7 antigen (Dayhoff, 1976; Orr et al., 1979). Based on sequence alignments, it has been proposed that both β_2-m and ac-2 will have a 4- on 3-stranded sheet sandwich structure similar to those in the immunoglobulin (IgG) constant domains.

First, the secondary structure was predicted by the automatic procedure of Robson and co-workers (Garnier et al., 1978). The decision constants chosen were applicable to proteins with more than 50% β-structure and less than 20% α-helix in agreement with circular dichroism measurements (Orr et al.

1979). With these parameters nearly all the β-strand regions were localised in trials on IgG domains. The predicted β-structure in β_2-m (see Table 4.6) corresponds to the known β-strands in IgG constant domains when the sequences are aligned as in Dayhoff (1976). From this alignment, with only slight modifications to the secondary structure, the hydrogen-bonding diagram in FIgure 4.16 was constructed. The predicted central residues of the hydrophobic patches have the correct anticomplementary direction. Furthermore, probable β-bulges are located in strands A and B in homologous positions to observed bulges in FAHC and FCH3. When this approach was followed with the ac-2 sequence, a plausible hydrogen-bonding diagram could not be constructed. Trials with a sequence alignment program similar to that used for the ac-2/IgG comparisons failed to obtain the alignment of residues between FAHC and FAHV that is indicated by their crystal structures (Saul et al., 1978). Accordingly, sections of the proposed sequence alignment of ac-2/IgG were modified to obtain a hydrogen-bonding diagram that had suitable features of hydrophobic patches and β-bulges (Figure 4.16). The diagrams were taken as the 'native' structure of β_2-m and ac-2, and the algorithm for predicting sheet sandwiches was then used to establish that these structures are one of a small list (<100) of allowed sandwiches compatible with the β-strand assignment (Table 4.5). From the hydrogen-bonding diagrams we have generated atomic co-ordinates of the strands in ac-2 and β_2-m which are shown in Figure 4.17.

7 Conclusion

Clearly, there are severe restrictions on the number of allowed topologies for all-β proteins. Therefore, the topological similarity between Cu;Zn superoxide dismutase and the immunoglobulins (Richardson et al., 1976) could adequately be explained as the selection of similar structures from the small list of allowed folds. Thus, to favour the alternative explanation

TABLE 4.6

Secondary Structure Prediction for β_2-Microglobulin and the Histocompatibility factor ac-2.

Protein	α-helix	β-strands
β_2-microglobulin		1-12
		22-30
		34-39
	41-47	48-50
		61-65
		69-71
		80-89
ac-2 fragment		5-9
		16-24
		30-35
		46-51
		58-66
		76-81
		88-89

FIGURE 4.16. Predicted Strand Alignment in AC-2 and β_2-Microglobulin

AC2 PREDICTED

A → -2 — ASP/PRO pro [LYS] thr HIS val

B ← O — tyr PHE/GLY leu ALA trp [CYS] arg [LEU] thr ALA gl

E → +2 — THR phe GLU lys TRP ala [ALA] val VAL va

D ← — gly ALA/PRO arg THR glu VAL leu GLU ti

G → O — LEU pro [LYS] pro LEU thr

F ← O — HIS gln VAL his [CYS] thr TYR

C → -2 — ILE thr [LEU] thr TRP gln

BETA 2 PREDICTED

A → -2 — ile GLN arg THR/PRO lys [ILE] gln VAL tyr SER arg

B ← O — his PHE/GLY ser VAL tyr [CYS] asn LEU phe

E → O — ser PHE tyr [LEU] leu [TYR] ser TYR thr

D ← — glu VAL lys

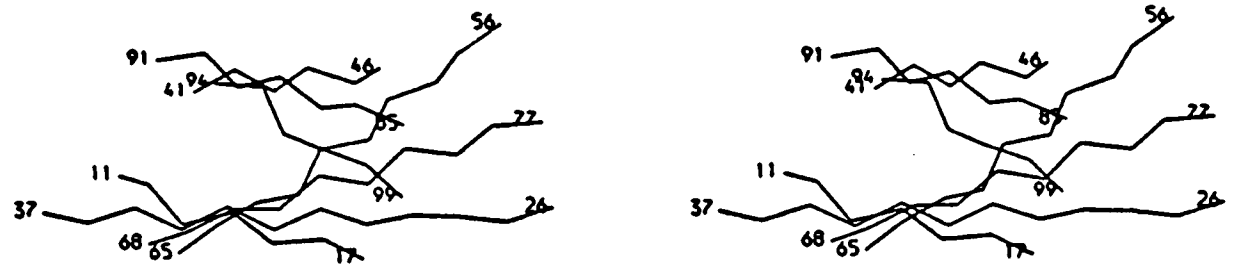
G → +2 — LEU ser GLN pro LYS ile [VAL] lys TRP

F ← O — HIS asn VAL arg [CYS] ala TYR

C → -2 — asp ILE glu [VAL] asp LEU

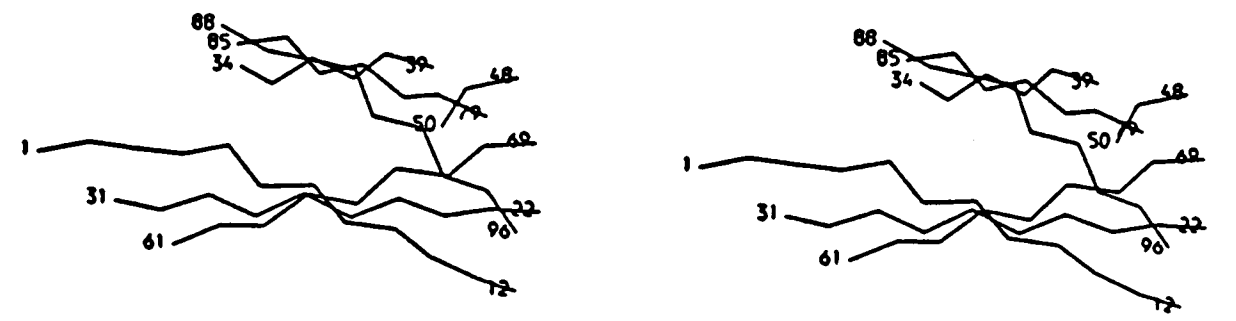
Putative central residues are boxed and ancillary hydrophobic residues contributing to the patch are underlined. Residues on the postulated internal faces of the sandwiches are shown in capital letters.

FIGURE 4.17. Predicted Structures for β_2-microglobulin and the Histocompatibility Factor AC-2.



Histocompatibility Factor AC-2. The sequence (Orr et al., 1979) was aligned to the immunoglobulin constant domains and together with a secondary structure prediction, a strand assignment was postulated. SHEETPERM produced a small set of alternatives from this strand assignment, one of which is shown above. This structure has an immunoglobulin-like topology.

β_2-microglobulin.



A similar analysis of the sequence of β_2-microglobulin (Dayhoff, 1976) yielded a strand assignment which provided the input for SHEETPERM. A small set of alternative structures including one with an immunoglobulin-like fold shown above was generated.

of divergence from a common ancestor requires distinct all-β proteins to have a very similar topology and highly significant sequence resemblances.

I should stress that the algorithm described in this Chapter is the first protein folding scheme to be uniformly applied to a variety of proteins of both known and unknown structures. Also, the predicted structures show a marked resemblance to the experimentally determined structures as seen both visually and through the r.m.s. deviations.

CHAPTER 5

βαβ PROTEINS

The third structural motif commonly observed in globular proteins is a parallel β-sheet covered on top and bottom by α-helices. The polypeptide chain often alternates regularly between α-helical and β-structure producing the generic chain repeat $(\alpha\beta)_n$. In this structure, the β-sheet is largely composed of hydrophobic residues. These hydrophobic side chains are buried by a coating of α-helices which have both hydrophobic and hydrophilic faces. Flavodoxin, Rhodenese, and many of the glycolytic enzyms fall into this general structural class. Thioredoxin, glyceraldehyde 3-phosphate dehydrogenase and carboxypeptidase A have mixed, not pure parallel, β-sheets but also exhibit an αβ repeat, albeit less regularly. These proteins will be a focus for developing concepts which unify the conclusions of the last two chapters.

1 Prelude

Pure parallel β-sheets are commonly found in proteins with an alternating pattern of secondary structure $(\beta\alpha)_n$. Nine such proteins were analysed to see what patterns of hydrophobic residues confer stability upon these globular polypeptide chains. The results of this analysis were developed into an algorithm for predicting the hydrogen bonding diagram of the β-sheet from a consideration of sequence and the location of secondary structure. A small list of structures was consistent with four topological and three strand alignment rules about $(\beta\alpha)_n$ proteins.

Proteins with mixed β-sheets were analysed and the rules developed for pure parallel β-sheets were extended to explain this class of structures.

An algorithm was developed to predict the hydrogen bonding diagram for mixed sheets with sequence and secondary structures as input. The results of these studies and their role in the prediction of protein structure are discussed.

2 Pure Parallel β-sheets

2.1 Definitions

Strand assignments and hydrogen bonding diagrams were produced from the original crystallographic papers describing the protein structure or from the Atlas of Protein Structure (Feldman, 1976). In every protein, the authors' definition of β-strands was revised in the light of a list of possible hydrogen bonds. Residues which had appropriate ϕ and ψ angles for β-structure but were not hydrogen bonded through a main chain atom to a sequentially distant main chain atom or which did not lie between two other residues bonded in this fashion were not considered part of a β-sheet. For the purpose of accessible area calculations, additional sheet edge residues were included to guarantee that all changes following helix packing were monitored. These assignments are compiled in Table 5.1.

2.2 Topological Properties of pure parallel β-sheets

A pure parallel β-sheet could theoretically adopt $\frac{1}{2}.2^{n-1}.n!$ strand topologies, where n is the number of strands. n strands could have n! strand orders but rotational symmetry reduces this by a factor of two. Each of the n-1 connecting loops could be left or right handed and so there are 2^{n-1} possibilities. Fortunately, many of the possible topologies are never seen. Several investigators (Sternberg & Thorton, 1976; Richardson, 1977) have explained the apparent preference for a few recurrent strand patterns through two topological constraints:

(1) The connection between two parallel β-strands is right-handed.

TABLE 5.1

Secondary Structure Assignments for βα proteins with pure parallel β-sheets.

Protein	β-strands		α-helices	
	N-terminus	C-terminus	N-terminus	C-terminus
Adenyl kinase	10	14	23	30
	35	38	53	62
	90	94	100	107
	114	118	144	164
	169	173	179	194
Alcohol dehydrogenase	193	199	170	187
	218	224	202	212
	238	243	229	236
	263	269	250	259
	287	293	275	283
	312	318		
Phospho-glycerate kinase N-terminal domain	17	22	41	52
	56	61	77	89
	91	96	101	109
	114	119	144	155
	158	163	173	178
	182	187		
Phospho-glycerate kinase C-terminal domain	207	212	189	202
	231	236	218	229
	277	282	261	275
	332	336	317	330
	367	371	348	365
	388	392	375	380
Lactate dehydrogenase	21	26	29	43
	45	51	55	70
	75	79		
	88	93	120	130
	130	134	139	151
	155	158	165	181
Rhodenese N-terminal domain	7	10	11	22
	29	33	42	50
	55	58	76	87
	94	98	107	119
	122	126	129	137
Rhodenese C-terminal domain	160	162	163	174
	176	181	183	189
	208	210	224	235
	242	246	251	264
	268	271	274	282

cont.

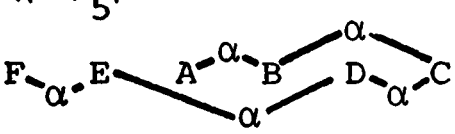
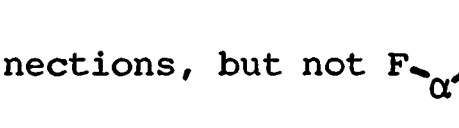
TABLE 5.1, cont.

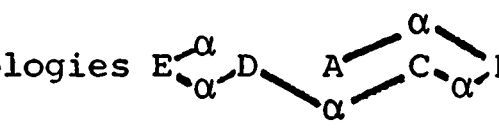
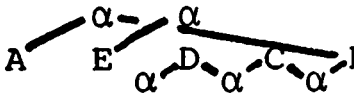
Protein	β-strands		α-helices	
	N-terminus	C-terminus	N-terminus	C-terminus
Triose Phosphate Isomerase	6	12	17	31
	38	42	44	55
	60	63	79	87
	89	93	105	120
	122	129	138	154
	159	167	177	196
	205	209	213	223
	227	231	237	246
Subtilisin	27	32	14	20
	45	49	64	73
	89	94	103	117
	120	125	132	145
	148	152	223	238
	174	177		
Flavodoxin	1	6	10	27
	29	35	66	74
	48	55	93	106
	80	87	124	138
	108	119		

Right handed βαβ units are found in 57 of 58 examples (Sternberg & Thorton, 1976). The right-handed rule reduces the number of topologies by a factor of 2^{n-1}.

(2) Pure parallel sheets are unknotted and have no more than one chain reversal (Richardson, 1977). As two chain reversals are required for a knot, the first half of this restriction is embodied in the second. Thus, ABEDC is an acceptable topology, but ABDEC is not.

From a survey of nine proteins with pure parallel β-sheets, FLAV, LDH, SUBT, RHOD, ADH, PGK, TIM and PK* (see Figure 5.1), two additional rules further restricting the set of allowed topologies were evident:

(3) The difference in the number of α-helices on the two faces of the β-sheet is never more than 1. Thus PGK with the form $(\beta\alpha)_5\beta$ in the N-terminal domain could have the topology  which puts two helices on the top and three on the bottom of the β-sheet assuming right-handed connections, but not  which forces four of the five helices to be on the bottom of the sheet.

(4) An α-helix trailing the final β-strand lies on top of the β-sheet and always to the left of the first strand when the first strand is placed to the left of the second strand. In AK, with the form $(\beta\alpha)_5$, the topologies  and  have two helices on the top and three on the bottom. However, the second would not be allowed because of the steric requirements of the α-helices as they pack against the β-sheet.

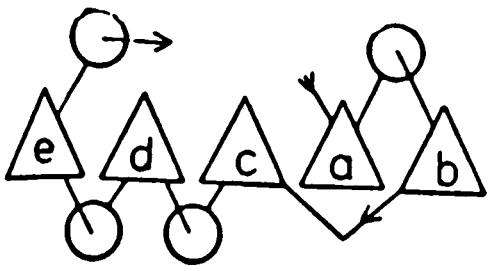
Of course, topologies generated which are compatible with these rules may not be physically reasonable since the polypeptide segment joining the

* The abbreviations used in this section are: Flavodoxin (FLAV), Lactate Dehydrogenase (LDH), Subtilisin (SUBT), Rhodenese (RHOD), Alcohol Dehydrogenase (ADH), Phosphoglycerate kinase (PGK), Triose Phosphate Isomerase (TIM), Pyruvate Kinase (PK).

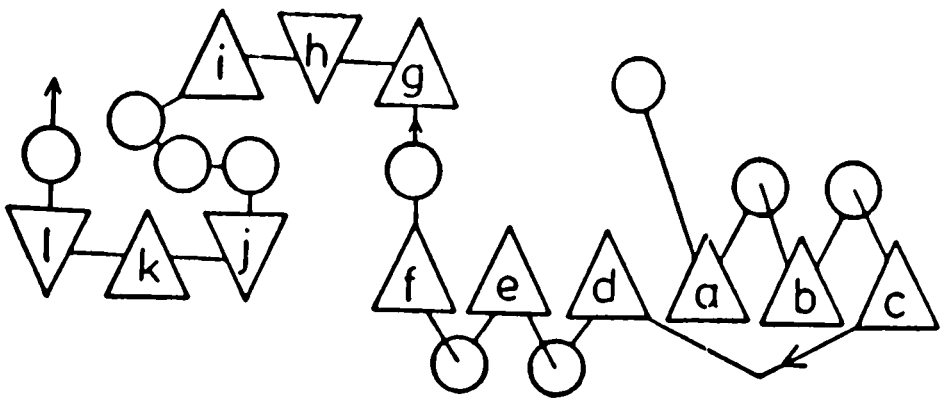
FIGURE 5.1

Schematic Diagram of Proteins with Pure Parallel β-Sheets.

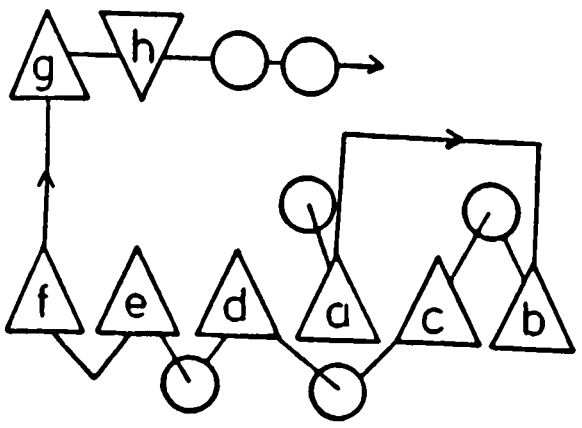
A schematic diagram of the proteins with pure parallel β-sheets considered in this work. Each strand is represented by a triangle whose apex points up or down according to whether the strand is viewed from the N- or C- terminus. A circle represents an α-helix. TIM and PK are drawn to accentuate their β-barrel structures.



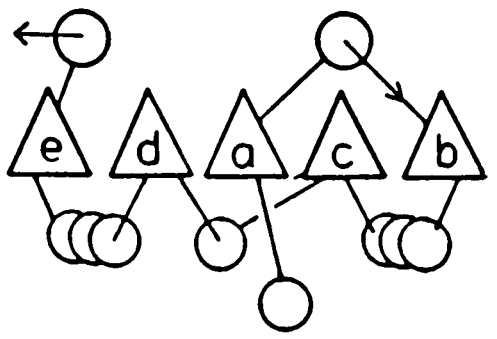
FLAV



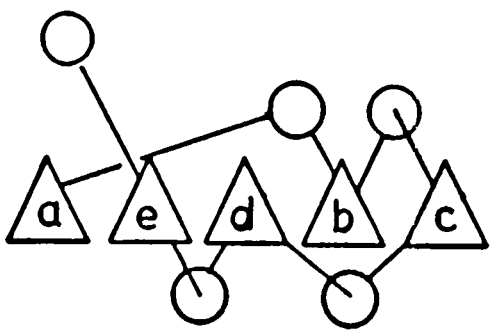
LDH



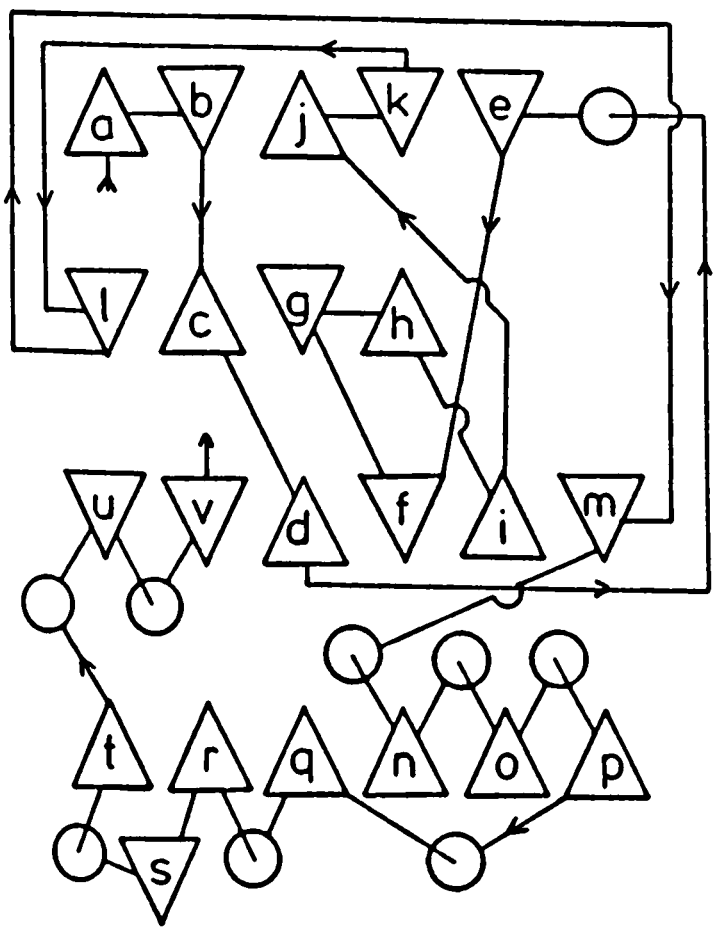
SUBT



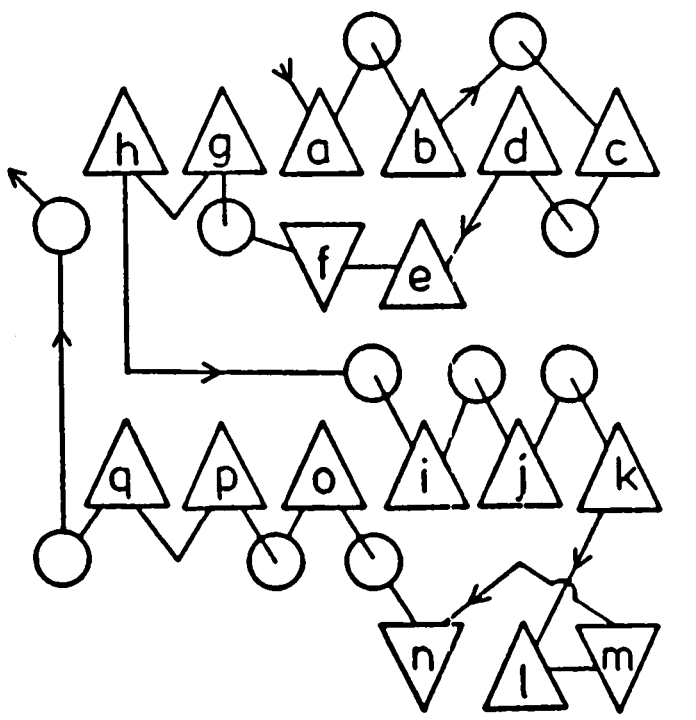
AK



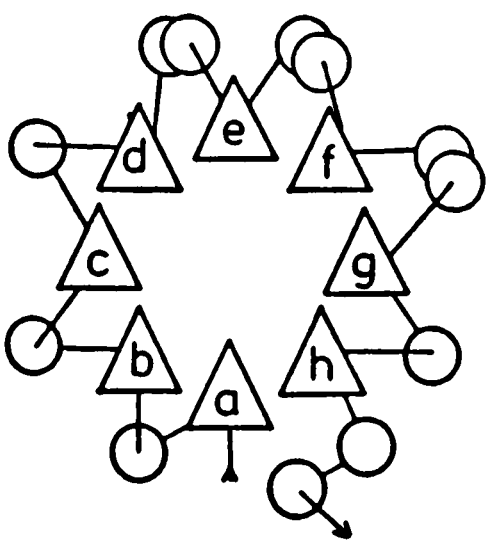
RHOD



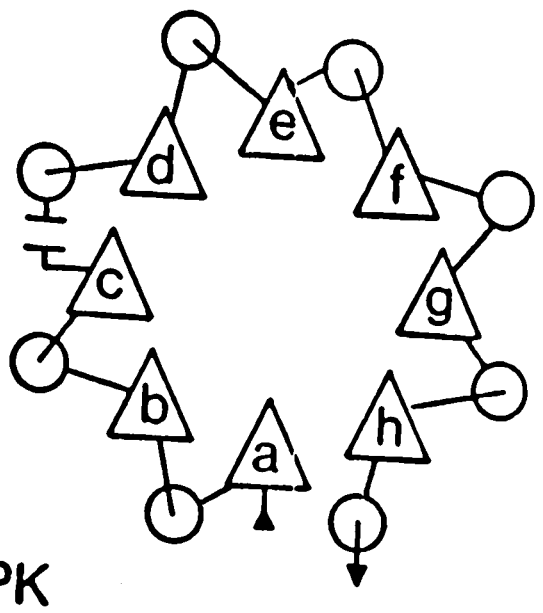
ADH



PGK



TIM



PK

secondary structure units may not be consistent with the link length required. An ideal parallel β-sheet was constructed and if the distances between the terminii of consecutive strands were greater than the length of the linking chain (1.5Å for helices and 3.3Å for coils) needed to join these points, the topology was rejected.

Table 5.2 demonstrates that although the number of possible topologies is vast, the number of allowed topologies is a manageable subset of these. This data is a result of the computer program BAB (see Appendix V) which has the assignments of secondary structure as input and outputs a list of topologies allowed by rules 1-4.

2.3 The Alignment of Residues in a pure parallel β-sheet - Analysis

A study of the disposition and juxtaposition of interacting residues on the internal face of the stacked pair of β-sheets indicated that hydrophobic residues clustered to form a pair of anticomplementary parallelograms on the buried sheet surfaces. This tendency suggested rules for phasing the residues of six of the strands within the two sheets. This effect was attributed to the characteristic twist of the β-sheet (Chothia, 1974) and so a similar pattern was expected for pure parallel β-sheets covered by α-helices.

In an effort to elucidate the packing pattern of residues on a pure parallel β-sheet, a detailed study of the non-polar accessible contact area (NPACA) at various levels of organisation of the protein was performed. This section focuses on the differences in NPACA for the sheet together with an α-helix packed upon it and the isolated sheet and helix. The per residue NPACA differences for the residues on the sheet following the removal of various helices are detailed in Figure 5.2. The fourteen diagrams shown correspond to the dissections which resulted in the largest losses in NPACA.

Two conclusions are evident from an inspection of these diagrams:

TABLE 5.2

The Effect of Topological Restrictions on Pure Parallel β-sheets

Protein		Number of Strands	Number of Possible Topologies[a]	Number of Topologies consistent with Rule 1	Number of Topologies consistent with Rules 2-4	Number of Topologies allowed by simple endpoint constraint
Flavodoxin		5	960	60	6	5
Adenyl Kinase		5	960	60	6	6
Phospho-glycerate kinase	N-domain	6	11520	360	10	10
	C-domain	6	11520	360	10	3
Alcohol dehydrogenase		6	11520	360	15	14
Triose Phosphate Isomerase		8	258480 + 1[b]	20160 + 1	20 + 1	18 + 1
Rhodenese	N-domain	5	960	60	6	6
	C-domain	5	960	60	6	6
Lactate dehydrogenase		6	11520	360	13	13
Subtilisin		6	11520	360	6 + 6[c]	6 + 6

[a] Computed as $\frac{1}{2}.2^{n-1}.n!$ where n is the number of strands.

[b] β-Barrel allowed if more than 6 strands in the sheet.

[c] Six additional topologies allowed if left handed βαβ is allowed between strands B and C.

FIGURE 5.2. Non-Polar Accessible Contact Area for the Packing of α-helices on Pure Parallel β-sheets.

ADENYLATE KINASE

(1) The residues on the top of the β-sheet which lose NPACA when Helix 1 (residues 23-30) is removed.

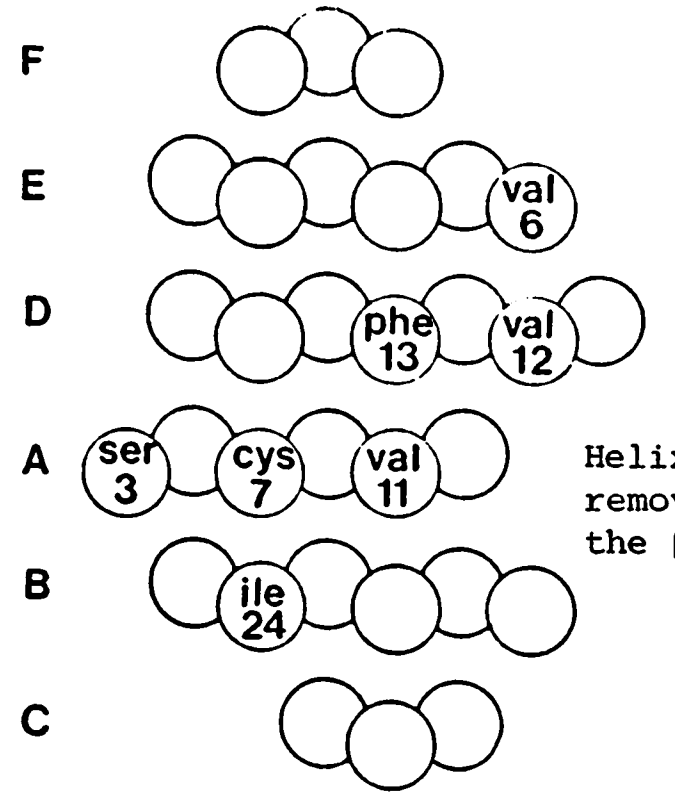
(2) The residues on the bottom of the β-sheet which lose NPACA when Helix 4 (residues 144-164 is removed.

(3) The residues on the top of the β-sheet which lose NPACA when the trailing helix (residues 179-194) is removed.

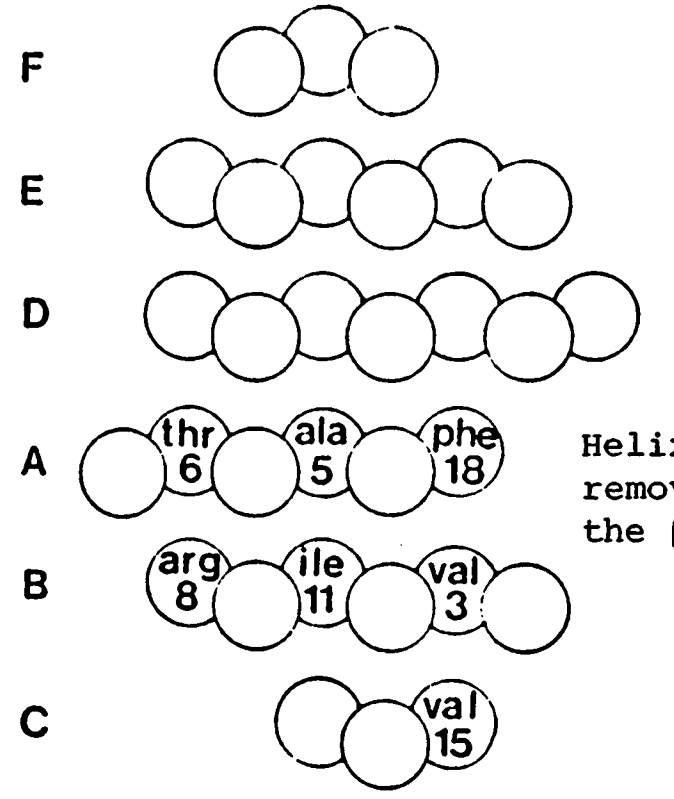
Note that the pattern of area changes on the top of the sheet is anticomplementary to the pattern on the bottom of the sheet.

FIGURE 5.2, cont.

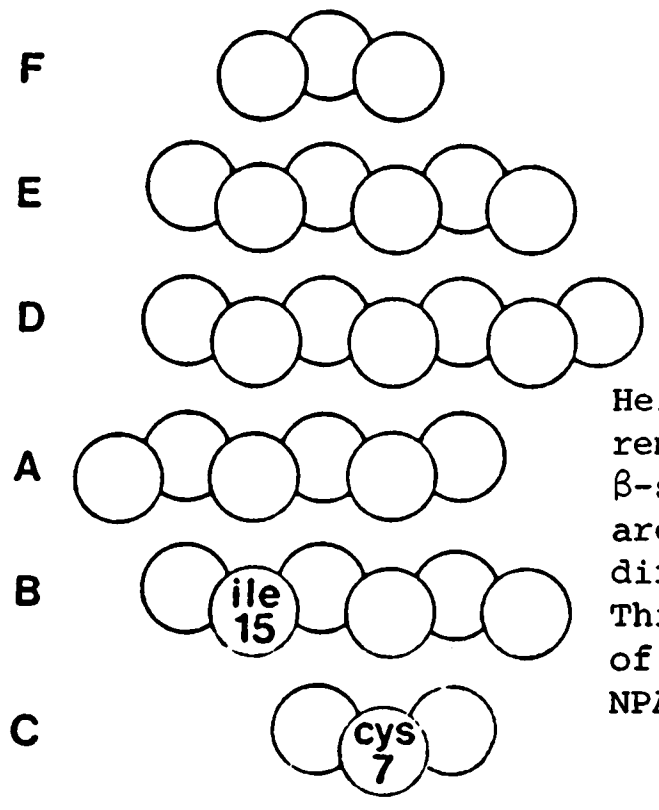
ALCOHOL
DEHYDROGENASE



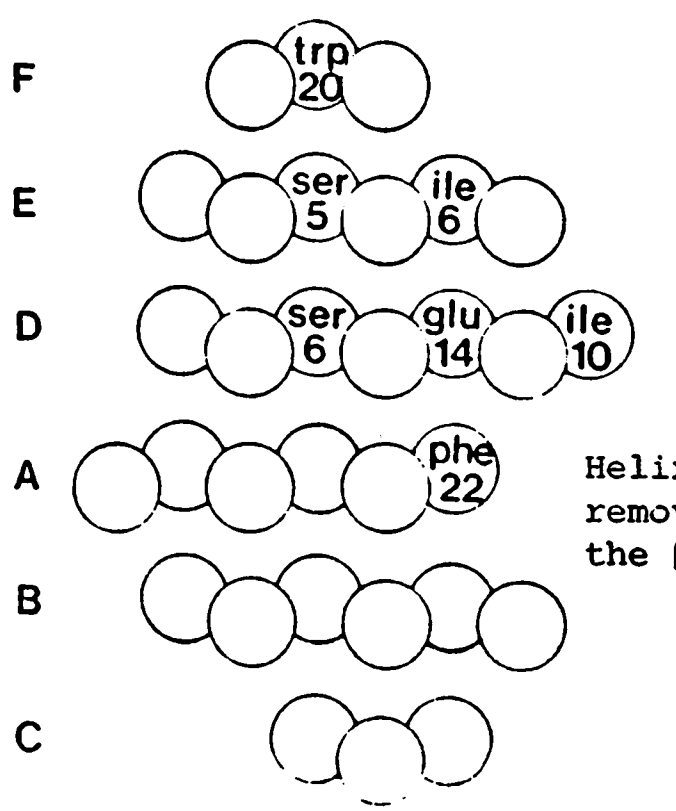
Helix 2 (residues 202-212) removed from the top of the β-sheet.



Helix 4 (residues 250-259) removed from the bottom of the β-sheet.



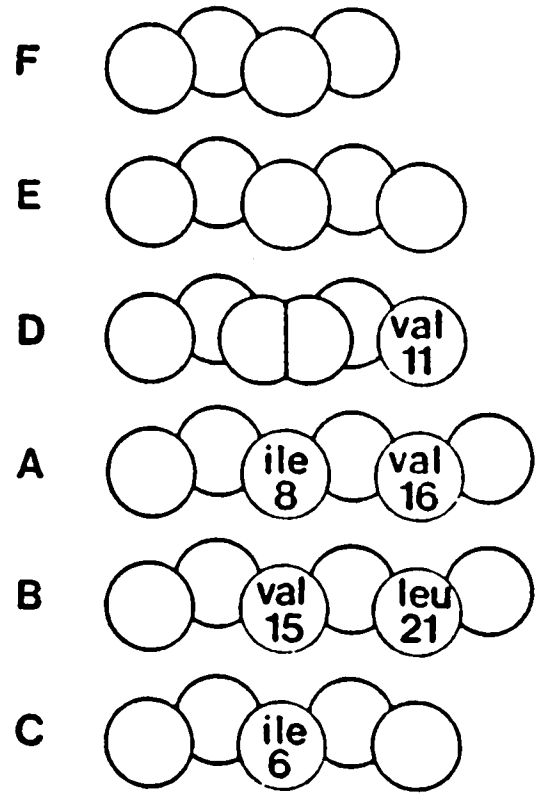
Helix 3 (residues 229-236) removed from the top of the β-sheet. The pattern of area changes is in the wrong direction in this example. This is probably a consequence of the limited nature of the NPACA change on helix packing.



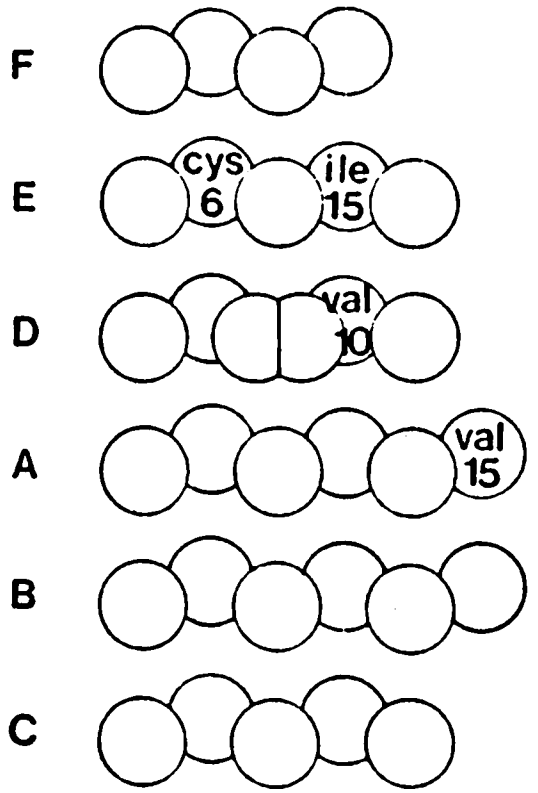
Helix 5 (residues 275-283) removed from the bottom of the β-sheet.

FIGURE 5.2, cont.

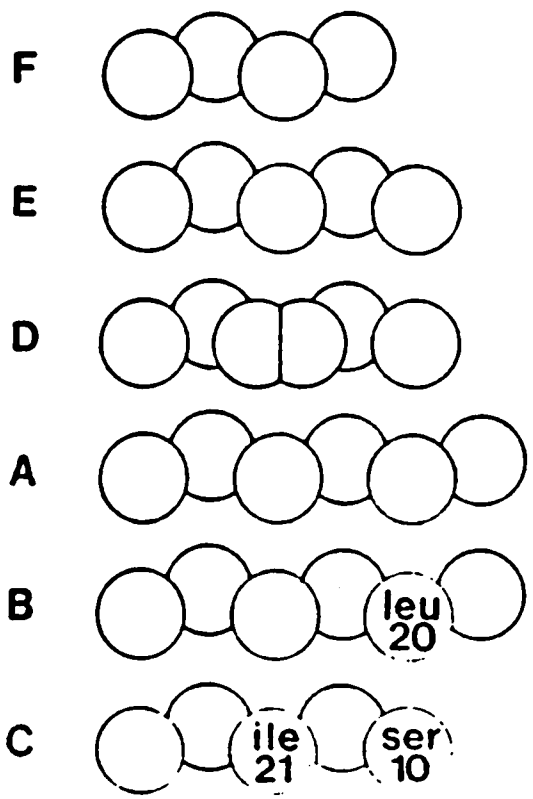
LACTATE
DEHYDROGENASE



Helix 1 (residues 29-43) removed from the top of the β-sheet.



Helix 3 (residues 120-130) removed from the bottom of the β-sheet.



Helix 2 (residues 55-70) removed from the top of the β-sheet.

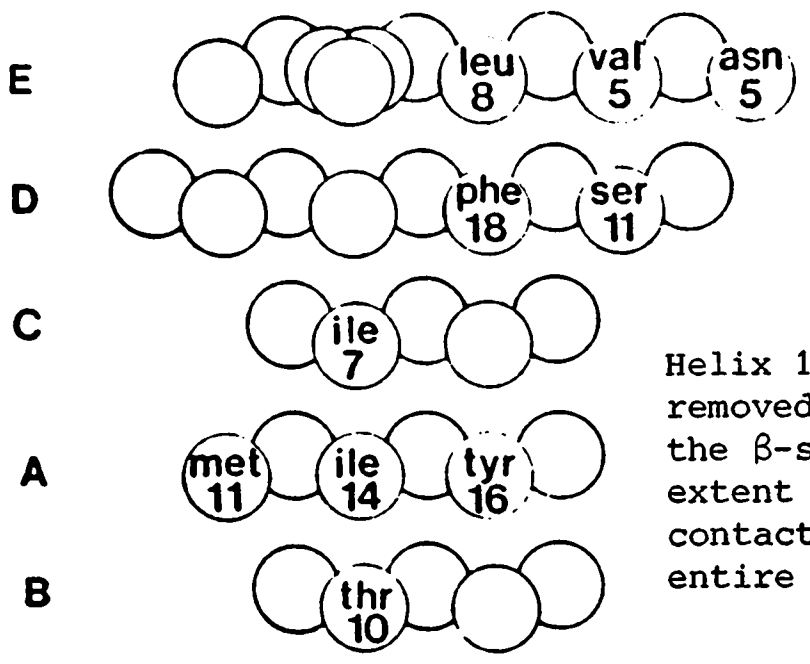
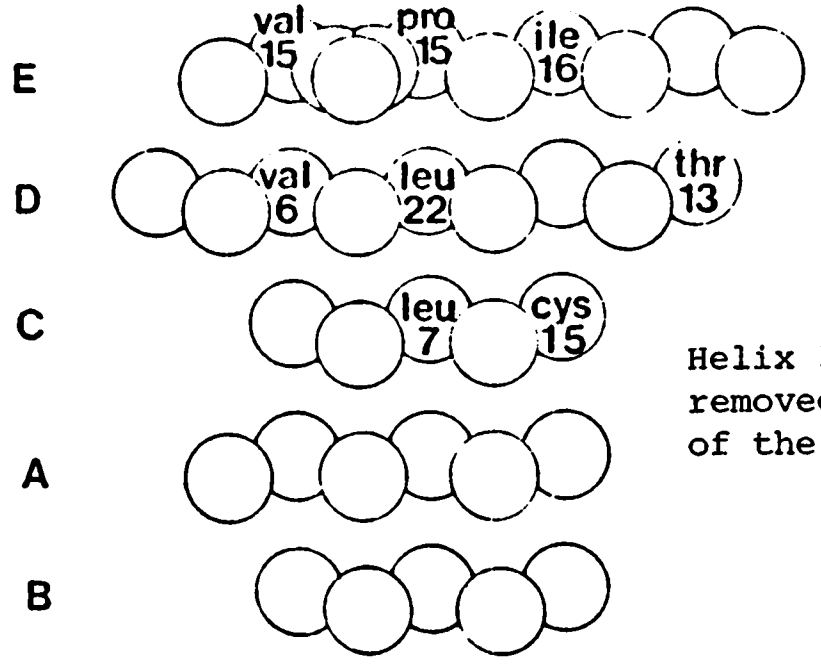


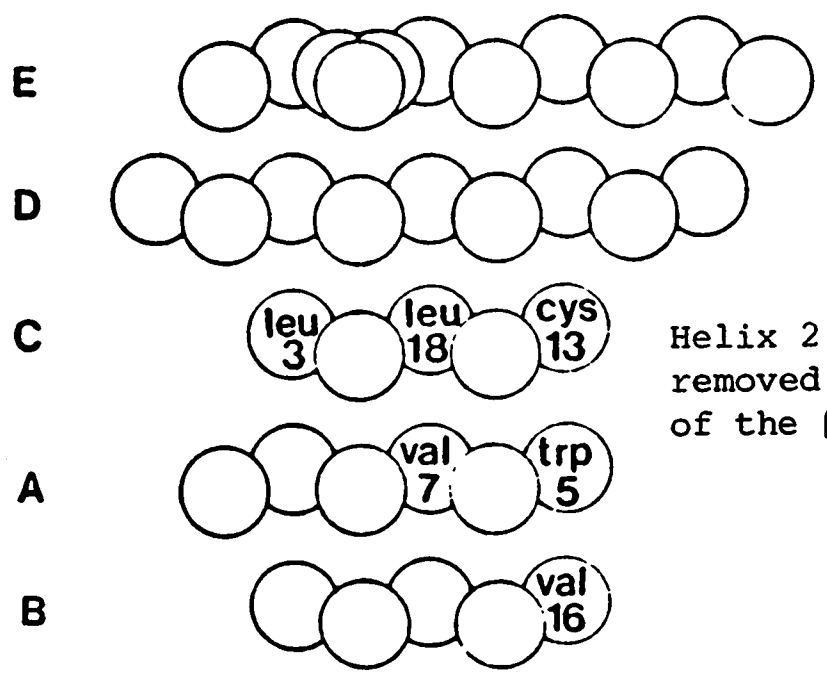
FIGURE 5.2 cont.

FLAVODOXIN

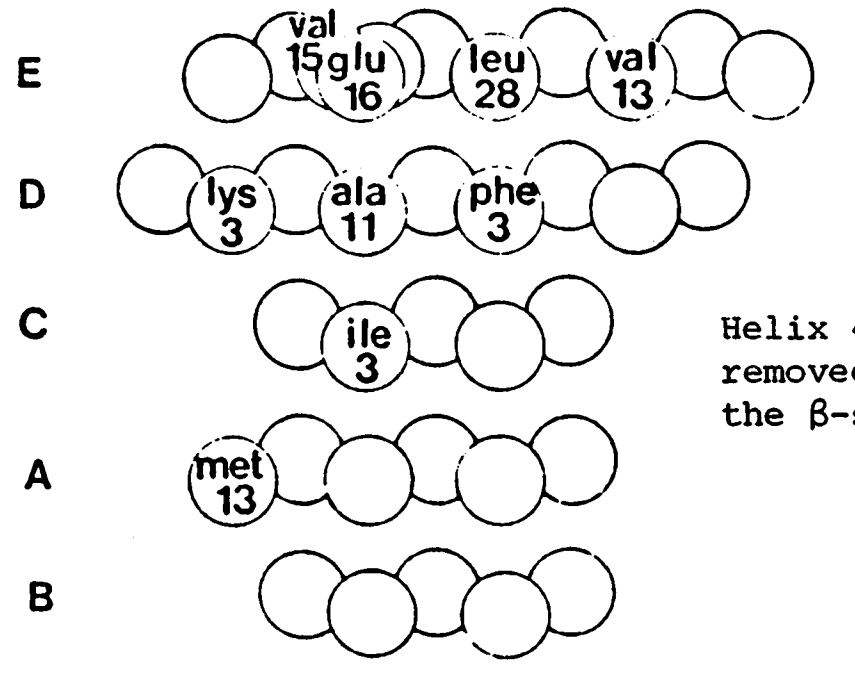
Helix 1 (residues 10-27) removed from the top of the β-sheet. Note the extent of the changes in contact area across the entire β-sheet.



Helix 3 (residues 93-106) removed from the bottom of the β-sheet.



Helix 2 (residues 66-74) removed from the bottom of the β-sheet.



Helix 4 (residues 124-138) removed from the top of the β-sheet.

(1) The residues with significant (>2Å²) NPACA changes upon removal of the α-helices are predominantly hydrophobic (83%).

(2) Although the NPACA changes are distributed over as many as five strands, the direction of the patch on the bottom surface of the sheet is anticomplementary to that on the top surface of the sheet. This pattern is clear in all but one case, the removal of helix 3 from the bottom surface of ADH. This aberration is probably the result of the limited changes in NPACA.

2.4 The Alignment of Residues in a pure parallel β-sheet - Prediction

Although a restricted set of topologies for pure parallel β-sheets provides a reasonable reduction in the problem of predicting the structure of a βαβ proteins, the arrangement of residues along the strand relative to neighbouring strands must also be identified. A survey of nine pure parallel β-sheets (Rhodenese (2), PGK (2), AK, ADH, Subtilisin, LDH, FLAV) guided by the conclusions of the analysis of NPACA loss, revealed three restrictions for allowed H-bonding patterns:

(5) There exists a central core of hydrophobic residues in two adjacent rows i, i+1, in all of the non-edge strands where i or i+1 is within one residue of the strand midpoint. Thus, in flavodoxin, the pair Ala-Leu in strand 4 aligns with Ile-Leu in strand 3 and Ile-Val in strand 1. The number of hydrophilics in the region is small and typically 0 or 1.

(6) Potential hydrogen bonding between strands is always within 1 of the maximum number possible. This is quantified by strand overlap.

(7) If a hydrophobic island is defined as an uninterrupted row of two or more hydrophobic residues which align, then the midpoints of these hydrophobic rows, h_i, must have

$$h_i > h_{i+2} > \quad .... \quad > h_k \qquad \text{and}$$

$$h_{i+1} < h_{i+3} < \ldots < h_k$$

for some row i. This implies that the groups of hydrophobics in alternating rows progress from the lower left to the upper right corner of the sheet while the intervening rows proceed from the upper left to the lower right hand corner. The residues involved in the lower left to upper right patch are on the top of the sheet when the strands run from right to left across the page and the first strand is above the second strand. Moreover, a stretch of hydrophobic islands must occupy at least 4 consecutive rows and no intervening rows are seen.

A computer program, SHEET (see Appendix V) was written to apply constraints 5-7. As input, this program uses the set of allowed topologies supplied by BAB together with the amino acid sequence and strand assignments. For each strand, four possible positions for the strand midpoints of each strand are sampled. This produces a list of 4^n hydrogen bonding diagrams where n is the number of strands for each topology. Each structure is tested against rules 5-7 and acceptable hydrogen bonding diagrams are output (see Figure 5.3). The computation time is short, ranging from less than 30 seconds for flavodoxin to about 50 minutes for triose phosphate isomerase on an ICL 2980. Certainly the time requirements are small compared to BUILD for α-helices or SHEETPERM for β-sheet sandwiches, but the set of possible structures is also highly restricted.

2.5 Results and Discussion

When these restrictions are applied to the set of all possible H-bonding patterns consistent with the allowed topologies, a small set of structures is permitted. More importantly, the exact H-bonding pattern is generated if discrepancies due to β-bulges are ignored. The level of the reduction is presented in Table 5.3.

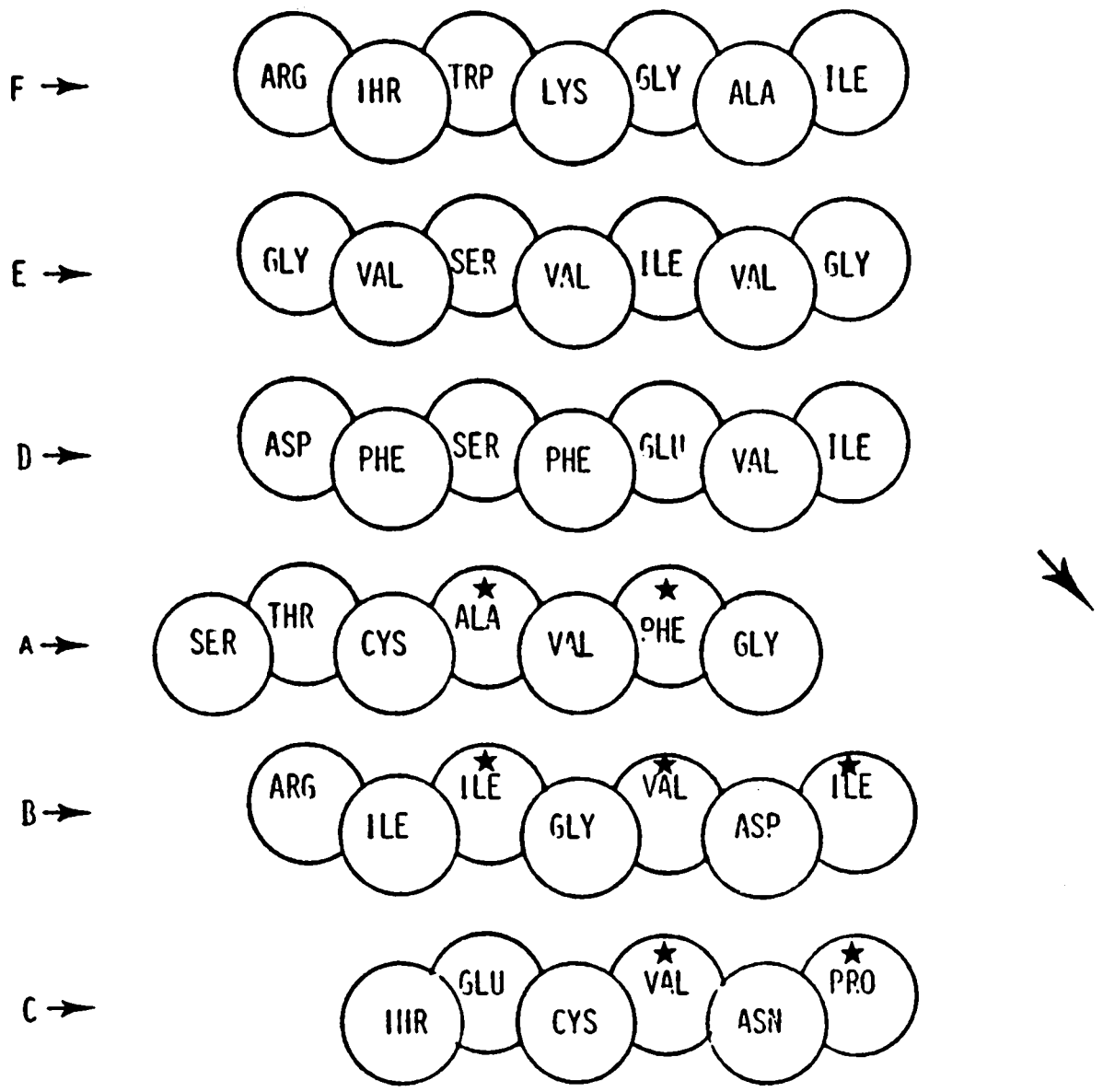
Two interesting anomalies are found when these rules are used to

FIGURE 5.3

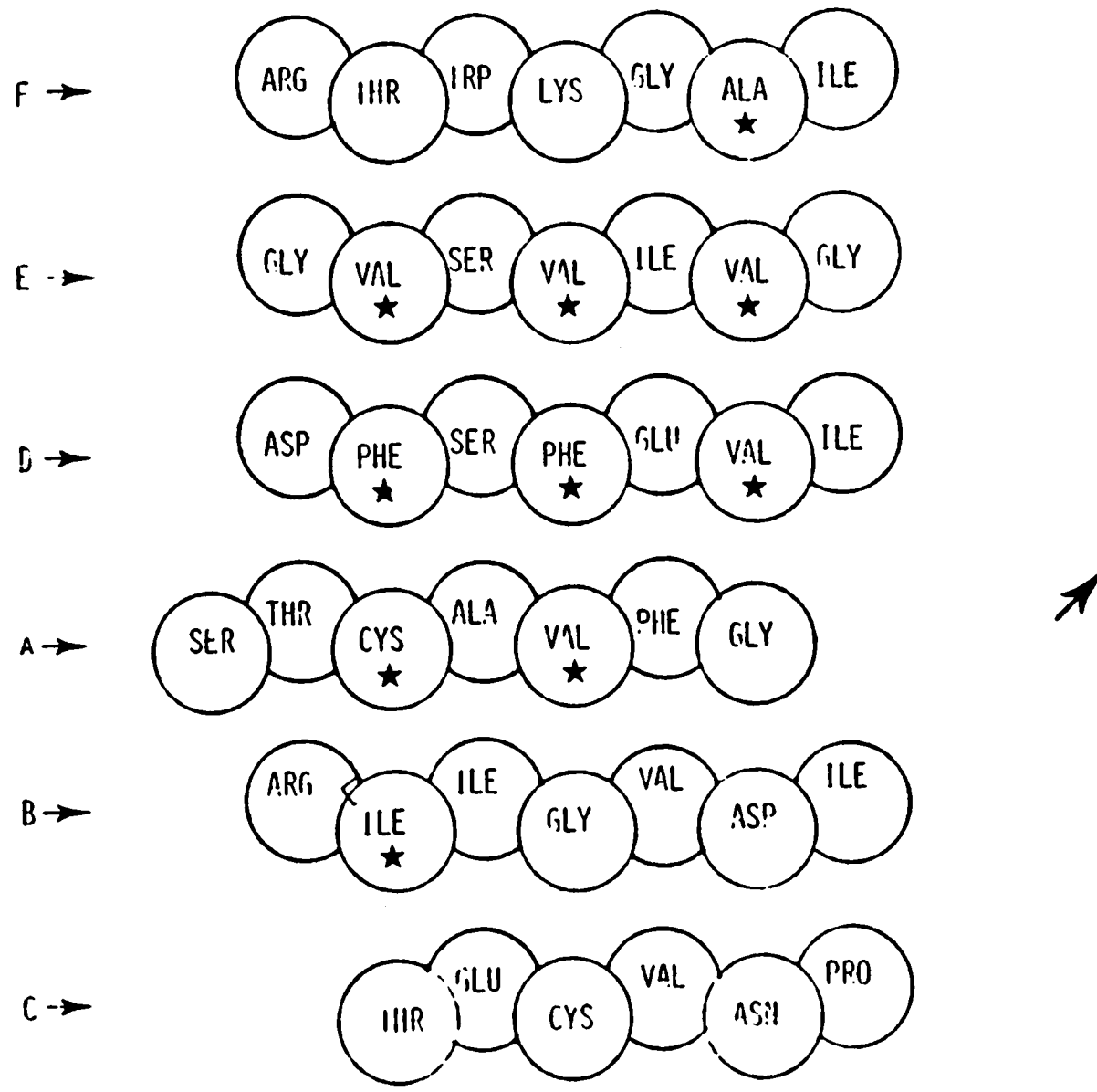
Bubble Diagrams of Pure Parallel β-sheets.

The strand alignment of nine pure parallel sheets in seven proteins is shown in the bubble diagrams on the facing and following pages. The residues which contribute to the hydrophobic patch predicted by rules 5 - 7 are shown with stars. Residues on the top face of the β-sheet which contribute to the patch are starred in the upper frame. Residues on the bottom face of the β-sheet which contribute to the patch are starred in the lower frame. Arrows indicate the direction of the constellation of hydrophobic residues and their generic anticomplementarity. These are also the alignments that are predicted by the algorithm SHEET.

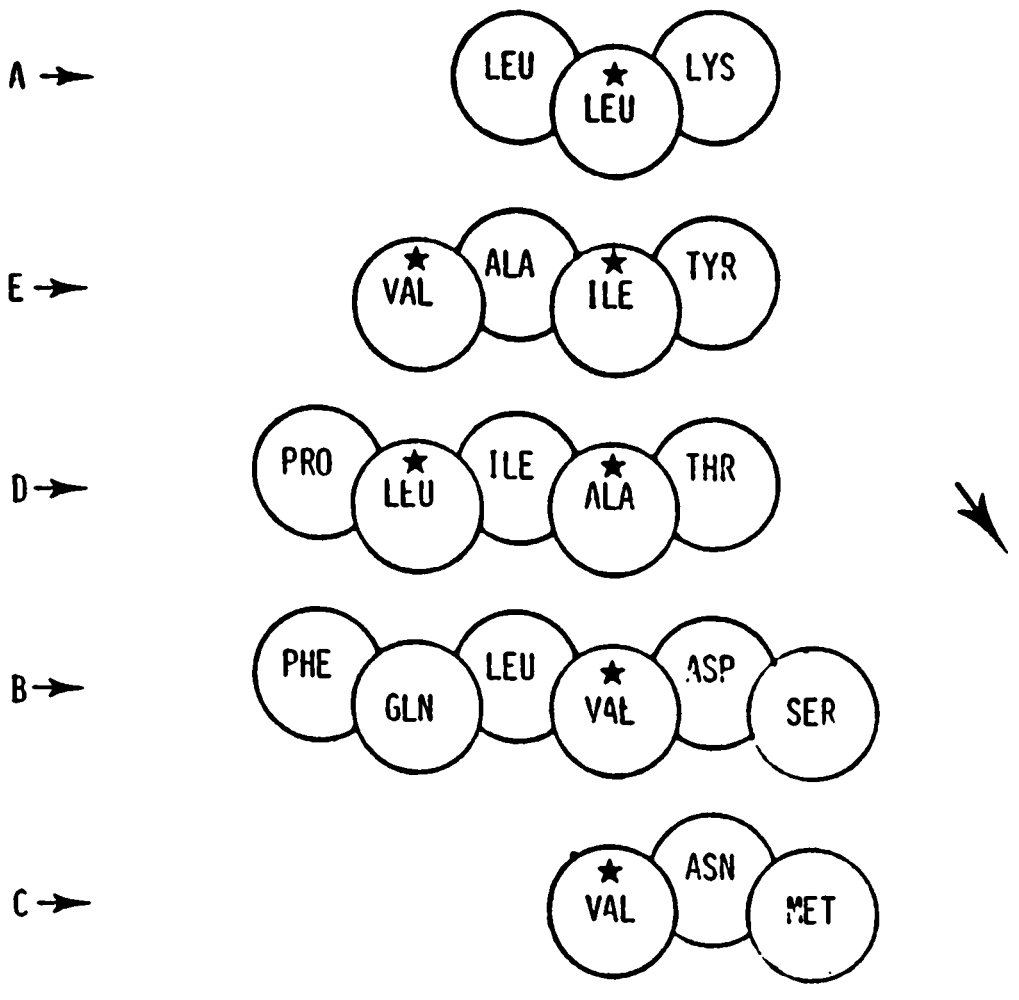
ALCOHOL DEHYDROGENASE



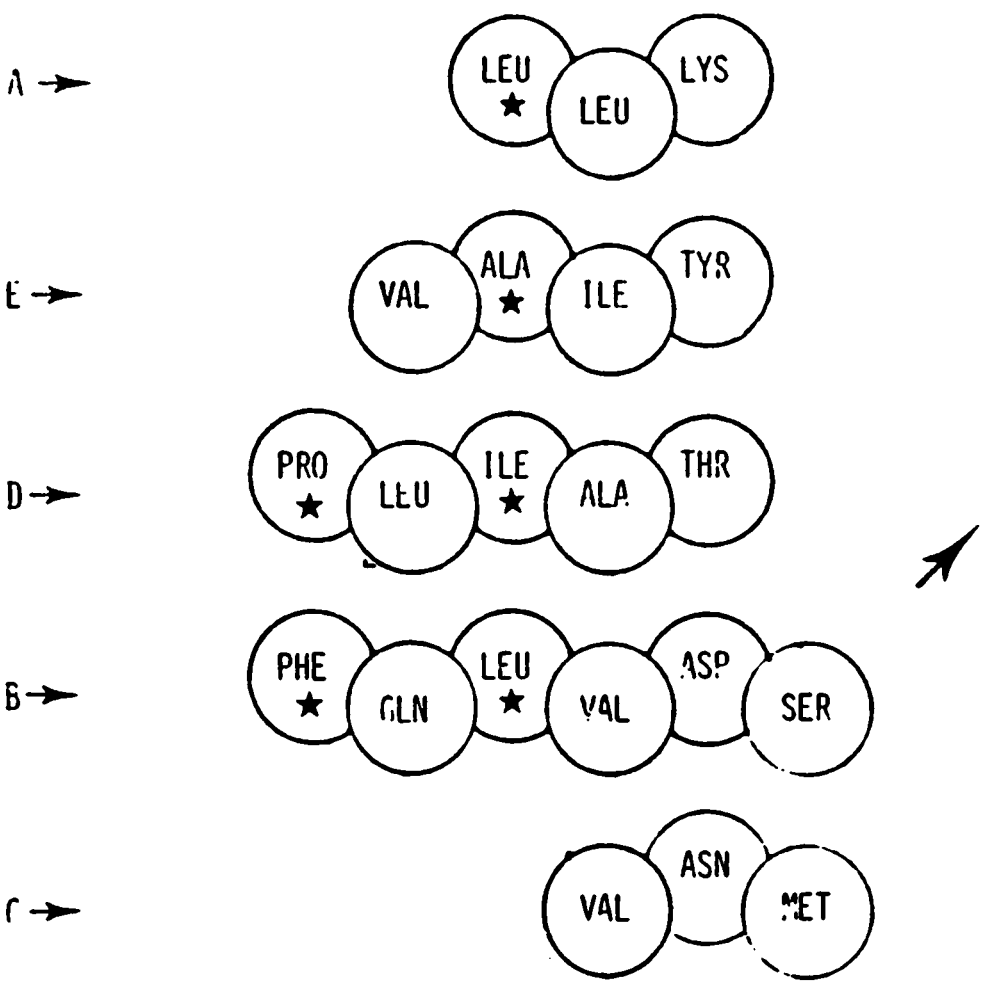
ALCOHOL DEHYDROGENASE



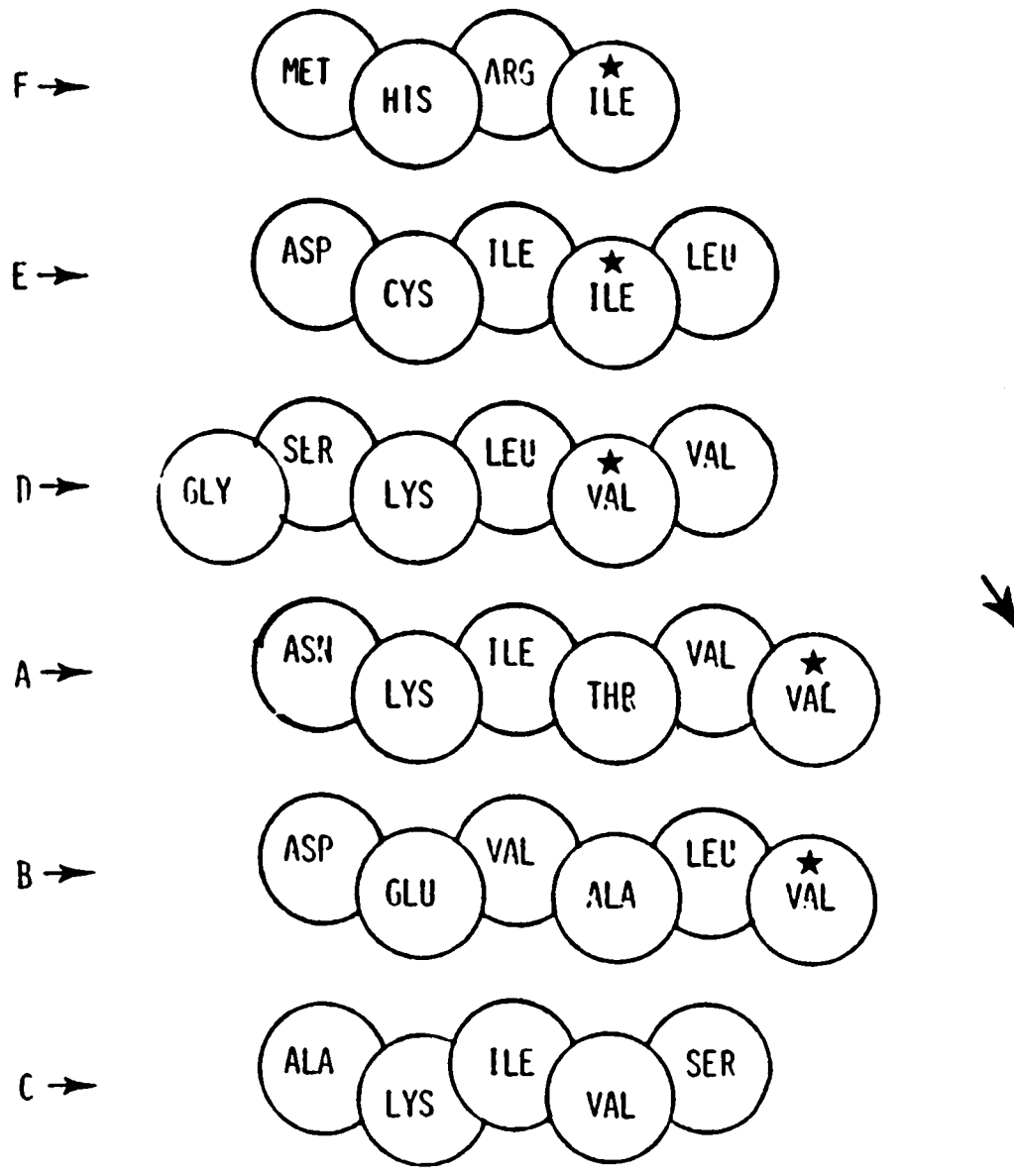
RHODENESE (C-TERMINAL DOMAIN)



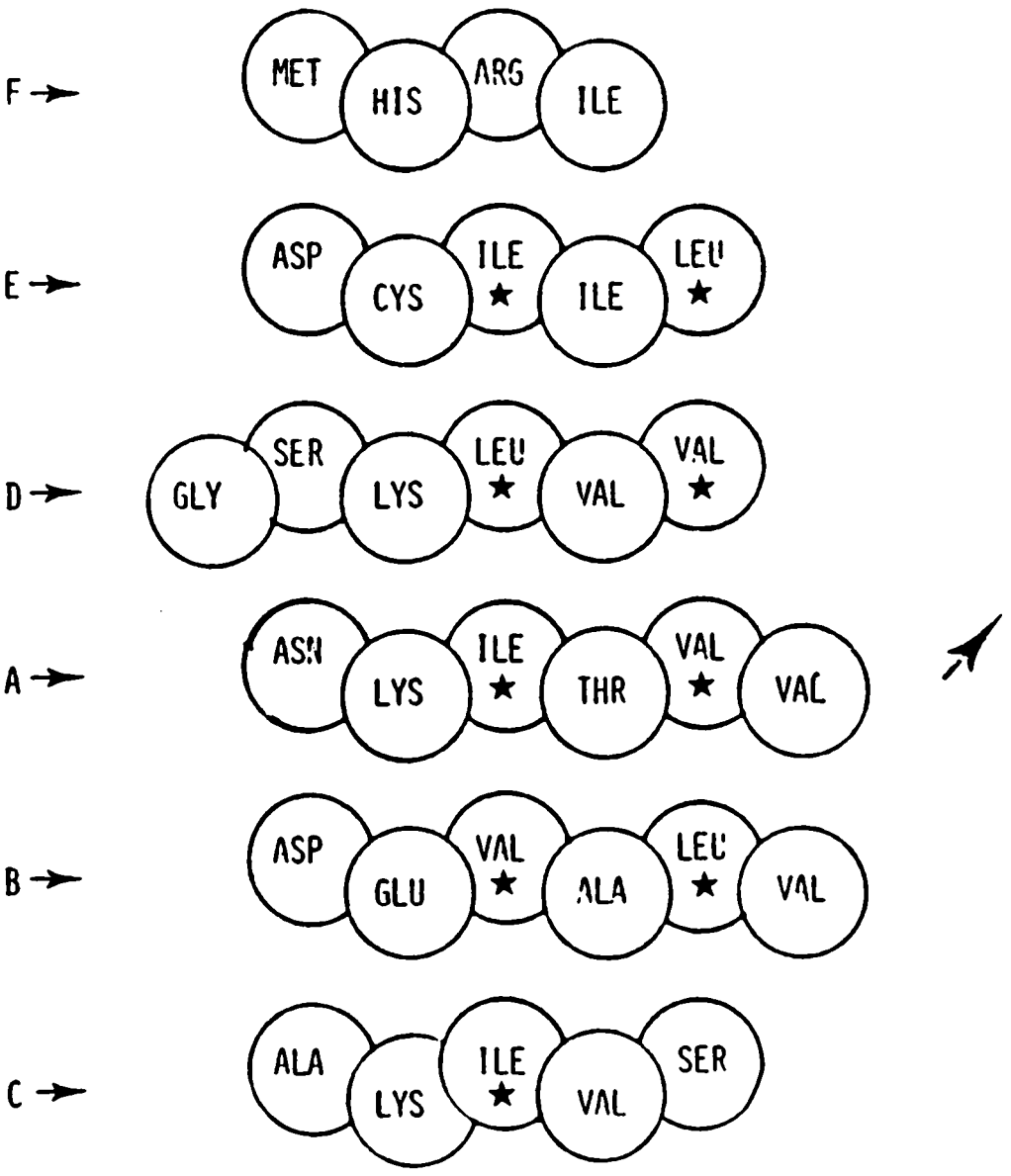
RHODENESE (C-TERMINAL DOMAIN)



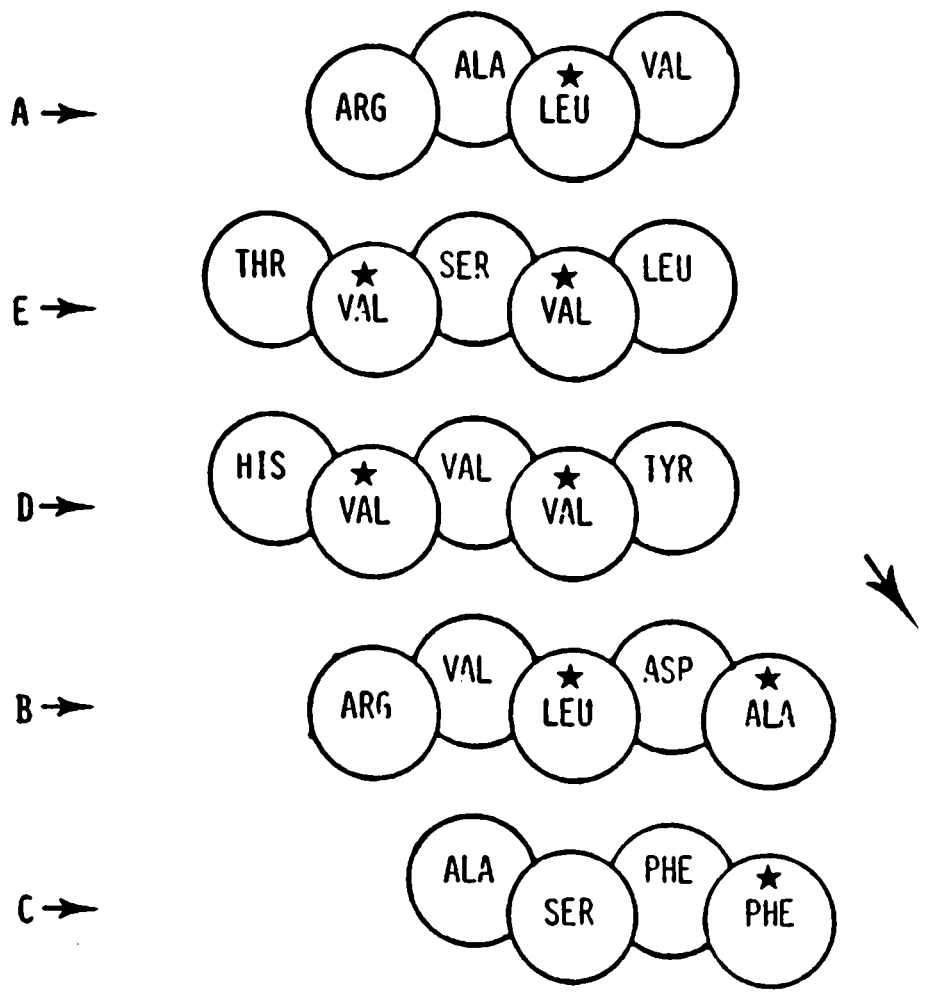
LACTATE DEHYDROGENASE



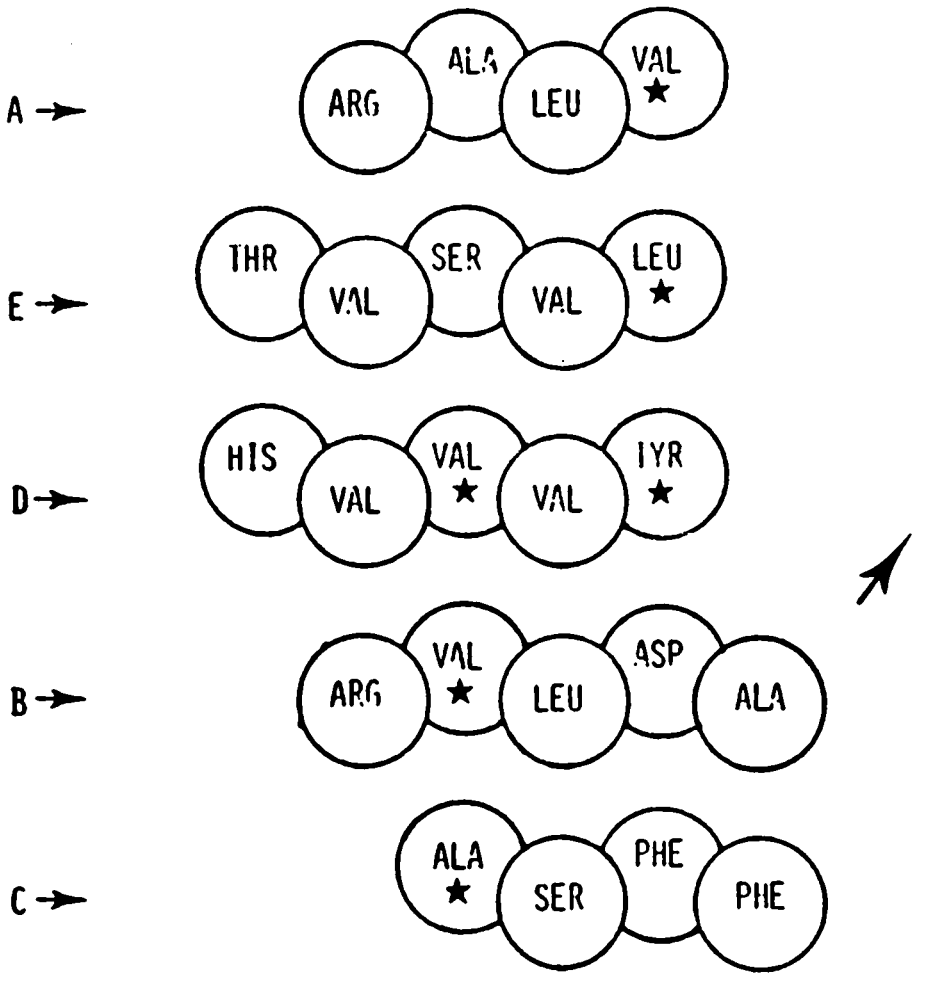
LACTATE DEHYDROGENASE



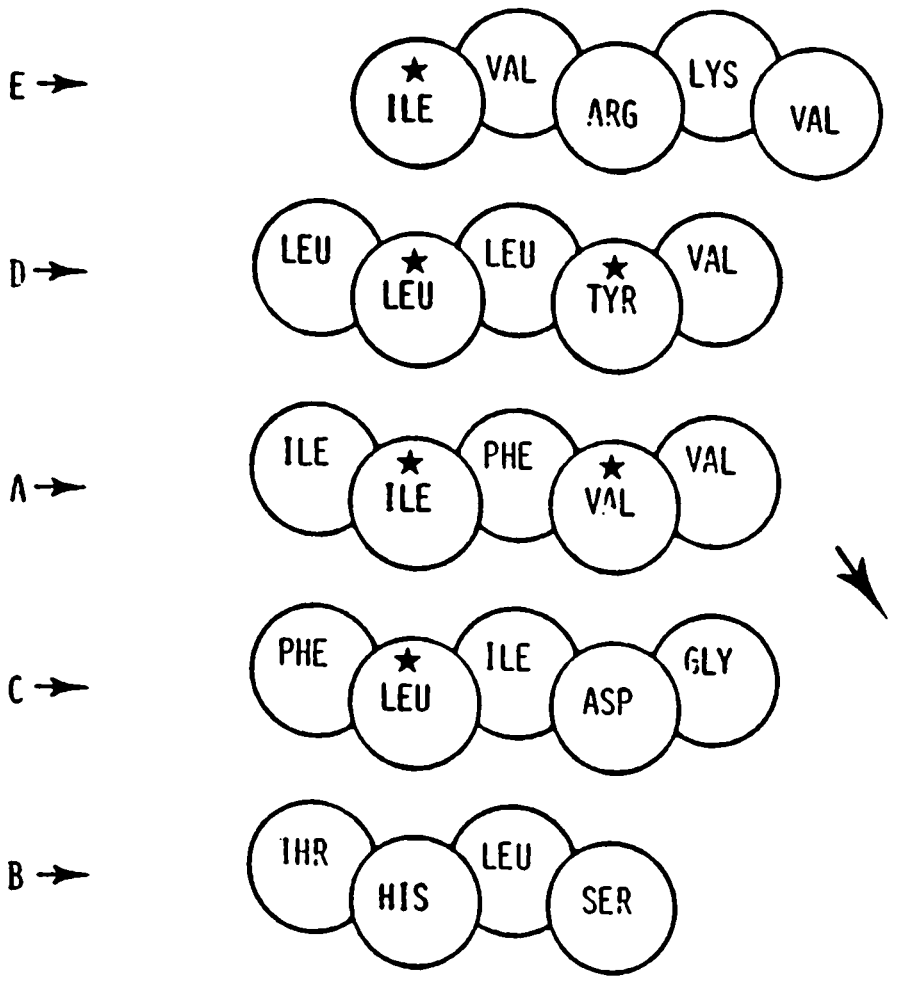
RHODENESE (N-TERMINAL DOMAIN)



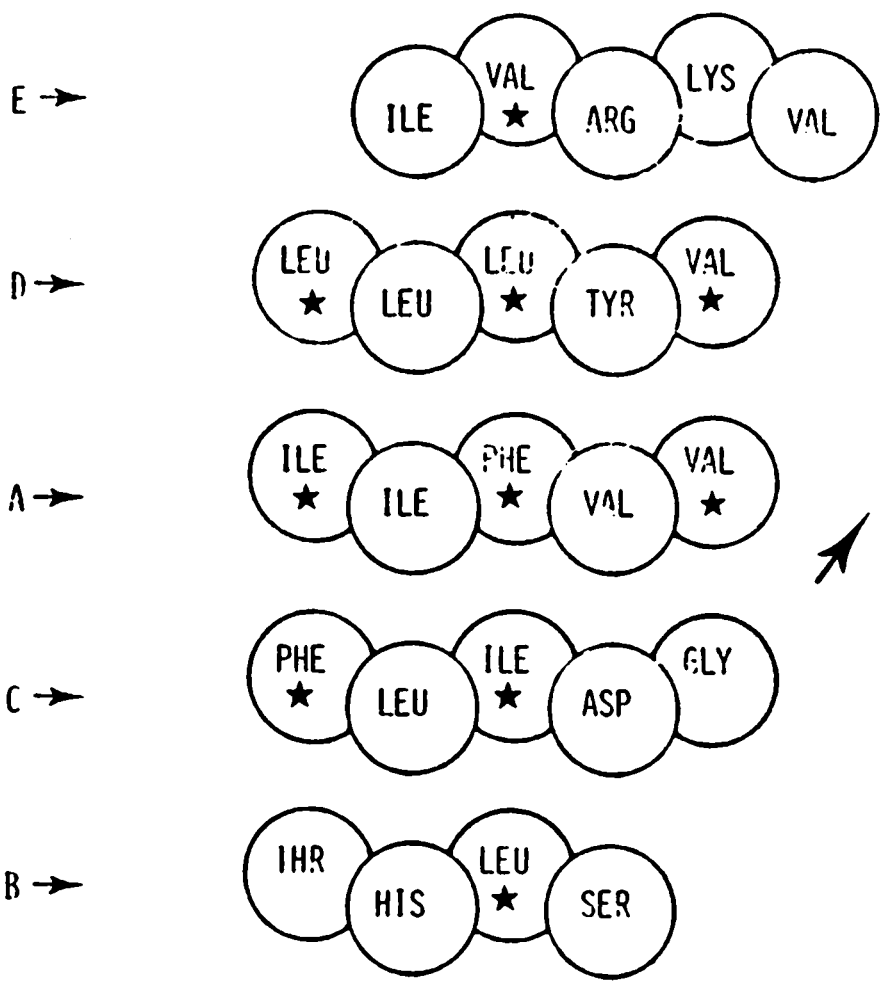
RHODENESE (N-TERMINAL DOMAIN)



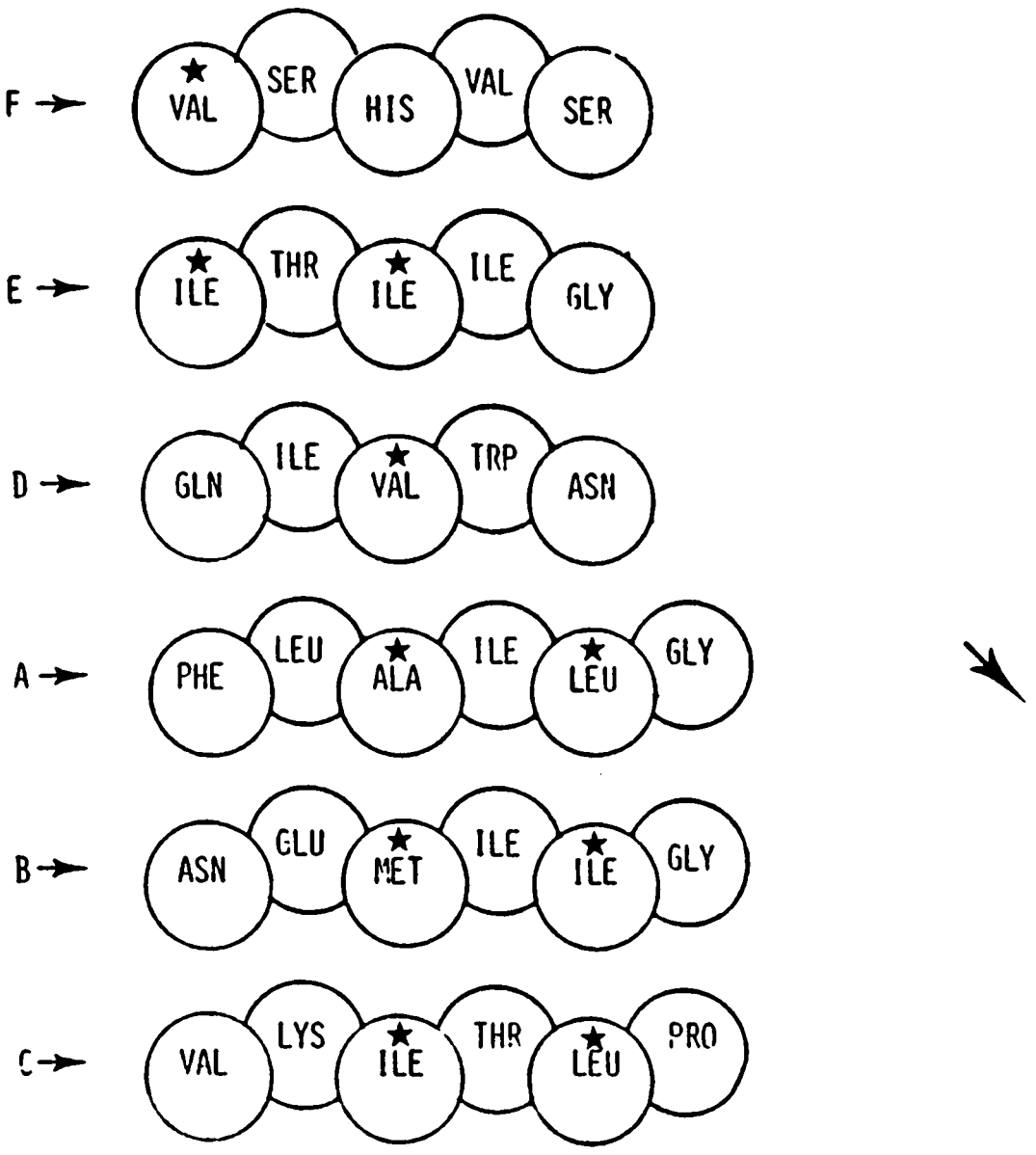
ADENYL KINASE



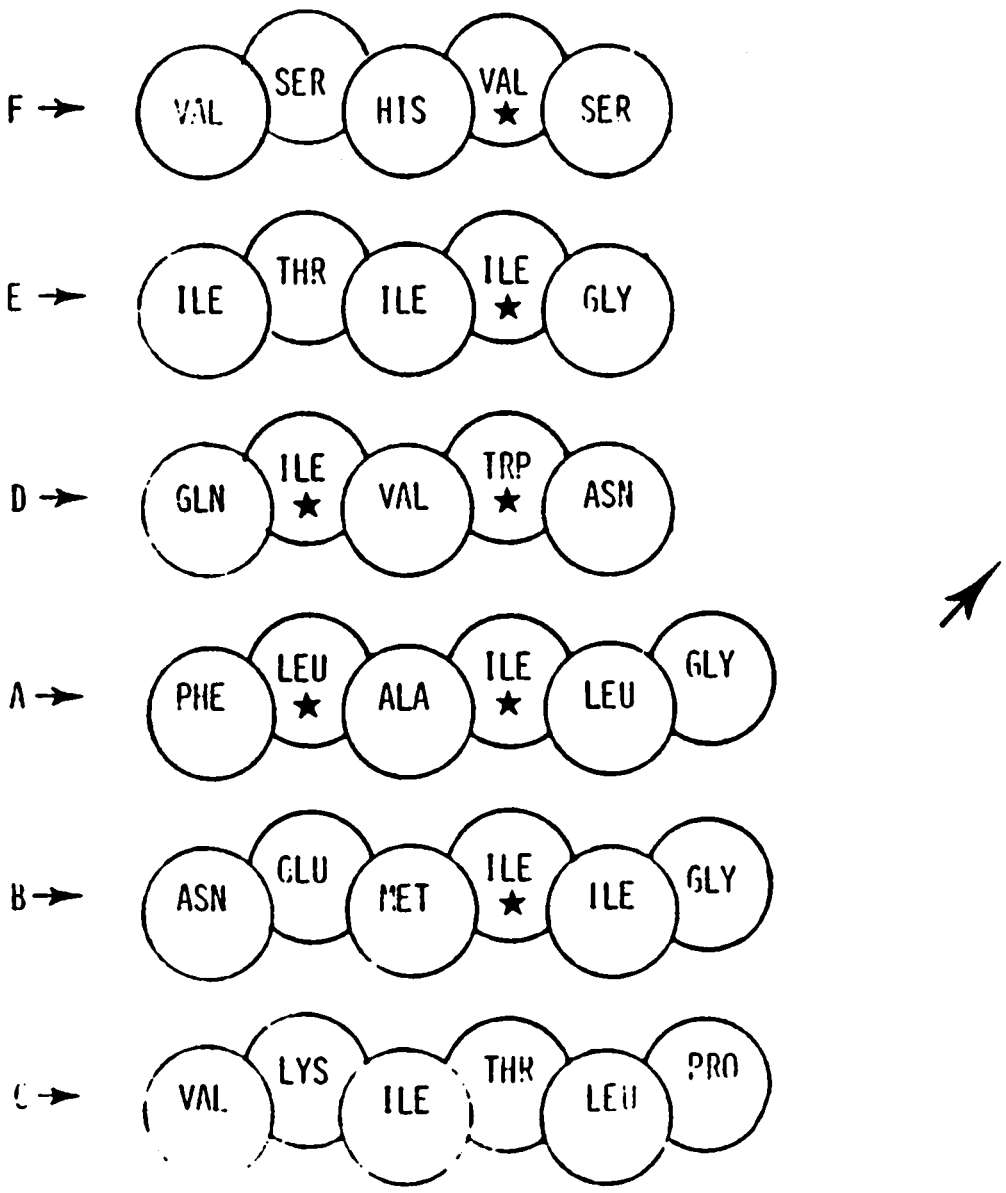
ADENYL KINASE



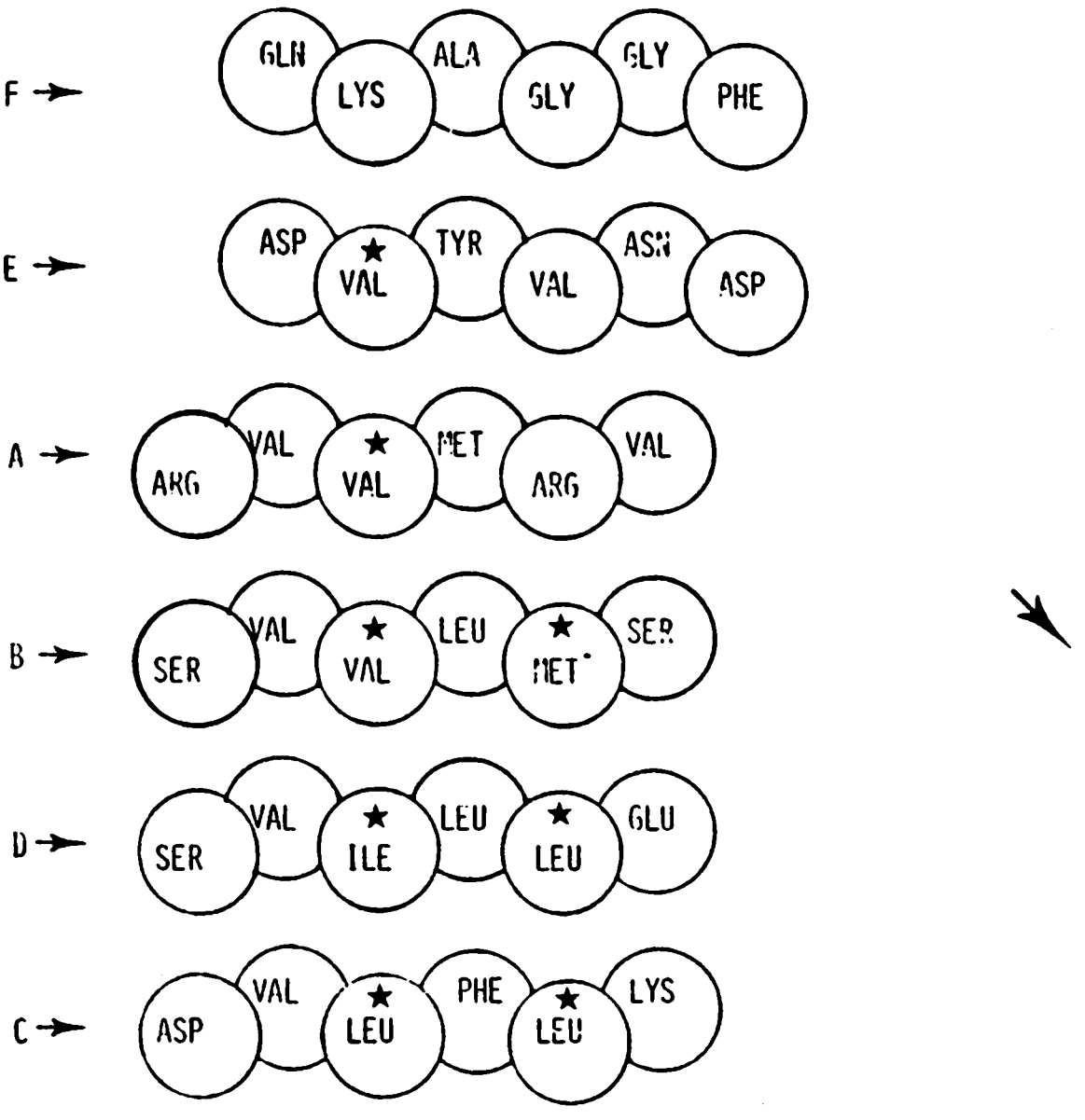
PHOSPHOGLYCERATE KINASE (C TERMINAL DOMAIN)



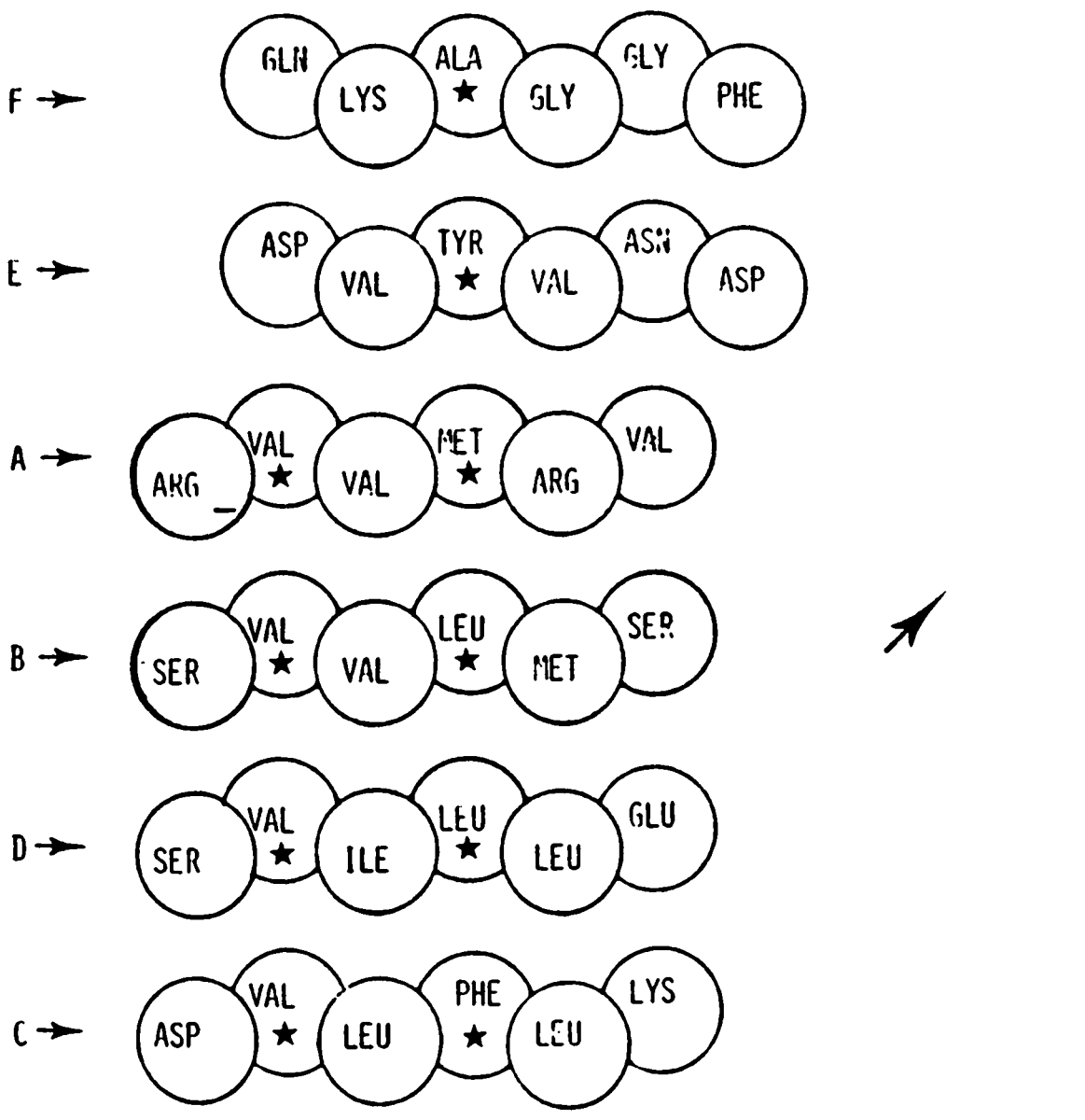
PHOSPHOGLYCERATE KINASE (C TERMINAL DOMAIN)



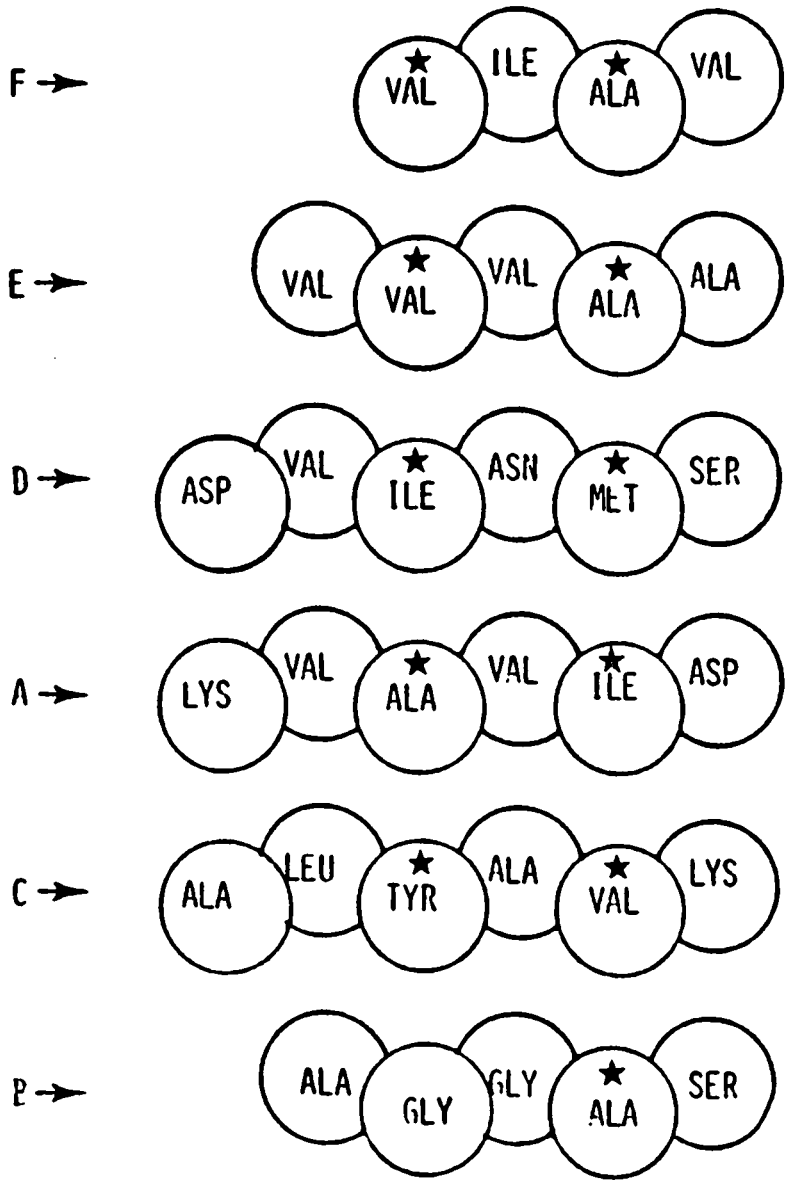
PHOSPHOGLYCERATE KINASE (N-TERMINAL DOMAIN)



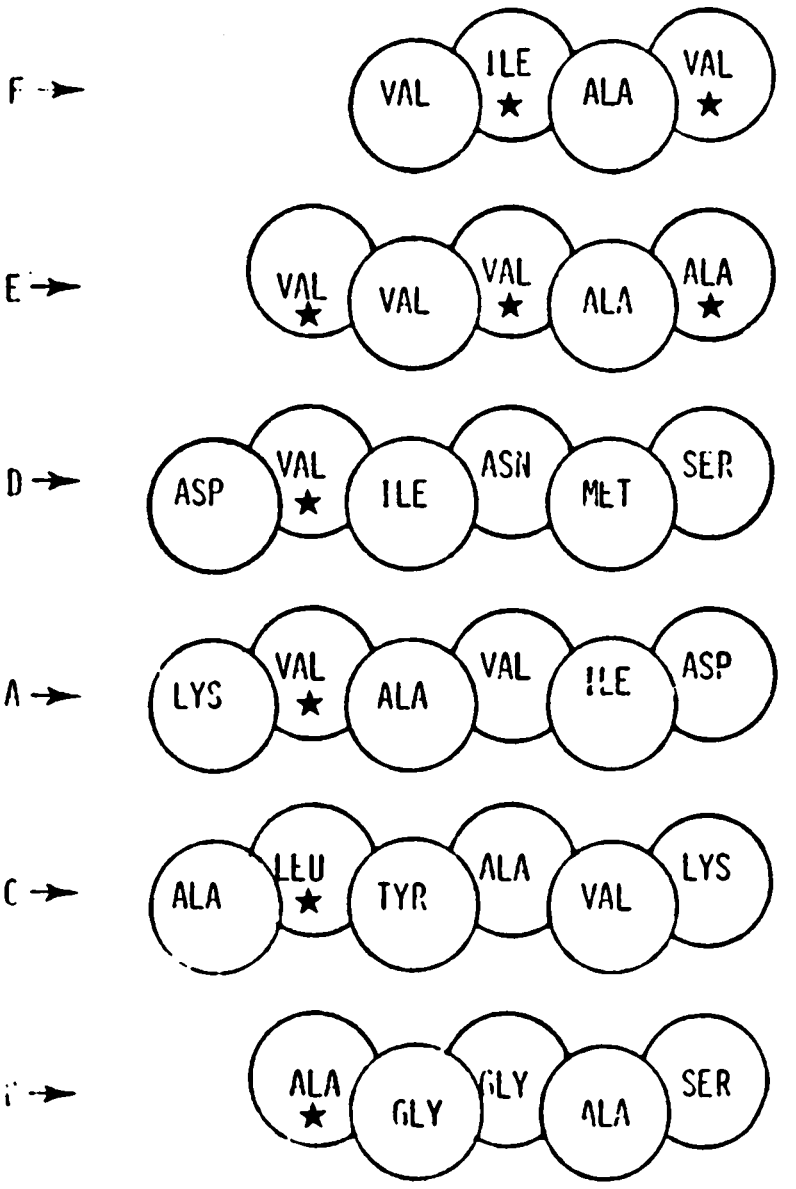
PHOSPHOGLYCERATE KINASE (N-TERMINAL DOMAIN)



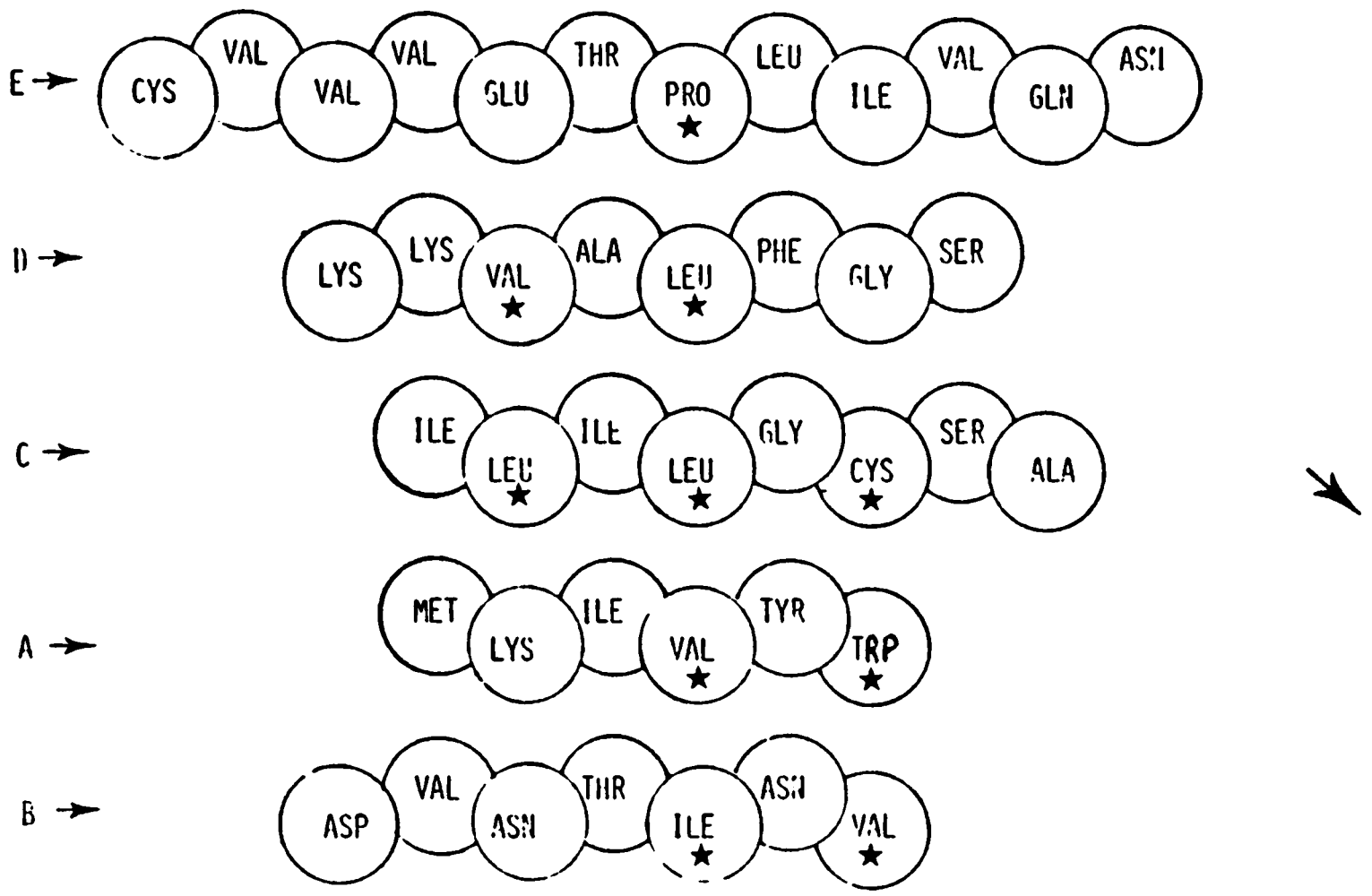
SUBTILISIN



SUBTILISIN



FLAVODOXIN



FLAVODOXIN

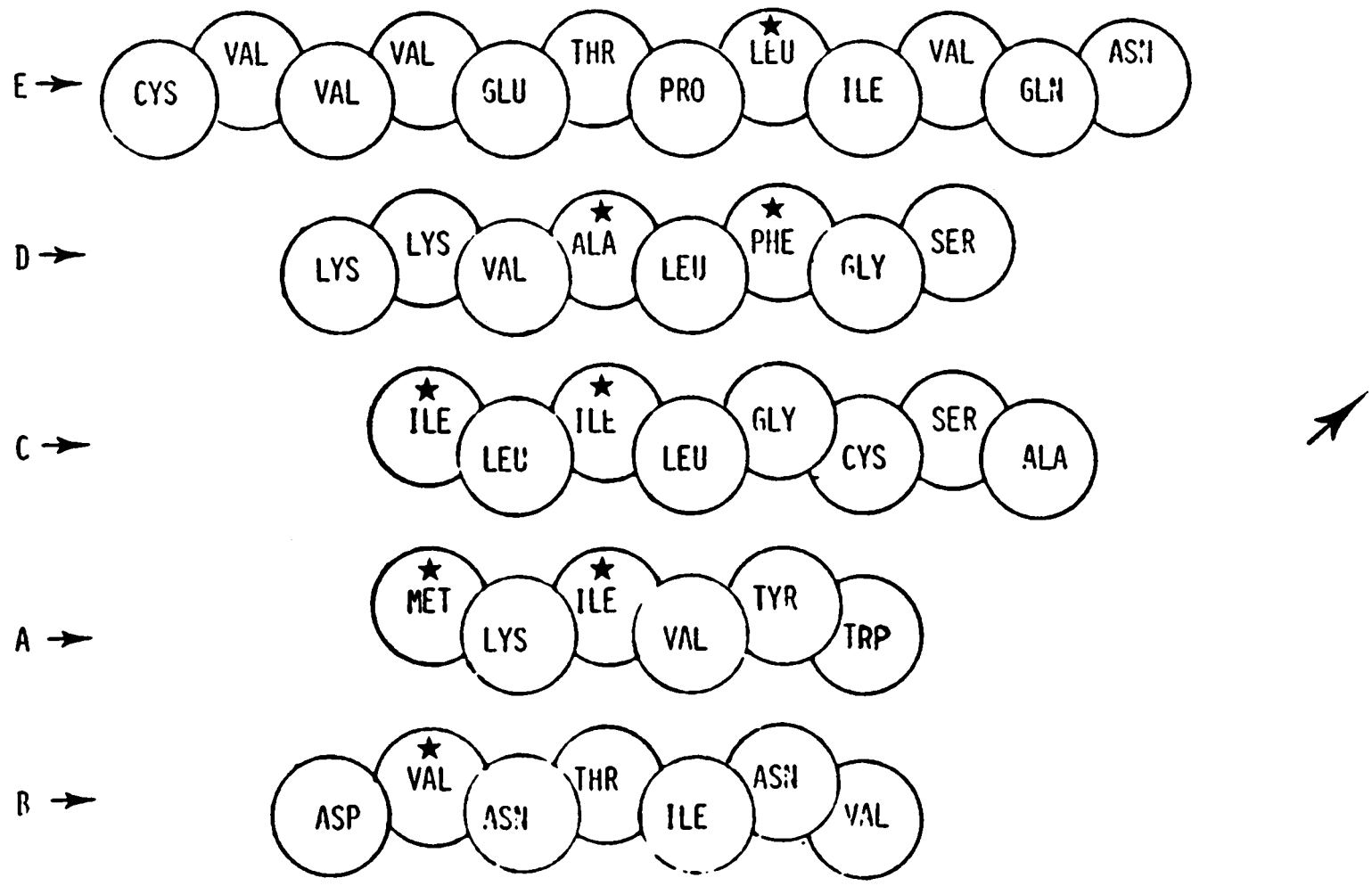


TABLE 5.3

The Value of Structural Restrictions 5-7 on Pure Parallel β-sheets

Protein		Number of Strands	Number of Possible H-bonding Patterns[a]	Number of H-bonding Patterns consistent with Rules 1-4 & End-point restrictions[b]	Number of H-bonding Patterns consistent with Rules 5-7
Flavodoxin		5	1.0×10^6	5120	8
Adenyl Kinase		5	1.0×10^6	6144	11
Phospho-glycerate kinase	N-domain	6	4.7×10^7	40960	47
	C-domain	6	4.7×10^7	12288	2
Alcohol dehydrogenase		6	4.7×10^{11}	57344	148
Triose Phosphate Isomerase		8	1.7×10^6	$1.2 \times 10^6 + 1$[c]	0 + 1
Rhodenese	N-domain	5	1.0×10^6	6144	67
	C-domain	5	1.0×10^6	6144	32
Lactate Dehydrogenase		6	4.7×10^7	53248	96
Subtilisin		6	4.7×10^7	6144 + 6144[d]	0 + 18

[a] Computed from $\frac{1}{2}4^n.2^{n-1}.n!$ where n is the number of strands and 4 possible alignments for each strand are allowed.

[b] See Table

[c] β-Barrel allowed if there are more than 6 strands in the sheet.

[d] Six additional topologies are allowed if a left handed βαβ connection is allowed for the long link between strands B & C.

analyse the βα structures of TIM and SUBT. TIM forms a β-barrel although its secondary structure might easily be accommodated as an 8-stranded parallel β-sheet with 4 helices on each side. The actual number of H-bonds in the β-barrel is 37 due to the highly staggered arrangement of the strands. There are, however, no pure parallel β-sheets with 35 or more H-bonds which satisfy rules 1-7 even though many possible topologies exist. SUBT is the only well documented structure with a left handed βαβ connection. If only the right handed connection were allowed, the second strand which has three hydrophilic residues of the five residues in the strand would have to be internalised to maintain an equal number of helices on each side of the sheet. This would rule out the existence of any potential hydrophobic core (rule 5). If the long connection between the second and third strands permits a left handed connection, then six additional topologies including the observed topology are allowed and the exact H-bonding pattern is produced.

3 Mixed β-Sheets

3.1 The Prediction of Strand Alignments in Mixed Sheets

A more difficult, though related, problem concerns the organisation of β-sheets which are neither pure parallel nor pure antiparallel. These mixed sheets seem to occur in proteins with a combination of β and α structure but always have at least one connecting loop which is too short to tolerate a parallel connection. Although some of these proteins resemble the $(\beta\alpha)_n$ proteins described in the previous section (e.g. B5C, TRDX, PGM$^+$, HKN$^+$, CPA, GPDCAT, GPDNAD), others do not (e.g. CARB, RNS, PAP)*.

$^+$ The chemical sequences of these structures have not been completely determined

* The abbreviations used are b_5-Cytochrome (B5C), thioredoxin (TRDX), carboxypeptidase A(CPA), the catalytic and NAD binding domains of glyceraldehyde 3-phosphate dehydrogenase (GPDCAT, GPDNAD), carbonic anhydrase (CARB), phosphoglycerate mutase (PGM), ribonuclease S (RNS), papain (PAP), and hexokinase (HKN).

3.2 Topological Properties of Mixed Sheets

In contrast to the pure parallel β-sheets where four topological rules severely restricted the number of reasonable paths for the polypeptide chain, much more diversity seems possible for mixed sheets. Although the connection between parallel β-strands remains right-handed (Sternberg & Thorton, 1976) and unknotted (Richardson, 1977), there can be more than one chain reversal (see Table 5.4). Moreover, the helices joining the strands do not form an equal covering of the sheet. Carboxypeptidase A has four helices on one side and two helices on the other side of its β-sheet (see FIgure 5.4).

A survey of some other topological properties suggests that adjacency is directly correlated with the size of the sheet (see Table 5.5). This is consistent with some earlier work on adjacency by Sternberg & Thorton (1977b). One curious relationship found in Table 5.5 is that the difference between the number of strands (S) and the number of helices between strands (H) is always greater than or equal to the number of antiparallel connections. This might suggest that helices promote parallel connections, but this is clearly not the case (see Table 5.5). One-third of the connecting loops with helices form antiparallel connections. S - H is also greater than or equal to the number of reversals for all of the mixed sheets other than b_5-cytochrome which is a fragment of a membrane bound protein. The existence of one more strand in the cleaved fragment could preserve this relationship. Both of these relationships are true for all parallel β-sheets. However, their existence may be fortuitous and largely due to the limitations of the data base.

Richardson (1977) has noted the tendency for β-sheets to minimise the number of mixed strands, strands with hydrogen bonding patterns which are characteristic of pure parallel sheets on one side and antiparallel sheets on the other. The number of possible mixed strands increases as the number of strands in the sheet. While the maximum number of mixed

TABLE 5.4

A SURVEY OF SOME TOPOLOGICAL PROPERTIES OF MIXED β-SHEETS

Protein	Number of Strands	Number of Helices between Strands	Number of Reversals[b]	Number of Antiparallel Connections	Adjacency	MI*
Cytochrome b_5	5	3	3	2	1	2
Thioredoxin	5	3	2	2	2	1
Ribonuclease S	5[d]	1	0	3	3	0
Phospho-glycerate mutase	6	4	1	2	2	1
Hexokinase I[a]	6	3	2	3	3	1
II	5	2	1	2	3	1
Papain	7[d]	3	2	3	3	0
Carboxy-peptidase A	8	4	3	4	3	2
Glyceraldehyde 3-phosphate dehydrogenase						
Catalytic	7	2	2	4	3	3
NAD Binding	9	3	2	2	5	1
Carbonic Anhydrase	10	4	6	6	5	2

* Mixed Interfaces

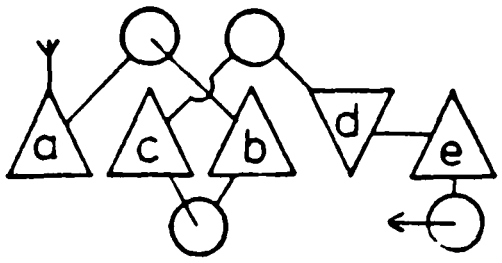
[a] I refers to the sheet with strands AKJINO and II to the sheet with DCBEH (see Figure 5.4).

[b] A reversal occurs when there is a change in the sense of the strand order e.g. CAB - starting at A, B is to the right of A but C is to the left of B. For pure parallel β-sheets, this number was at most 1.

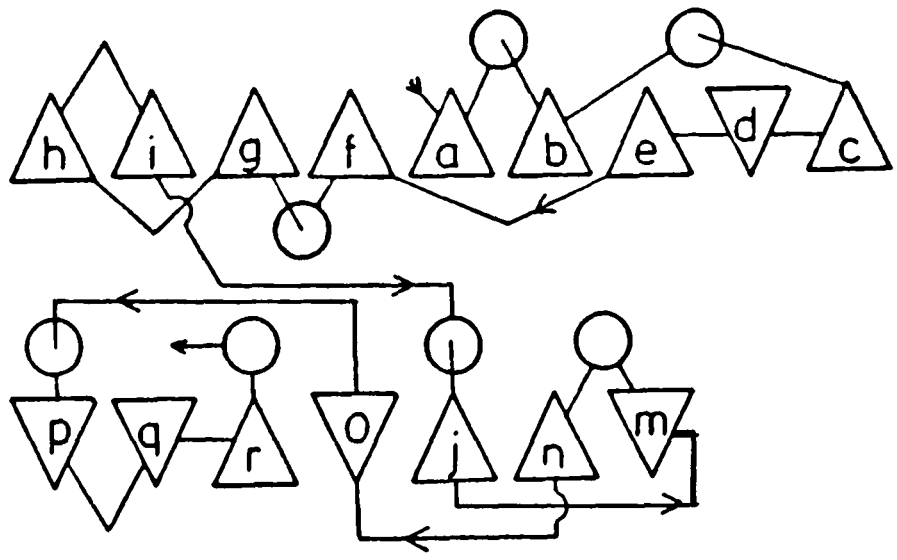
[c] Adjacency is computed as the number of sequential strands that are nearest neighbours in a sheet.

[d] Two sequential distinct segments form one effective strand.

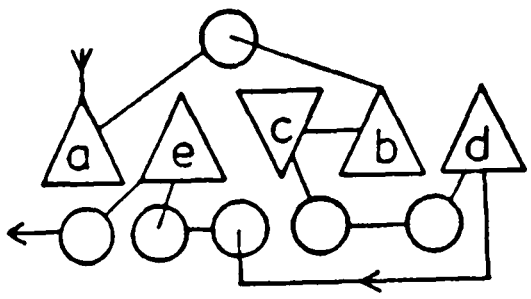
FIGURE 5.4. Schematic Diagrams of "Mixed Sheet" Proteins.



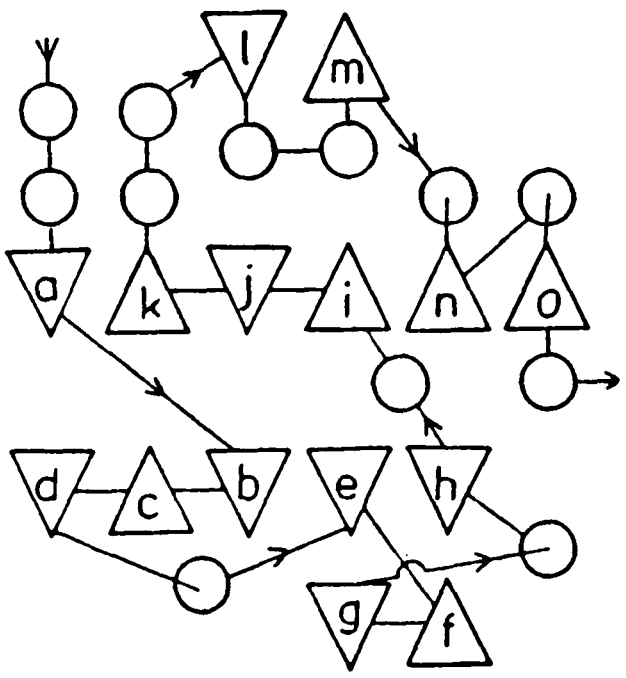
TRDX



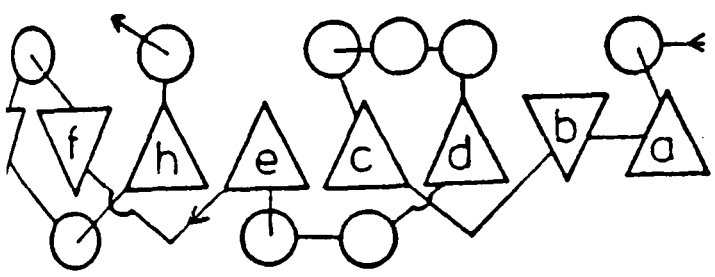
GPDH



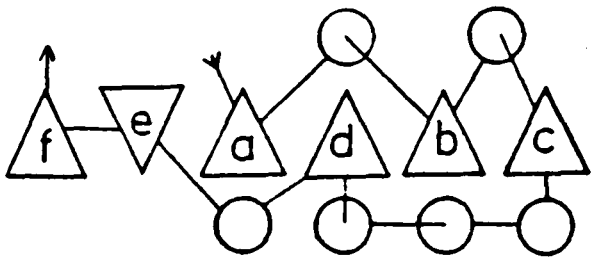
CYT B5



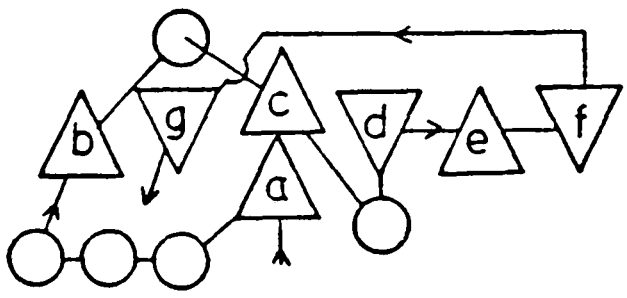
HKN



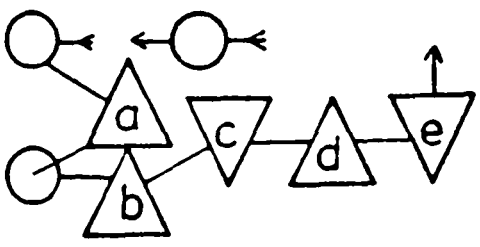
CPA



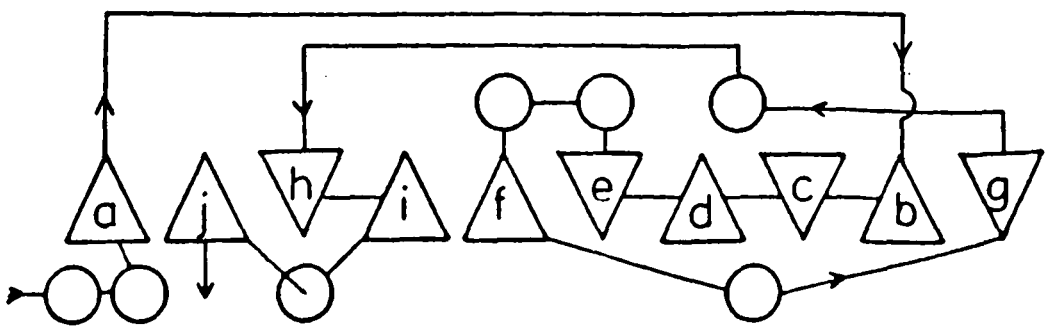
PGM



PAP



RNS



CARB

Schematic diagrams of the "mixed sheet" proteins considered. β-sheets are viewed along their strand direction. Each strand is represented by a triangle whose apex points up or down according to whether the strand is viewed from the N- or C-terminus. A circle represents an α-helix.

TABLE 5.5. Connecting Loops in Mixed β-sheets.

Protein & Connecting Loop		Loop Length and Structure	
		Parallel	Anti-parallel
Cytochrome b_5	A-B	14-coil & Helix	
	B-C		3-coil
	C-D		20-coil & Helix
	D-E	22-coil & Helix	
Thioredoxin	A-B	14-coil & Helix	
	B-C	24-coil & Helix	
	C-D		17-coil & Helix
	D-E		5-coil
Carboxypeptidase A	A-B		13-coil
	B-C		7-coil
	C-D	38-coil & Helix	
	D-E	81-coil & Helix	
	E-F		4-coil
	F-G	25-coil & Helix	
	G-H		24-coil & Helix
Glyceraldehyde	A-B		27-coil
3-phosphate	B-C		16-coil & Helix
dehydrogenase	C-D		5-coil
Catalytic	D-E	26-coil & Helix	
Domain	E-F	16-coil	
	F-G		10-coil
Glyceraldehyde	A-B	20-coil & Helix	
3-phosphate	B-C	24-coil & Helix	
dehydrogenase	C-D		3-coil
NAD binding	D-E		3-coil
domain	E-F	16-coil	
	F-G	19-coil & Helix	
	G-H	7-coil	
	H-I	14-coil	
Carbonic Anhydrase	A-B	14-coil	
	B-C		2-coil
	C-D		17-coil
	D-E		18-coil
	E-F		16-coil & Helix
	F-G		23-coil & Helix
	G-H	14-coil & Helix	
	H-I		17-coil
	I-J	34-coil & Helix	
Papain	A-B	103-coil & Helices	
	B-C	16-coil & Helix	
	C-D	30-coil & Helix	
	D-E		2-coil
	E-F		10-coil
	F-G	15-coil	

cont.

TABLE 5.5, cont.

Protein & Connecting Loop		Loop Length and Structure	
		Parallel	Anti-Parallel
Ribonuclease S	A-B*	12-Coil & Helix	
	B-C		19-coil
	C-D		11-coil
	D-E		5-coil
Phosphoglycerate Mutase	A-B	44-coil & Helix	
	B-C	23-coil & Helix	
	C-D	98-coil & Helices	
	D-E		31-coil & Helix
	E-F		13-coil

* A & B form different parts of the same strand.

strands is n-2, where n is the number of strands, the number of mixed strands observed is always less than n/2 (see Table 5.5) and frequently much smaller.

An analysis of strand orders in all known β-sheets reveals one further restriction: if i, j, k and ℓ are sequential strands in a sheet, then the strand order is never k, i, ℓ, j or j, ℓ, i, k. These convoluted strand orders will be known as pretzels or reverse pretzels. In the two extreme situations, it is easy to understand why this highly convoluted strand order is never seen: if all strands are parallel, a knot is created (Richardson, 1977) and if all strands are antiparallel a crossover of connecting loops is forced (Ptitsyn et al., 1979). Moreover, the fact that this order is never observed is very significant as approximately one half of all possible 6-stranded sheets contain pretzels or reverse pretzels. As the number of strands in the sheet increases, the fraction of strand orders which are allowed by this rule decreases rapidly (see Figure 5.5).

These observations can be formulated into three rules which restrict the number of allowed strand orders and topologies:

M1) The connection between parallel strands is right handed;

M2) The number of mixed strands is minimal. In the prediction of an unknown structure, this value would be increased from 1 to 2 and so on until a reasonable structure is found. In practice, this value is set to the observed value; and

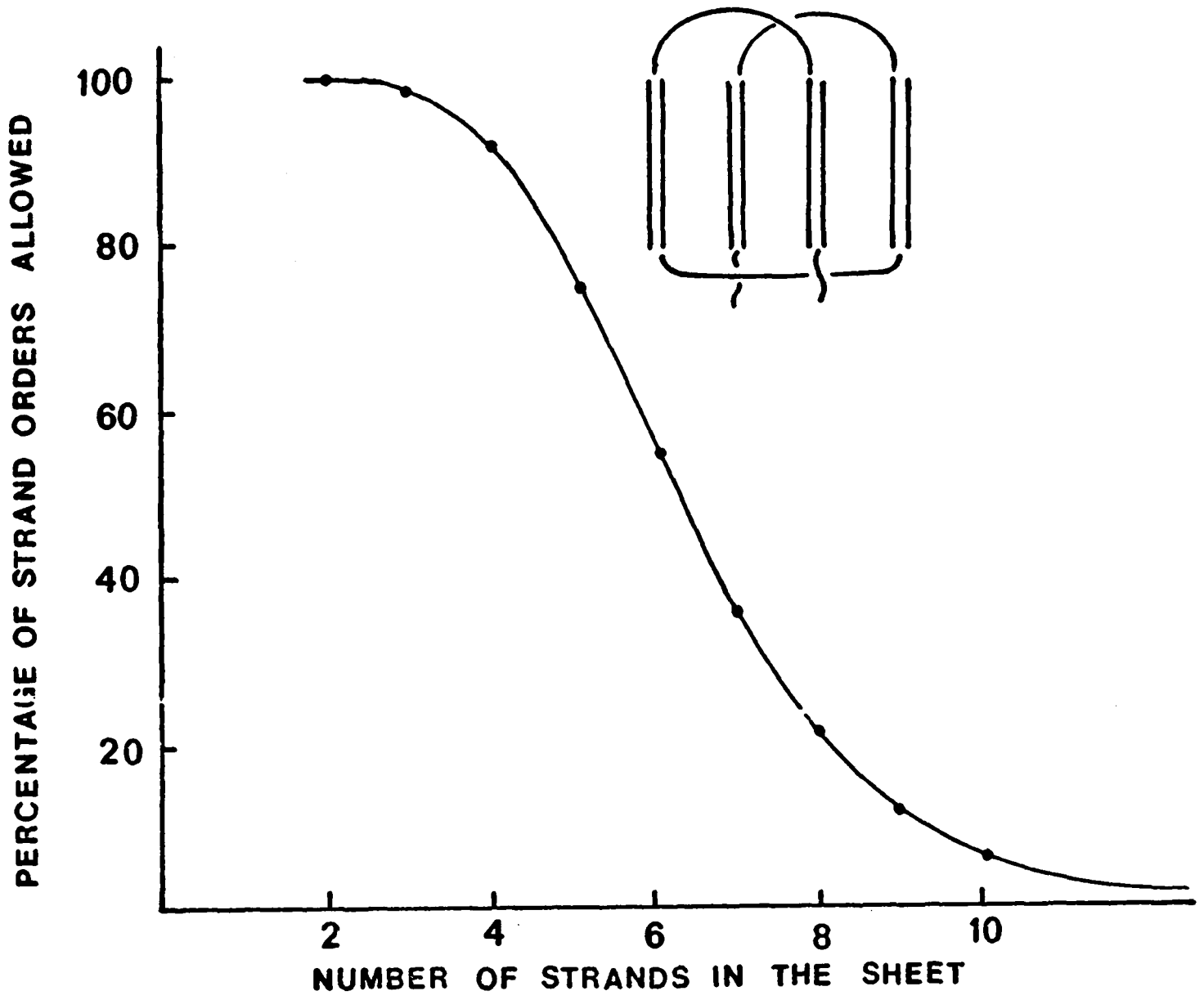
M3) No pretzels or reverse pretzels are permitted.

3.3 The Alignment of Residues in Mixed β-sheets - Prediction

Although mixed sheets do not share many of the restrictive topological properties of pure parallel sheets, they do have similar arrangements of hydrophobic residues on the two faces of the β-sheet. The lack of a regularly repeating βα unit precludes the formation of a uniform covering of the β-sheet by helices. Thus, rules 5-7 from pure parallel sheets will

FIGURE 5.5

Probabilities for Pretzel Strand Orders in β-sheets.



All possible strand orders were generated for β-sheets with 2 to 10 strands. The percentage of sheets which did not have a pretzel strand order is plotted as a function of sheet size. A pretzel occurs when four strands have the sequential order a,b,c,d and the sheet order b,d,a,c or c,a,d,b as shown in the schematic β-sheet in the upper right hand corner of the graph. Although pretzels are potentially very frequent in sheets with six or more strands, they are never observed in protein structure.

have to be generalised slightly. A survey of 8 mixed sheets (B5C, TRDX, CPA, GPDCAT, GPDNAD, RNS, PAP and CARB) revealed three restrictions for allowed H-bonding patterns.

M4) There exists a central core of hydrophobic residues in two adjacent rows, i, i+1, in all of the non-edge strands where i is within one residue of the strand midpoint. The number of hydrophilics in this region is small and typically 1 or 2 but is 6 in the case of the 10-stranded sheet of carbonic anhydrase which is covered by only 5 helices.

M5) Potential hydrogen bonding between strands is always within 2 of the maximum number possible. This is quantified by strand overlap.

M6) If a hydrophobic island is defined as an uninterrupted row of two or more hydrophobic residues which align, then the mid-points of these islands must have:

$$h_i > h_{i+2} > \ldots. > h_k$$

$$h_{i+1} < h_{i+3} < \ldots. < h_{k+1}$$

for some i. This implies that the groups of hydrophobics in alternating rows progress from the lower left to the upper right corner of the sheet while the intervening rows proceed from the upper left to the lower right hand corner. The residues involved in the lower left to upper right patch are on the top of the sheet when the strands run from right to left across the page and the first strand is above the second strand.

A compute program, MIXSHEET (see Appendix V) was written to apply constraints M1 - M6. As input, the program uses the set of all strand orders which could have an overlap consistent with rule M2 together with the amino acid sequence and strand assignments for each strand, 4 possible positions for the strand midpoint and 2 different directions. This produces a list of 8^n hydrogen bonding diagrams where n is the number of strands for each strand

order. Each structure is tested against rules M1 - M6 and acceptable hydrogen bonding diagrams are output (see Figure 5.6). The computation time ranges from 30 seconds for TRDX to over 10 hours for one strand order for carbonic anhydrase on an ICL 2980. Unless severe restrictions on topologies or strand orders can be found this approach is not practical for sheets with more than 8 strands. For hexokinase whose sequence is not completely known and so not included in this work, the additional problem of segregating one set of strands into two sheets arises. Ribonuclease S and papain present a slightly different problem. In each of these proteins, one strand in the sheet is composed of two sequentially distinct segments. Clearly, more insight into the nature of mixed sheets will be required to resolve these additional complications.

3.4 Results and Discussion

MIXSHEET was uniformly applied to five β-sheets in four proteins: TRDX, B5C, CPA and the NAD binding and catalytic domains of GPD. Carbonic anhydrase was not studied because the computation time required would be approximately 270 hours on an ICL 2980. Attempts to increase the efficiency of the MIXSHEET algorithm and simplify the problem of examining a 10-stranded sheet with 1.9×10^{18} possible structures are in progress.

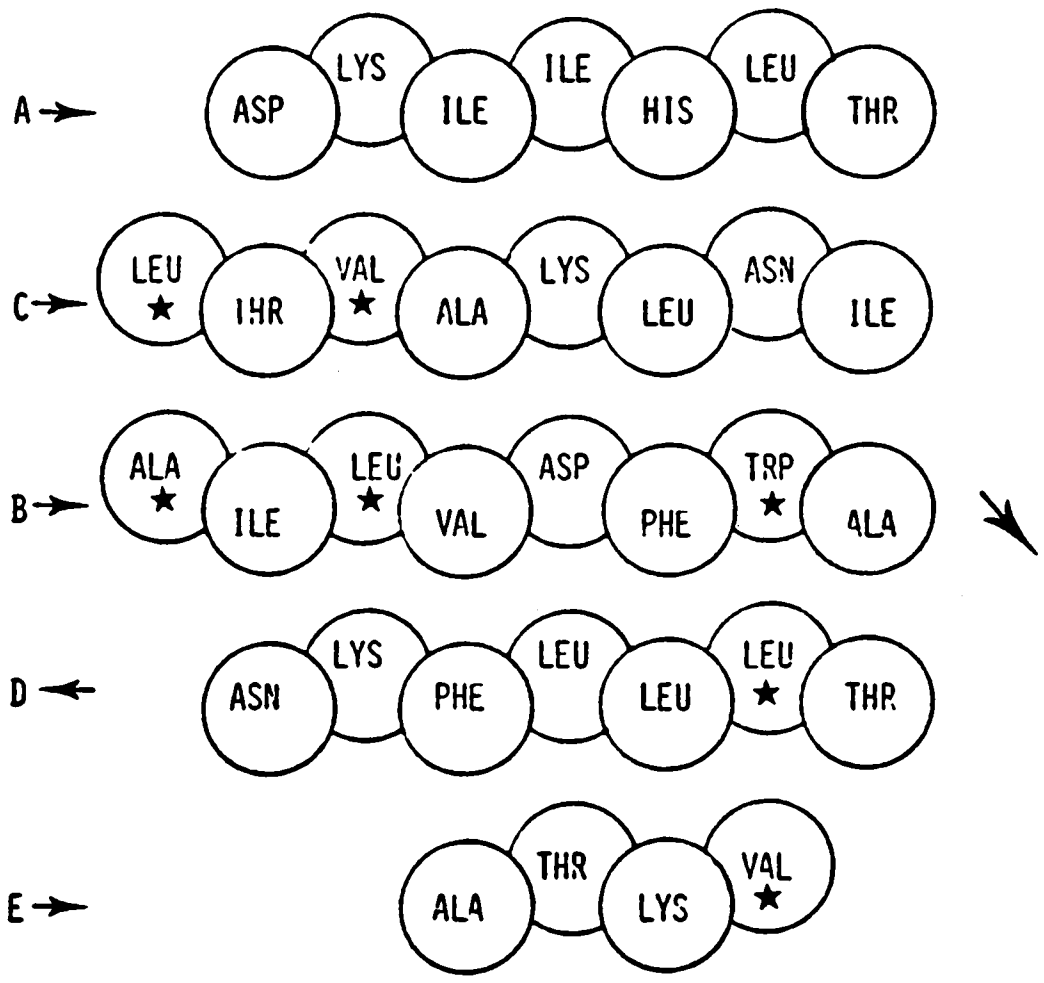
Table 5.6 presents the results for the five β-sheets studied. In spite of the drastic increase in the number of possible structures, the number of allowed structures remained reasonably small. In every case, the exact hydrogen bonding pattern was one of the allowed structures. Since the existence of helical segments linking the strands in the sheet is ignored for all intensive purposes in this calculation, it is surprising that the results are so favourable. Perhaps some of the alternative hydrogen bonding patterns which are consistent with rules M1 - M6 will leave hydrophobic clusters of residues uncovered by helices, or will force some helices to pack against a hydrophilic portion of the sheet. Hopefully, the superiority

FIGURE 5.6

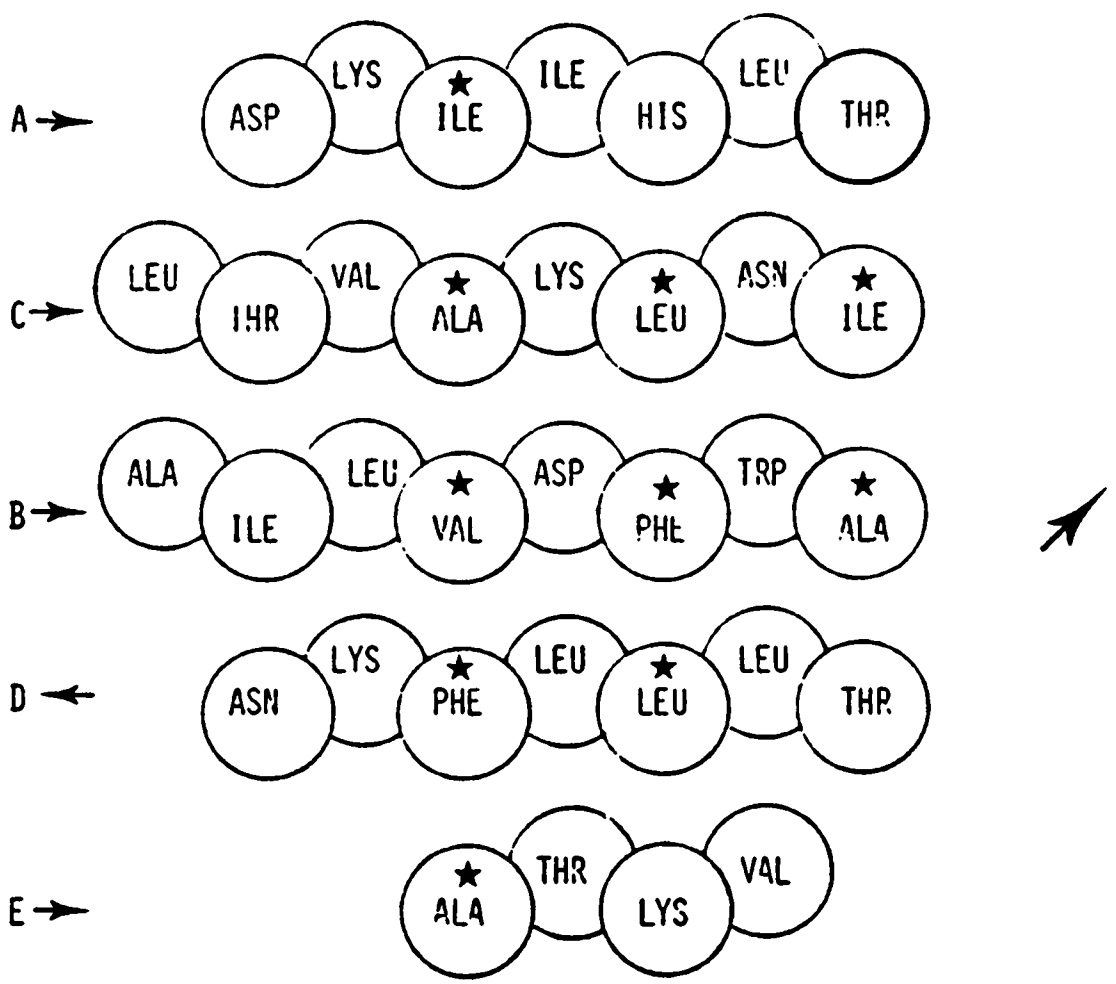
Bubble Diagrams of Mixed β-Sheets

The strand alignment of eight mixed β-sheets in seven proteins is shown in the bubble diagrams on the facing and following pages. The residues which contribute to the hydrophobic patch predicted by rules M4 - M6 are shown with stars. Residues on the top face of the β-sheet which contribute to the patch are starred in the upper or left frame. Residues on the bottom face of the β-sheet which contribute to the patch are starred in the lower or right frame. Arrows indicate the direction of the constellation of hydrophobic residues and their generic anti-complementarity. For thioredoxin, cytochrome b_5, carboxy-peptidase A, and both domains of glyceraldehyde 6-phosphate dehydrogenase, these are the alignments that are predicted by MIXSHEET. The size of carbonic anhydrase and the strand discontinuities in ribonuclease S and papain make their analysis by MIXSHEET impossible at present. These three proteins still do fit rules M1 - M6.

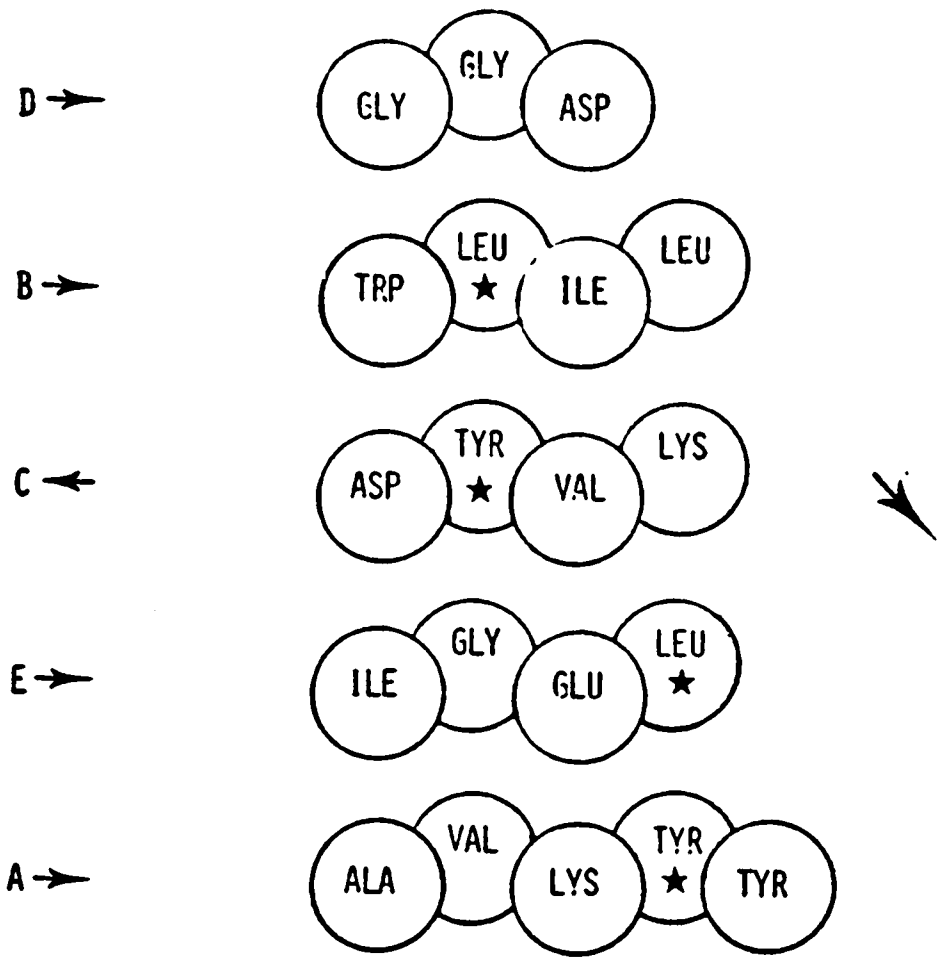
THIOREDOXIN



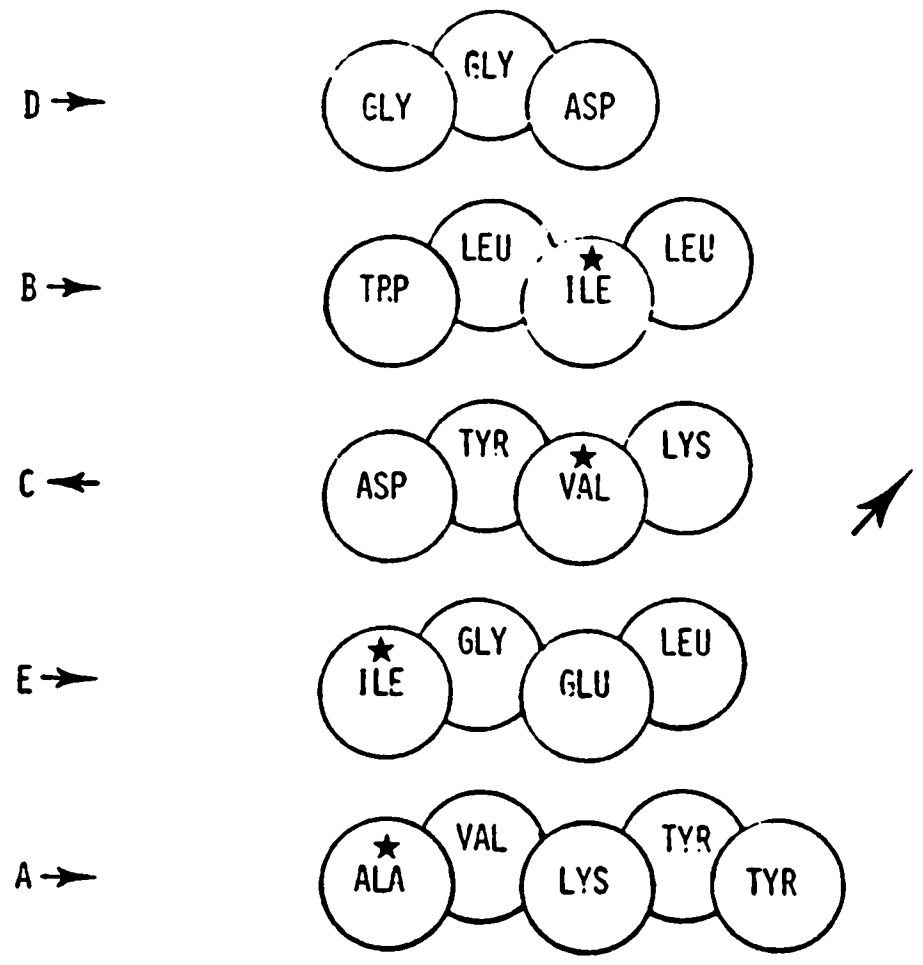
THIOREDOXIN



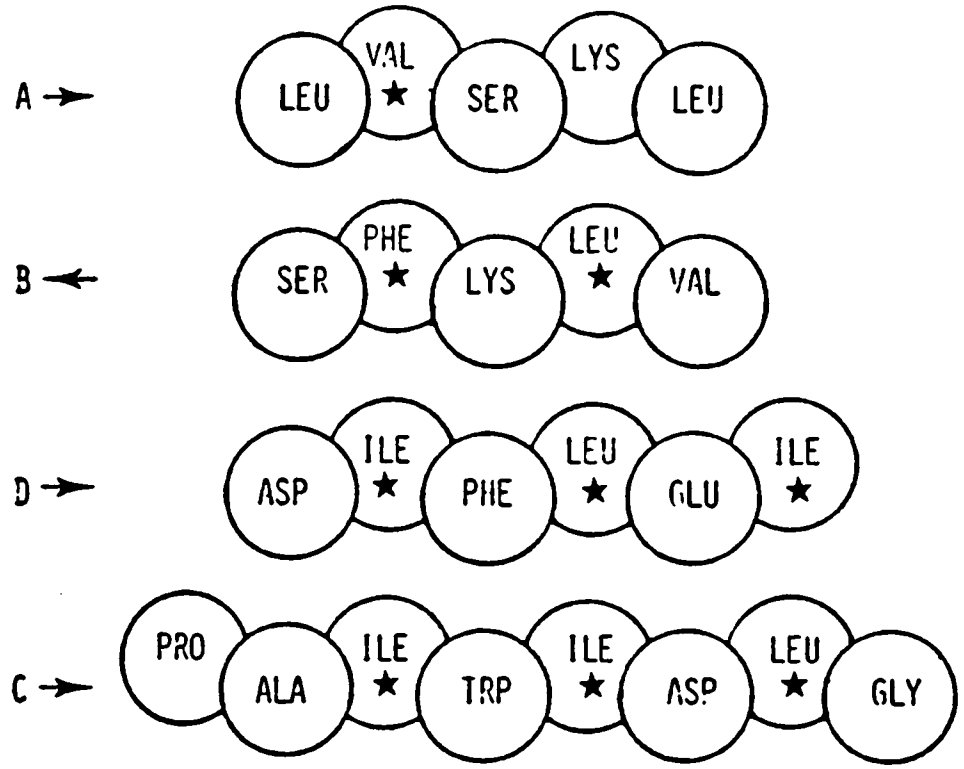
CYTOCHROME B5



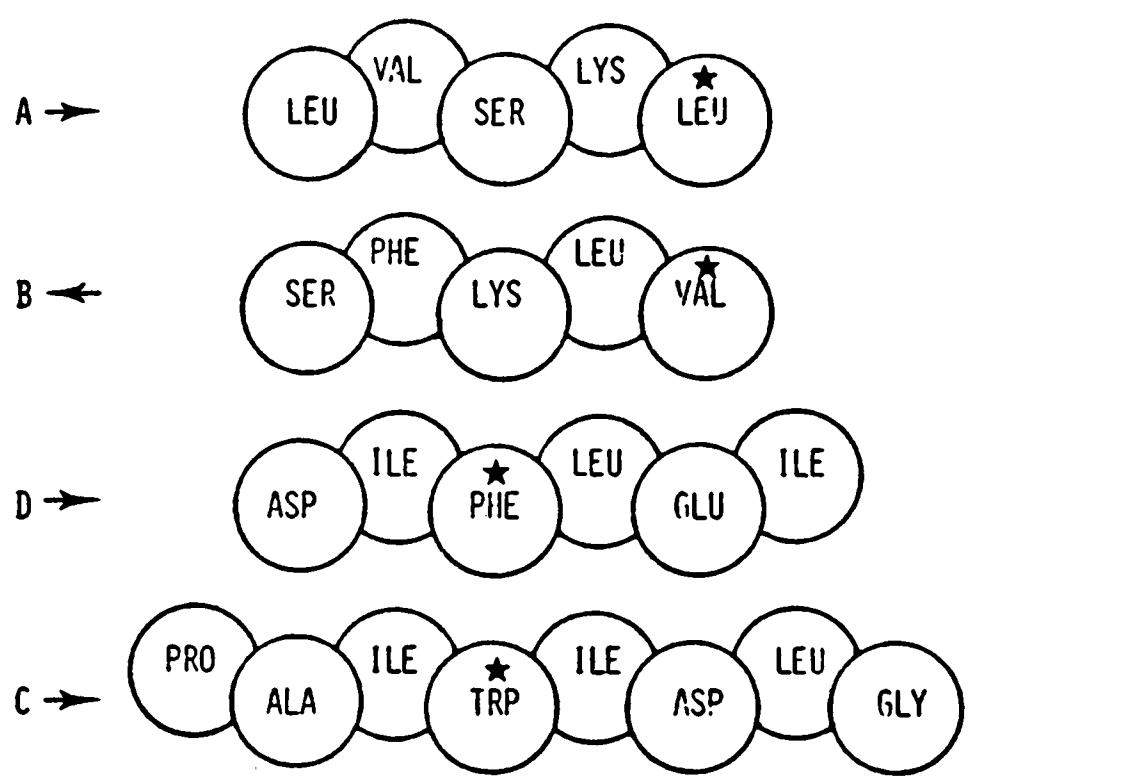
CYTOCHROME B5



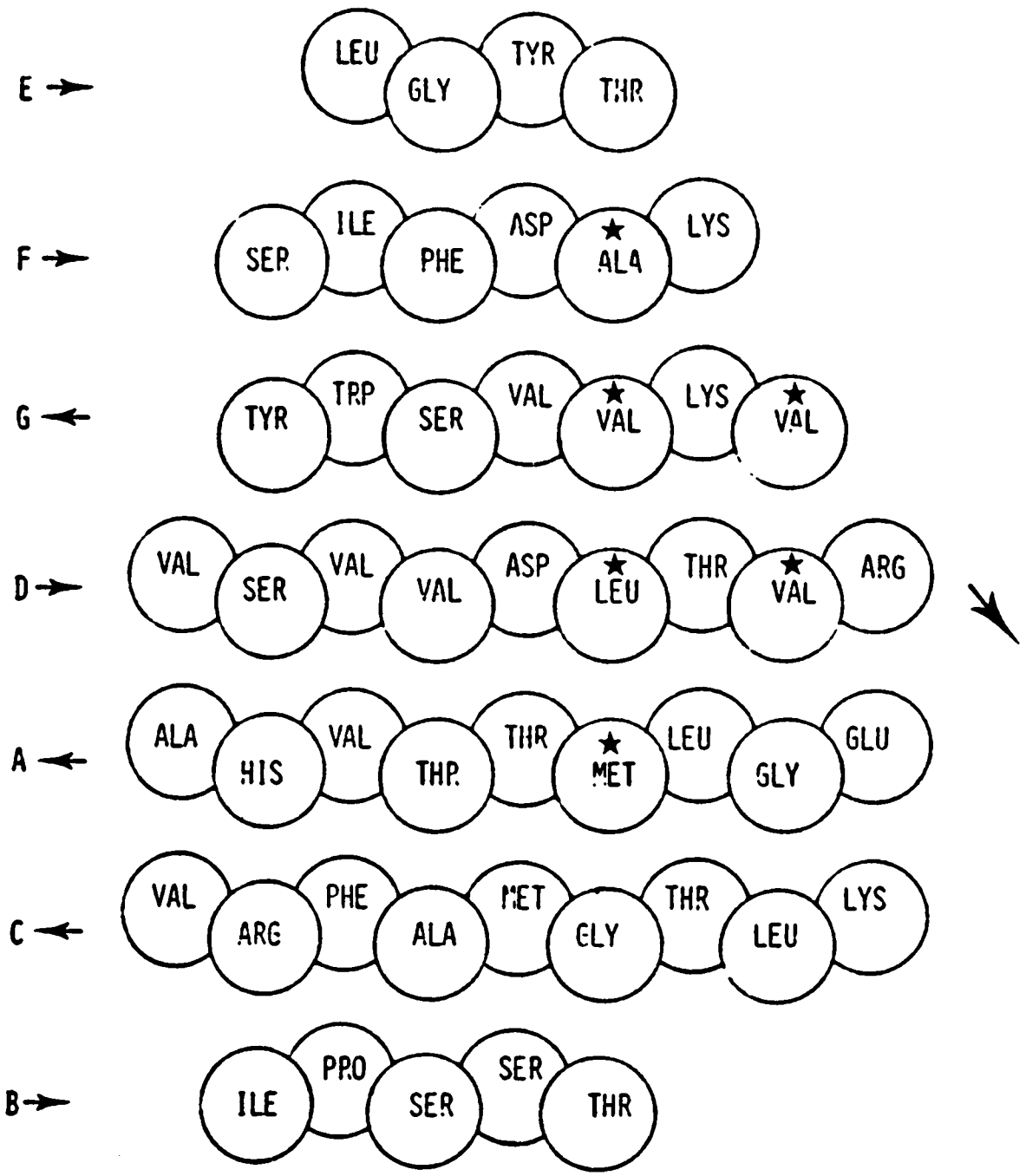
CARBOXYPEPTIDASE A



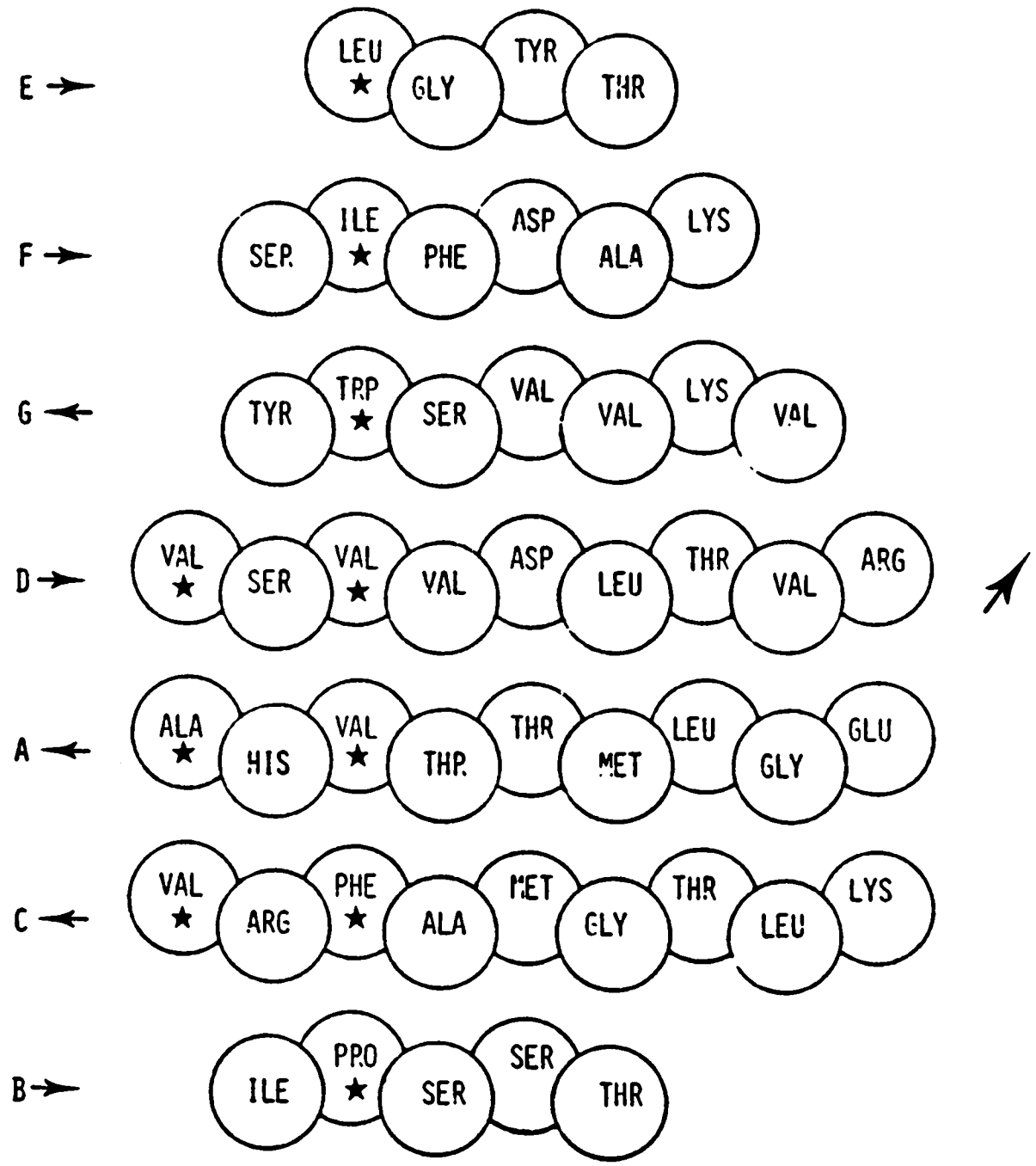
CARBOXYPEPTIDASE A



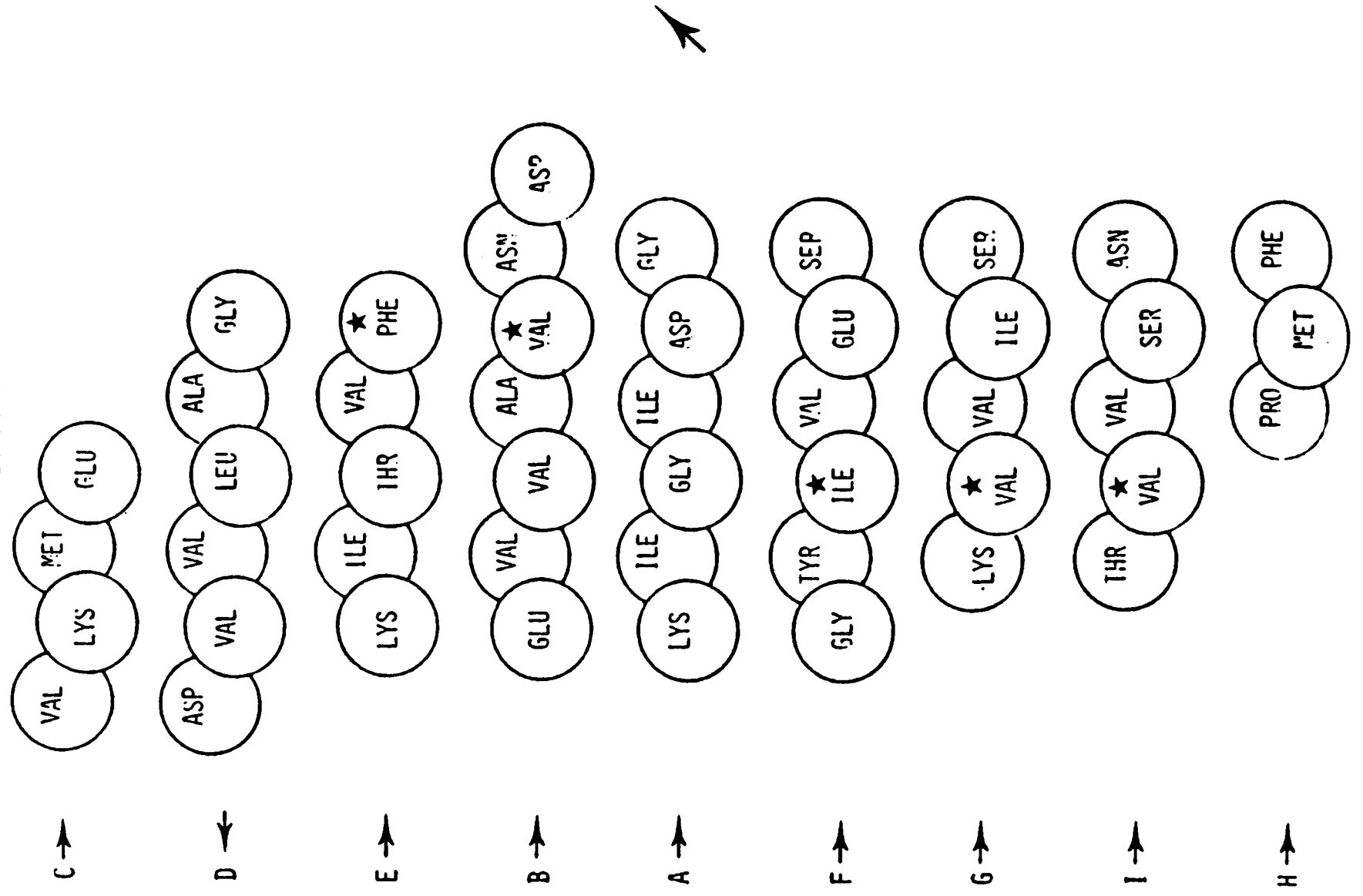
GLYCERALDEHYDE 3-PHOSPHATE DEHYDROGENASE
CATALYTIC DOMAIN



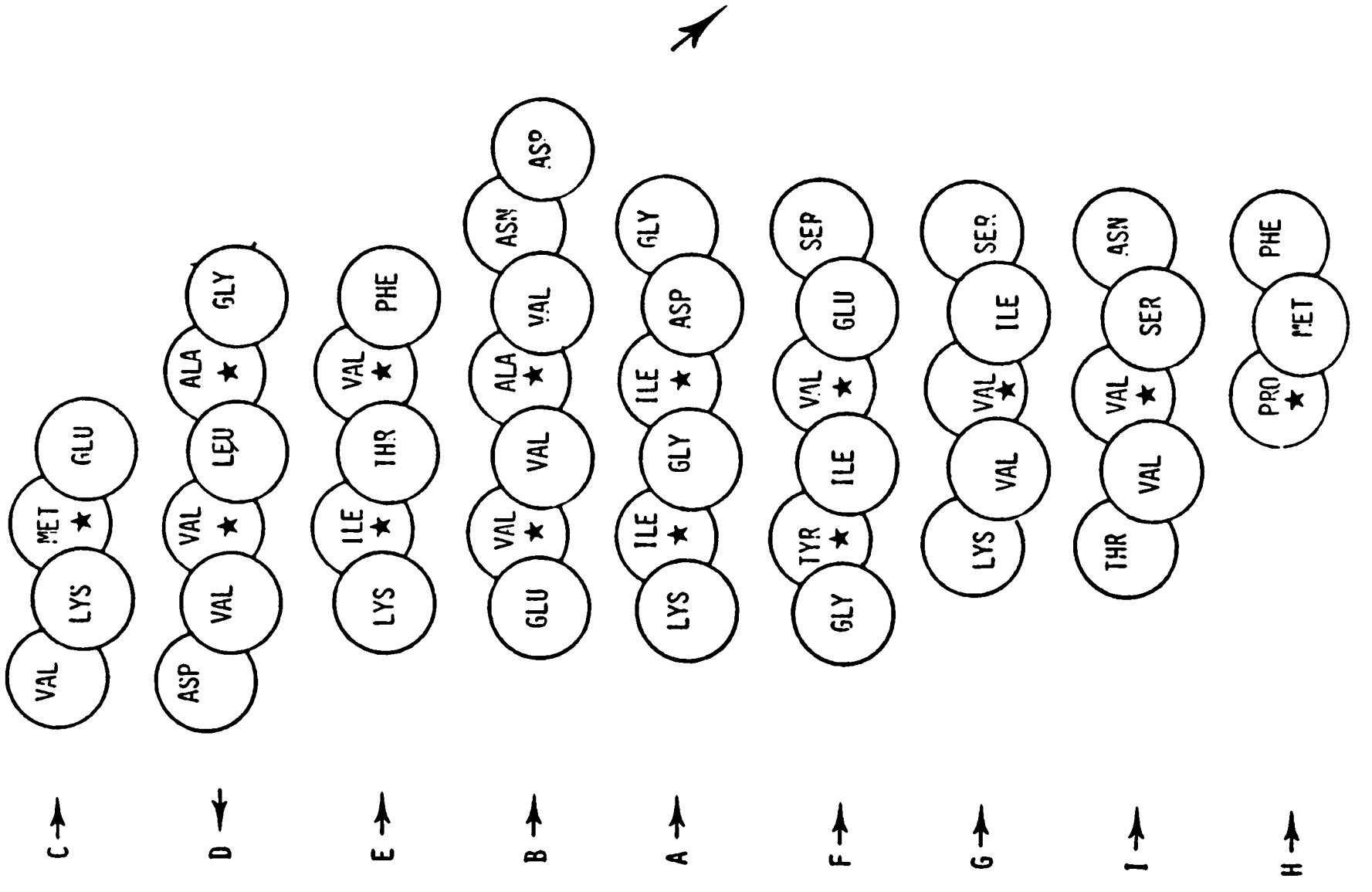
GLYCERALDEHYDE 3-PHOSPHATE DEHYDROGENASE
CATALYTIC DOMAIN



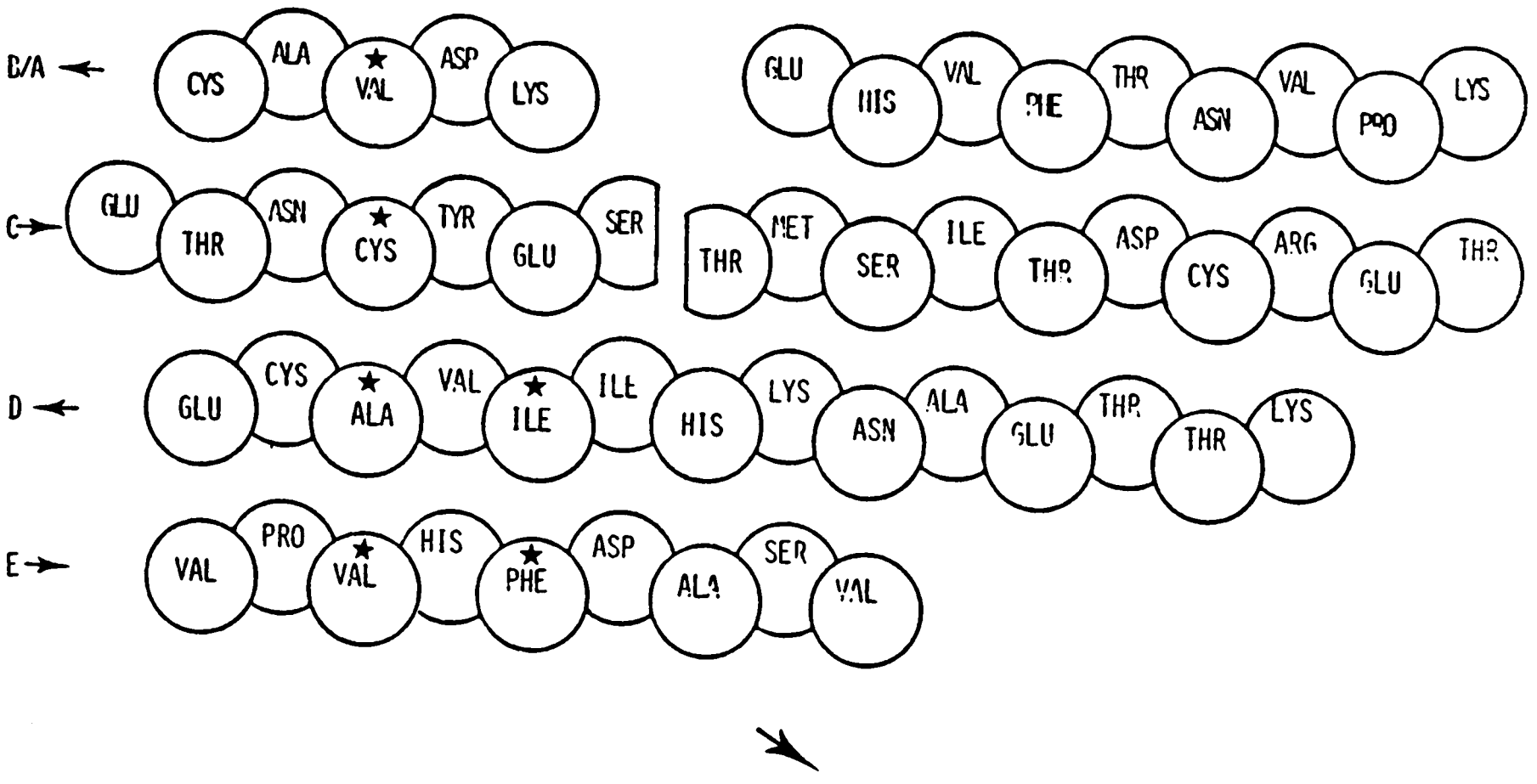
GLYCERALDEHYDE 3-PHOSPHATE DEHYDROGENASE
(NAD BINDING DOMAIN)



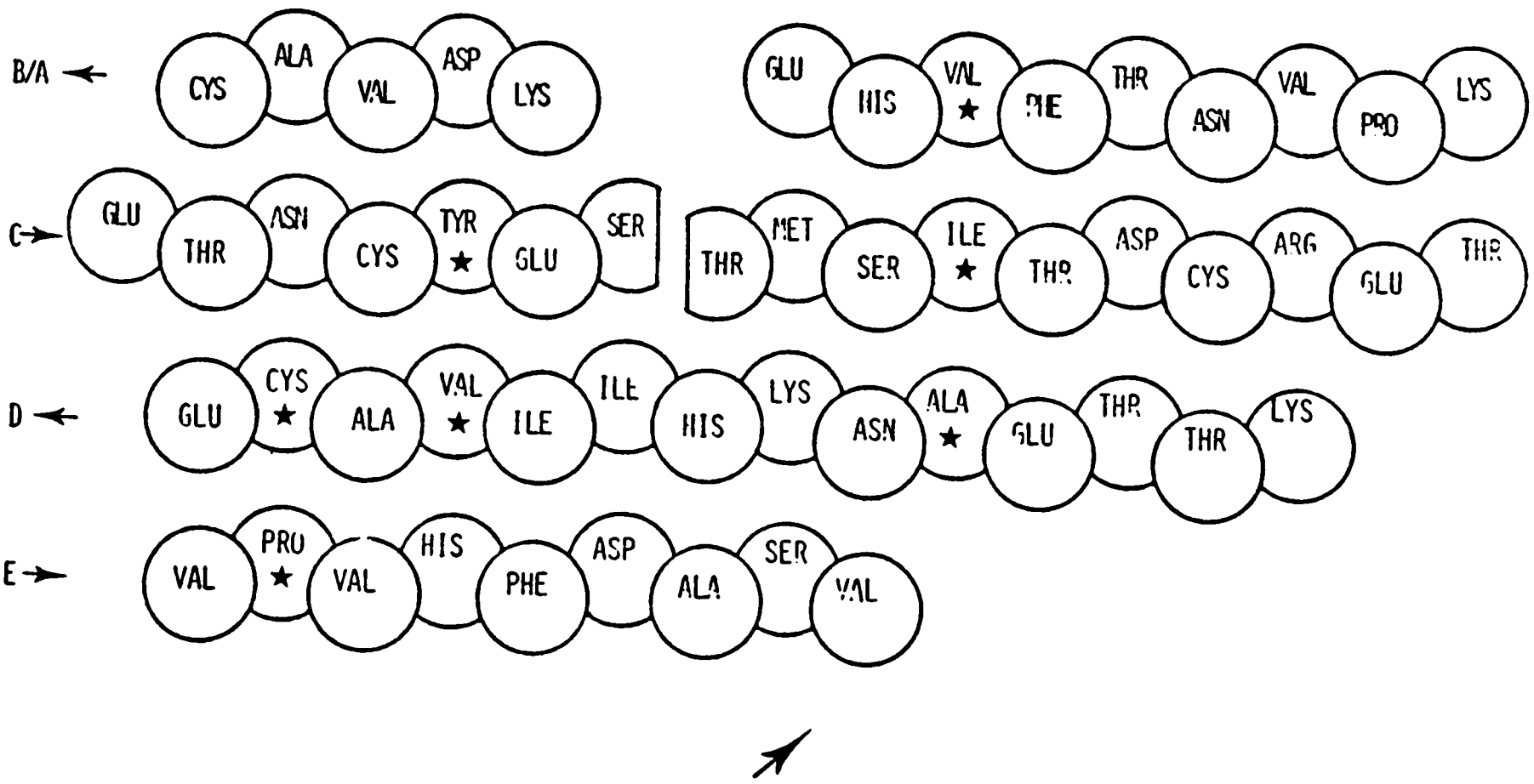
GLYCERALDEHYDE 3-PHOSPHATE DEHYDROGENASE
(NAD BINDING DOMAIN)



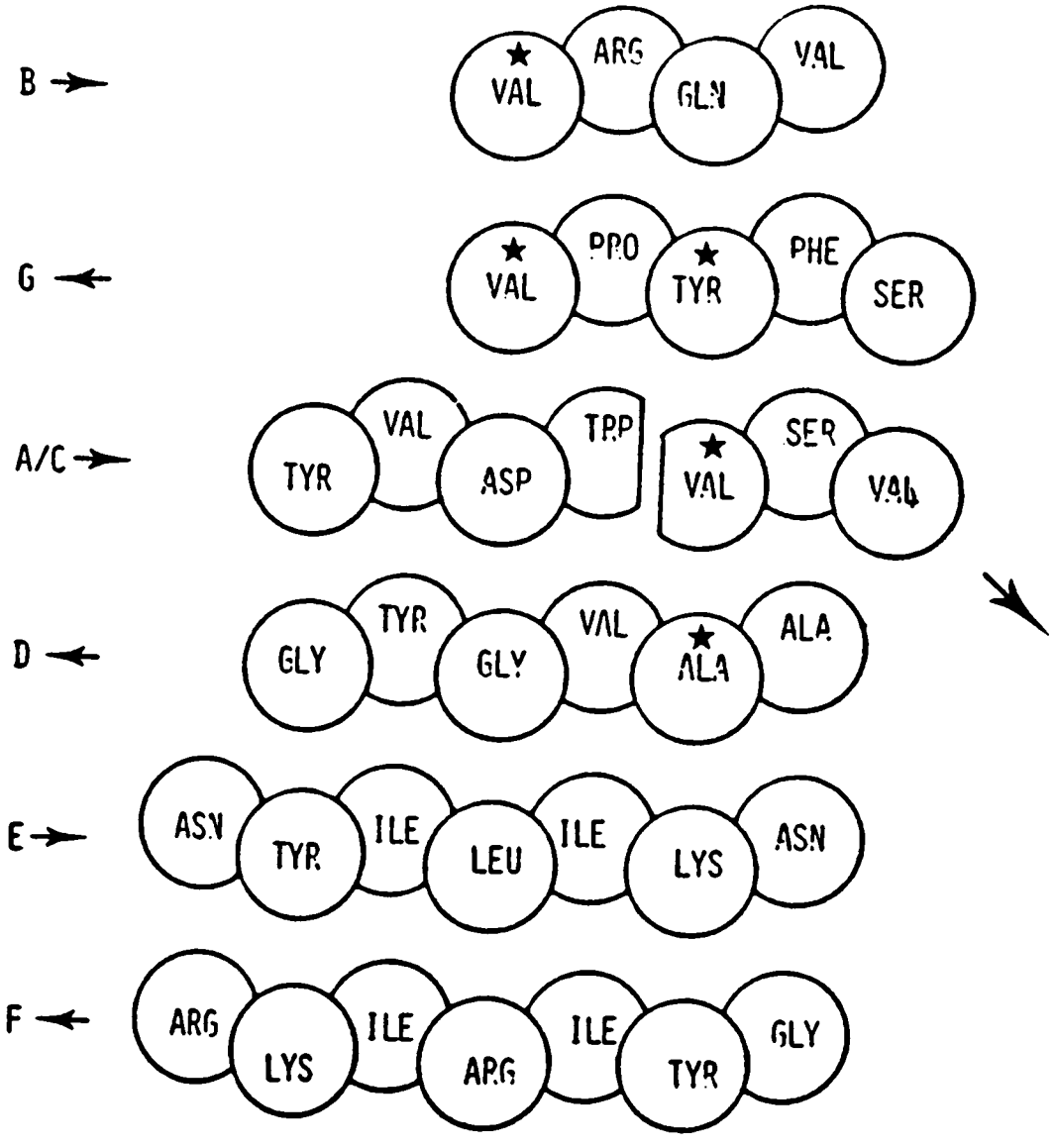
RIBONUCLEASE S



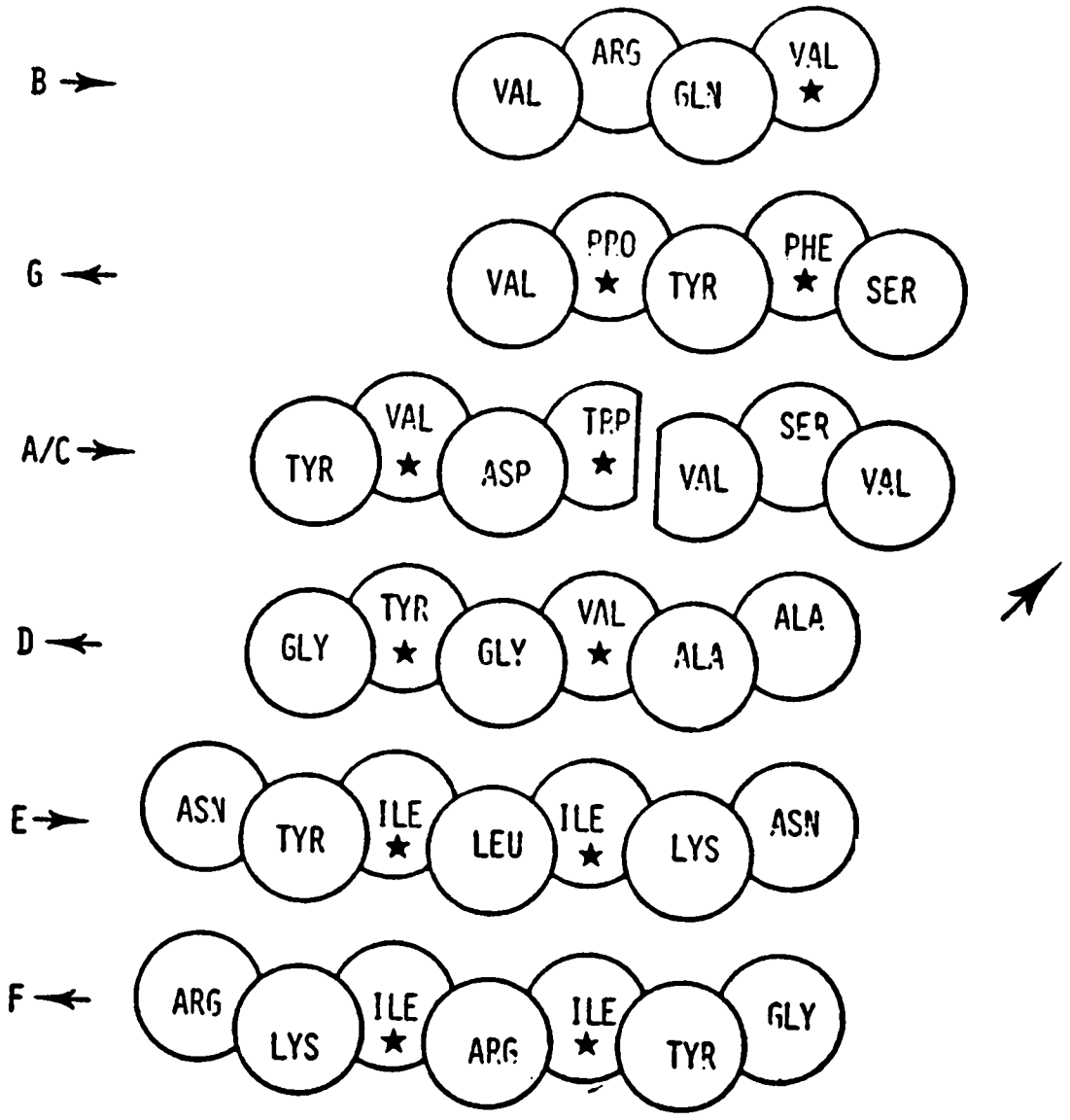
RIBONUCLEASE S



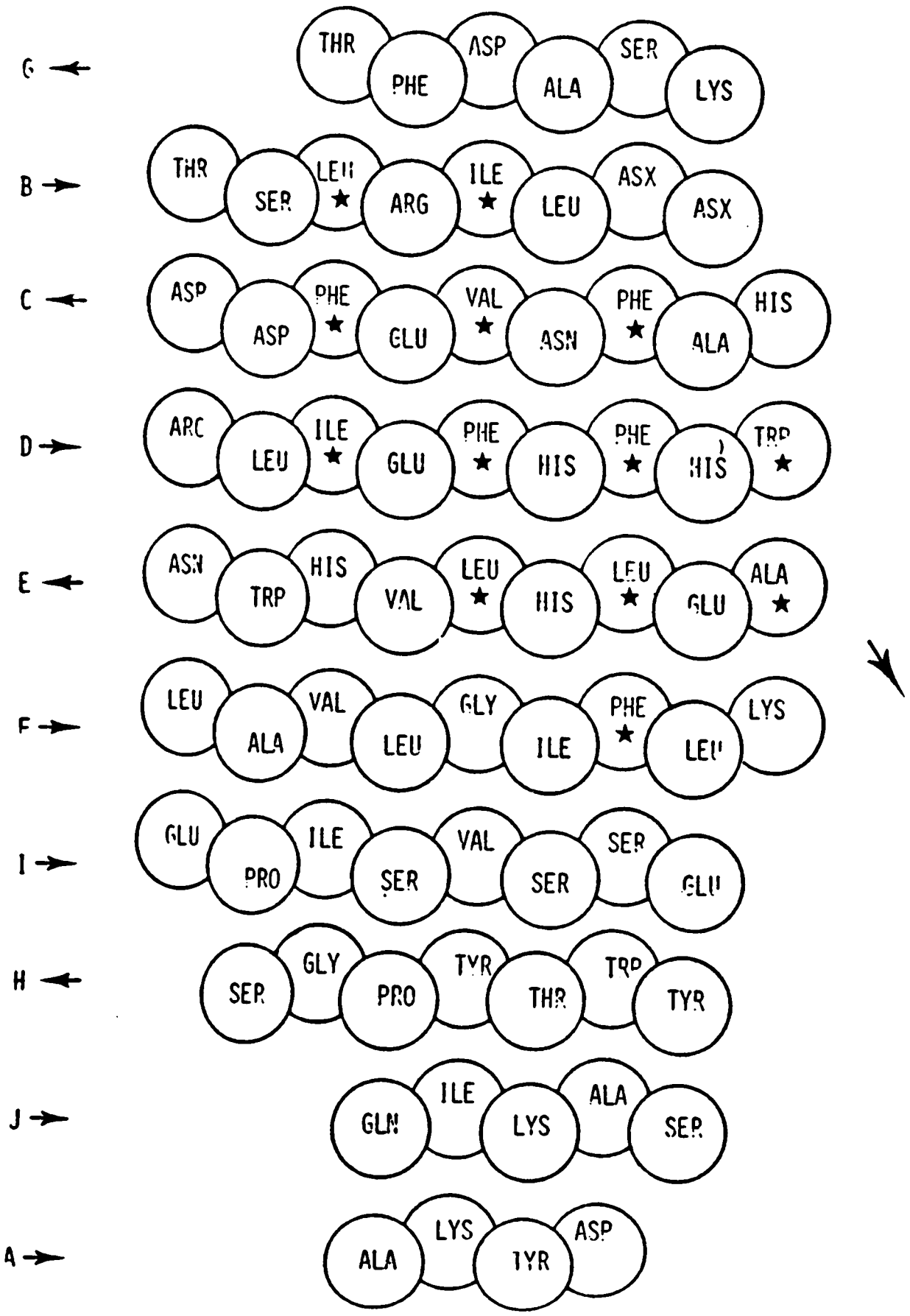
PAPAIN



PAPAIN



CARBONIC ANHYDRASE



CARBONIC ANHYDRASE

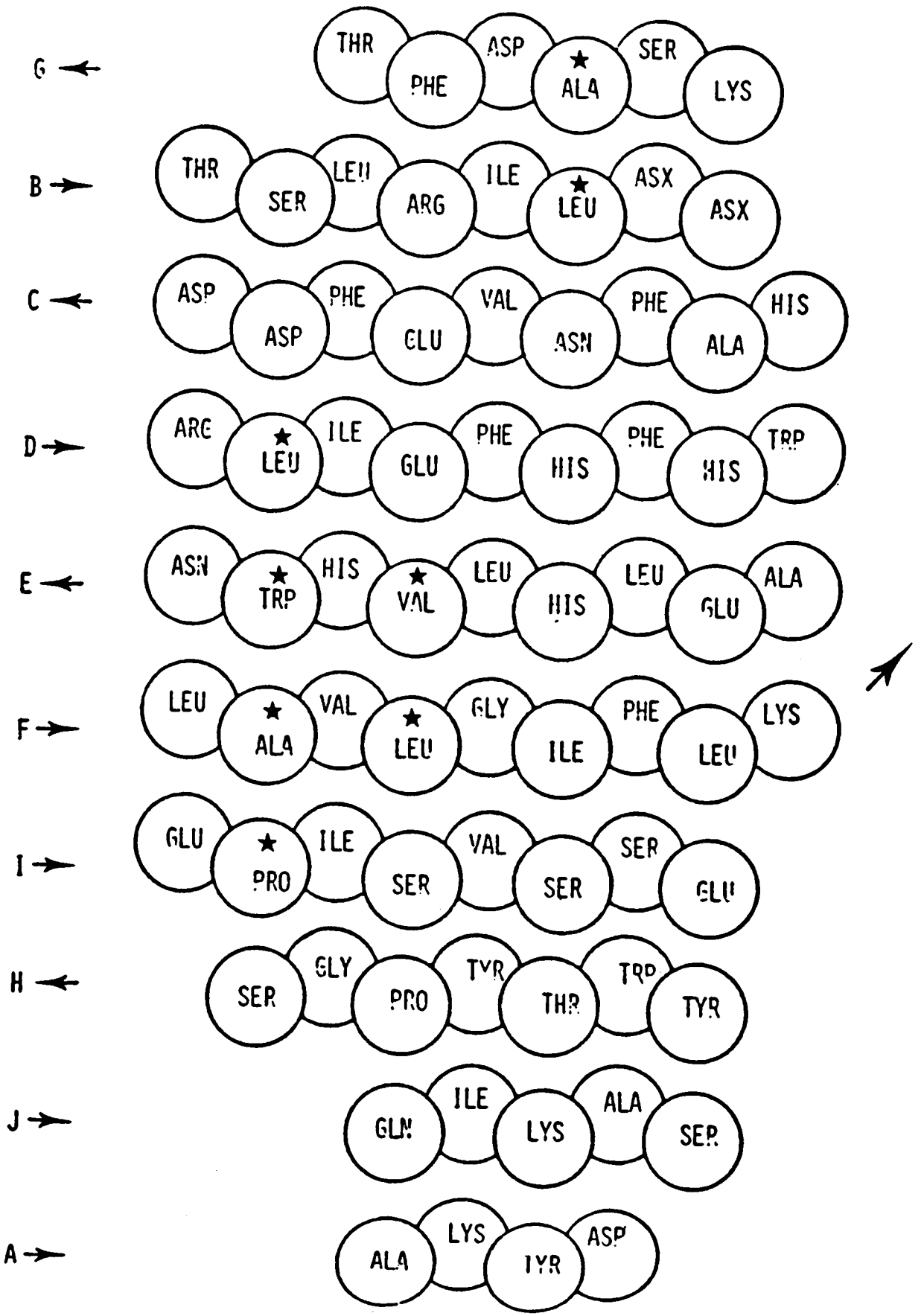


TABLE 5.6

MIXED SHEET CALCULATIONS

Protein	Number of Strands	Number of Possible Structures[a]	Number of Allowed Structures
TRDX	5	6.3×10^{7}	63
B5C	5	6.3×10^{7}	510
CPA	8	8.7×10^{13}	2388
GPDCAT	7	6.8×10^{11}	649
GPDNAD	9	1.2×10^{16}	1513
CARB	10	1.9×10^{18}	b

[a] Computed as $8^{n}.2^{n-1}.n!$ where n is the number of strands, 8^{n} is the number of parallel and antiparallel alignments, 2^{n-1} is the number of sides for connections and n! is the number of strand orders.

[b] This computation is beyond the capabilities of the facilities at Oxford and will be evaluated at a later date using the Rutherford Computing Laboratory.

of the native-like structure over the other alternatives will become visible as the quality and extent of the protein structure representation improves.

4 Postlude

The calculations on the structure of βα proteins presented in this Chapter are far from complete. They are certainly the most ambitious calculations attempted in this thesis. As only a small fraction of the residues in the proteins studied are in β-structure, the impact of this work is muted. However, the number of possible structures for β/α proteins is significantly larger than for α/α or β/β proteins. For the β-sheet of carbonic anhydrase, 1.9×10^{18} structures would have to be sampled. Possible positions for the helical residues will certainly augment this figure. Given the preliminary nature of the approach presented in this Chapter, the results can only be considered encouraging. First, the number of β-sheets consistent with the topological and structural rules developed are small. Moreover, it is easy to envisage a β-sheet as the structural core of a β/α protein onto which helices may be placed.

CHAPTER 6

CONCLUSION

In this thesis, I have examined the myriad of structural themes in proteins in an effort to derive unifying principles of organisation. The two major conclusions of this work are:

(1) The structure of many proteins can be approximated by the packing of units of secondary structure;

and

(2) A systematic application of rules which restrict the path of the polypeptide chain to all combinations of secondary structure assemblies severely limits the number of reasonable tertiary folds*.

As this search has global character, the existence of a good approximation to the crystallographically determined native structure is guaranteed; only its uniqueness is in question. Presumably, increasingly detailed and refined representations of these approximate folds will enhance the superiority of the native-like structure over all other candidates.

Implicit in this approach to the prediction of protein structure from amino acid sequence information is a hierarchic condensation model of folding:

(1) Primary sequence determines secondary structure;

(2) Secondary structure determines the tertiary fold; and

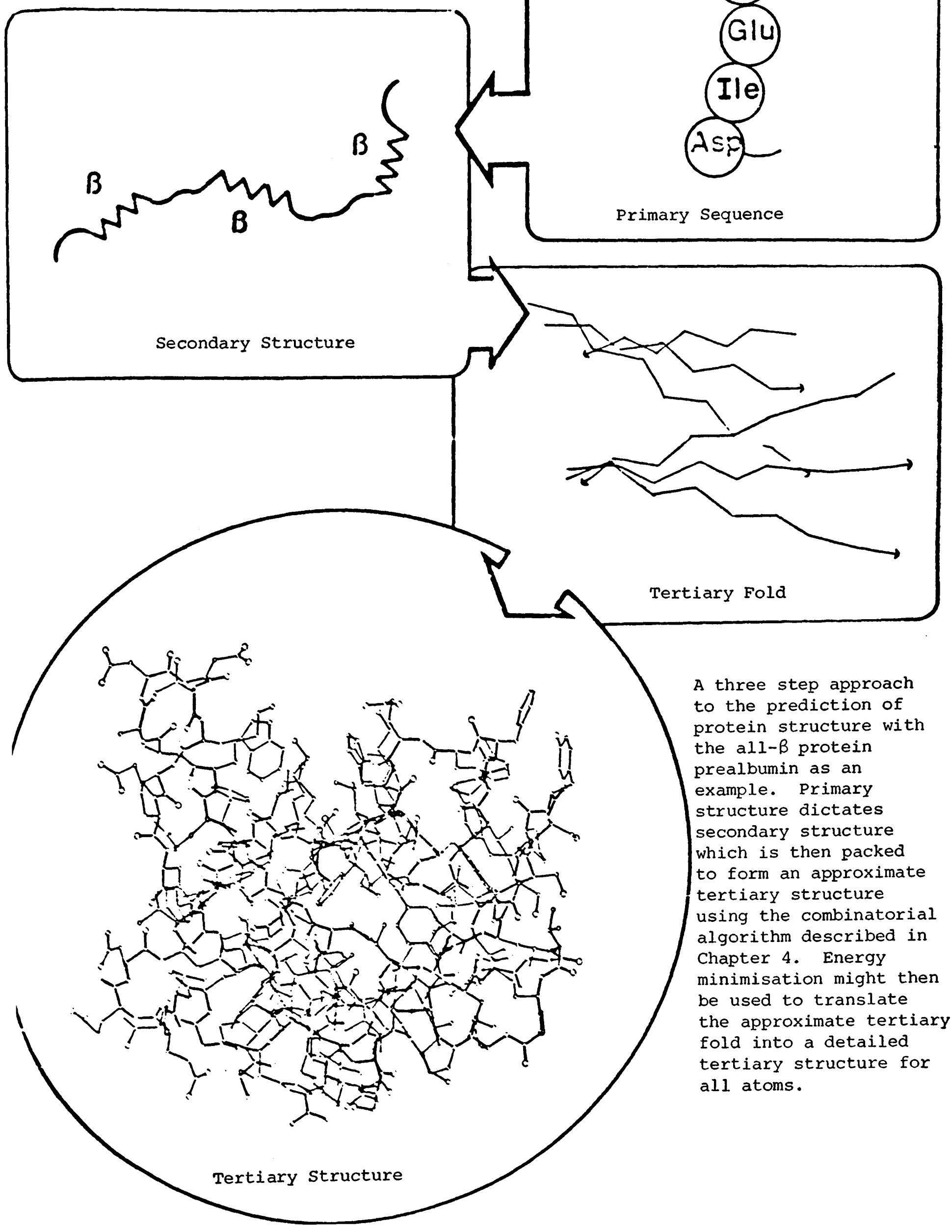
(3) The tertiary fold determines the tertiary structure.

This thesis has concentrated on the second link in this chain.

* A tertiary fold is the approximate path of the polypeptide backbone through the protein structure which adequately defines general features of the structure, e.g. internal and external residues.

FIGURE 6.1

Hierarchic Condensation Model for the Prediction of Protein Structure.



A three step approach to the prediction of protein structure with the all-β protein prealbumin as an example. Primary structure dictates secondary structure which is then packed to form an approximate tertiary structure using the combinatorial algorithm described in Chapter 4. Energy minimisation might then be used to translate the approximate tertiary fold into a detailed tertiary structure for all atoms.

The hierarchic condensation model is a simplification of the protein folding problem. Obviously, there is some feedback between levels of organisation. Perhaps the rules governing the packing of secondary structure will provide some insight into the secondary structure prediction problem. Certainly some sequences facilitate and others inhibit the packing of secondary structure.

The third part of this model, the complete assignment of atomic positions will no doubt come from an energy minimisation procedure. Improvements in potential functions will increase the radius of convergence of these procedures. As this expands, it may be hoped that a combinatorial algorithm for predicting the tertiary fold from only considerations of amino acid sequence will be there to meet it.

REFERENCES

Anderson, C.M., Zucker, F.H. and Steitz, T.A. (1979) Science 204, 375-380.

Anfinson, C.B. (1972) Biochem. J. 128, 737.

Anfinsen, C.B., Haber, E., Sela, M. and White, F.H. (1961) Proc. Natl. Acad. Sci. U.S.A. 47, 1309-1314.

Anfinsen, C.B. and Scheraga, H.A. (1975) Adv. Prot. Chem. 29, 205-300.

Antonini, E. and Brunori, M. (1971) in Hemoglobin and Myoglobin in their Reactions with Ligands, pp. 40-54.

Arnone, A., Bier, C.J., Cotton, F.A., Day, V.W., Hazen, E.E. Jr., Richardson, D.C., Richardson, J.S. and Yonath, A. (1971) J. Biol. Chem. 246, 2302-2316.

Artymiuk, P. (1979) D. Phil. Thesis, University of Oxford.

Baldwin, J. and Chothia, C. (1979) J. Mol. Biol. 129, 175-220.

Baldwin, R.L. (1980) in Protein Folding (ed. Jaenicke, R.) Elsevier/North Holland Biomedical Press, Amsterdam.

Banks, R.D., Blake, C.C.F., Evans, P.R., Haser, R., Rice, D.W., Hardy, G.W., Merrett, M. and Phillips, A.W. (1979) Nature 279, 773-777.

Banner, D.W., Bloomer, A.C., Petsko, G.A., Phillips, D.C., Pogson, C.I., Wilson, I.A., Corran, P.N., Furth, A.J., Milman, J.D., Offord, R.E., Priddle, J.D. and Waley, S.G. (1975) Nature 255, 609-614.

Beale, D. and Feinstein, A. (1976) Q. Rev. Biophys. 9, 135-180.

Beghin, F. and Dirkx, J. (1975) Arch. Int. Physiol. Biochim. 83, 167-168.

Bennett, W. and Steitz, T.A. (1978) Proc. Natl. Acad. Sci. U.S.A. 75, 4848-4852.

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T. and Tasumi, M. (1977) J. Mol. Biol. 112, 535-542.

Blake, C.C.F., Geisow, M.J., Oatley, S.J., Rerat, B. and Rerat, C. (1978) J. Molec. Biol. 121, 339-356.

Bloomer, A.C., Champness, J.N., Bricogne, G., Staden, R. and Klug, A. (1978) Nature 276, 362-368.

Bragg, W.L. (1921) Proc. Cambridge Phil. Soc. 17, 43.

Brandts, J.F., Halvorson, H.R. and Brennan, M. (1975) Biochemistry 14, 4953-4963.

Brill, A.S. and Sandberg, H.E. (1967) Proc. Natl. Acad. Sci. U.S.A. 57, 136.

Burnett, R.M., Darling, G.D., Kendall, D.S., LeQuesne, M.E., Mayhew, S.G., Smith, W.W. and Ludwig, M.L. (1974) J. Biol. Chem. 249, 4383-4392.

Chothia, C. (1973) J. Mol. Biol. 75, 295-302.

Chothia, C. (1974) Nature, 248, 338-339.

Chothia, C. (1976) J. Mol. Biol. 105, 1-14.

Chothia, C., Levitt, M. and Richardson, D. (1977) Proc. Natl. Acad. Sci. U.S.A. 74, 4130-4134.

Chou, P.Y. and Fasman, G.D. (1974) Biochemistry 13, 222-244.

Cohen, F.E., Richmond. T.J. and Richards, F.M. (1979) J. Mol. Biol. 132, 275-288.

Cohen, F.E. and Sternberg, M.J.E. (1980a) J. Mol. Biol. 137, 9-22.

Cohen, F.E. and Sternberg, M.J.E. (1980b) J. Mol. Biol. 138, 321-333.

Cohen, F.E., Sternberg, M.J.E. and Taylor, W. (1980a) in Protein Folding (ed. Jaenicke, R.) Elsevier/North Holland Biomedical Press, Amsterdam. 131-148.

Cohen, F.E., Sternberg, M.J.E. and Taylor, W.R. (1980b) Nature in press.

Colman, D.M., Jansinius, J.N. and Matthews, B.W. (1972) J. Mol. Biol. 70, 701-724.

Creighton, T.E. (1978) Prog. Biophys. Molec. Biol. 33, 231-297.

Creighton, T.E. (1979) J. Mol. Biol. 129, 235-264.

Crick, F.H.C. (1953) Acta Cryst. 6, 689-697.

Crippen, G.M. (1974) J. Theor. Biol. 45, 327.

Crippen, G.M. (1975) J. Theor. Biol. 51, 495.

Crippen, G.M. (1977) J. Comp. Phys. 24, 1-12.

Crippen, G.M. (1978) J. Mol. Biol. 126, 315-332.

Crippen, G.M. (1979) Int. J. Pep. Prot. Res. 13, 320-326.

Crute, M.B. (1959) Acta Cryst. 12, 24.

Cummings, A.L. and Eyring, E.M. (1975) Biopolymers 14, 2107-2114.

Daniels, A. (1977) D. Phil. Thesis, University of Oxford.

Dayhoff, M.O. (1976) Atlas of Protein Sequence and Structure, Vol.5, supplement 2. pp. 168-169. National Biomedical Research Foundation, Washington D.C.

Diamond, R. (1966) Acta. Cryst. 21, 253-266.

Dickerson, R.E. (1979) Chronological Summary of Protein Structure Determinations, unpublished.

Dickerson, R.E., Takano, T., Eisenberg, D., Kallai, O.B., Samson, L., Cooper, A. and Margoliash, E. (1971) J. Biol. Chem. 246, 1511-1535.

Dirkx, J. (1972) Arch. Int. Physiol. Biochim. 80, 185-187.

Deisenhofer, J. and Steigemann, W. (1974) in Proteinase Inhibitors pp. 484-496, Springer-Verlag, Berlin.

Deisenhofer, J. and Steigemann, W. (1975) Acta. Cryst. Sect. B 31, 238.

Drenth, J., Jansonius, J.N., Koekoer, R. and Wolthers, B.G. (1971) Adv. Prot. Chem. 25, 79-115.

Dreyfus, M. and Pullman, A. (1970) Theoret. Chim. Acta 19, 20.

Erlund, H., Nordstrom, B., Zeppezauer, E., Sonderlund, G., Ohlsson, I., Boiwe, T., Soderberg, B.-O., Tapia, O., Branden, C.-I. and Akeson, A. (1976) J. Mol. Biol. 102, 27-59.

Epstein, H.F., Schecter, A.N., Chen, R.F. and Anfinsen, C.B. (1971) J. Mol. Biol. 60, 499-508.

Feldmann, R.J. (1976) AMSOM Atlas of Macromolecular Structure on Microfiche, Tracor Jitco, Rockville, Md.

Finkelstein, A.V. and Ptitsyn, O.B. (1971) J. Mol. Biol. 62, 613-624.

Fischer, E. (1894) Chem. Ber. 27, 2985-2993.

Fisher, H.F. (1964) Proc. Natl. Acad. Sci. 51, 1285-1291.

Flory, P.J. (1969) Statistical Mechanics of Chain Molecules John Wiley and Sons, New York.

Garnier, J., Osguthorpe, D.J. and Robson, B. (1978) J. Mol. Biol. 120, 97-120.

Gibson, K.D. and Scheraga, H.A. (1971) Proc. Natl. Acad. Sci. U.S.A. 58, 420.

Gō, N. and Taketomi, H. (1979) Int. J. Peptide Protein Res. 13, 235-252.

Goel, N.S. and Ycas, M. (1979) J. Theor. Biol. 77, 253-305.

Greer, J. (1971) J. Mol. Biol. 59, 107-126.

Hagler, A.T. and Honig, B. (1978) Proc. Nat. Acad. Sci. U.S.A. 75, 554-558.

Hagler, A.T., Huler, E. and Lifson, S. (1974) J. Am. Chem. Soc. 96, 5319-5327.

Hagler, A.T., Lifson, S. and Dauber, P. (1979) J. Am. Chem. Soc. 101, 5122-5130.

Hagler, A.T. and Lapiccirela, A. (1976) Biopolymers 15, 1167-1200.

Hagler, A.T. and Moult, J. (1978) Nature 272, 222-226.

Hamlin, J. and Zabin, I. (1972) Proc. Natl. Acad. Sci. U.S.A. 57, 484-495.

Hammes, G.G. and Roberts, P.B. (1969) J. Am. Chem. Soc. 91, 1812-1816.

Harrison, S.C. and Blout, E.R. (1965) J. Biol. Chem. 240, 299-303.

Hartsuck, J.A. and Lipscomb, W.N. (1971) in The Enzymes (Boyer, P.D. ed.) Vol. 3, 3rd edit., pp. 1-56, Academic Press, New York.

Havel, T.F., Crippen, G.M. and Kollman, P.A. (1979) Biopolymers 18, 73-81.

Hendrickson, W.A., Klippenstein, G.L. and Ward, K.B. (1975) Proc. Natl. Acad. Sci. U.S.A. 72, 2160-2164.

Hendrickson, W.A., Love, W.E. and Karle, J. (1973) J. Mol. Biol. 74, 331-361.

Hermans, J. and Ferro, D. (1971) Biopolymers 10, 1121-1138.

Hirschmann, R., Nutt, R.E., Veber, D.F., Vitali, R.A., Varga, S.L., Jacob, T.A., Holly, F.W. and Denkewalter, R.G. (1969) J. Am. Chem. Soc. 81, 501-506.

Holbrook, J.J., Liljas, A., Steindel, J. and Rossman, M.G. (1975) in The Enzymes (Boyer, P.D., ed.) Vol. 11, 3rd edit., pp. 191-292, Academic Press, New York.

Honzatko, R.B., Monaco, H.L. and Lipscomb, W.N. (1979) Proc. Natl. Acad. Sci. U.S.A. 76, 5105-5109.

Hopfinger, A.J. (1973) Conformational Properties of Macromolecules, Academic Press, New York.

Huber, R., Deisenhofer, J., Colman, P.M., Matsushima, M. and Palm, W. (1976) Nature 264, 415-420.

Imoto, T., Johnson, L.N., North A.C.T., Phillips, D.C. and Rupley, J.A. (1972) in The Enzymes (Boyer, P.D. ed.) Vol. 7, 3rd edit., pp. 665-868, Academic Press, New York.

Johnson, R.N., Bradbury, J.H. and Appelby, C.A. (1978) J. Biol. Chem. 253, 2148-2154.

Kabat, E.A. and Wu, T.T. (1973a) Proc. Natl. Acad. Sci. U.S.A. 70, 1473-1477.

Kabat, E.A. and Wu, T.T. (1973b) Biopolymers 12, 751-774.

Kabat, E.A. and Wu, T.T. (1974) Proc. Natl. Acad. Sci. U.S.A. 71, 4217-4220.

Karplus, M. and McCammon, J.A. (1979a) Nature 277, 578.

Karplus, M. and McCammon, J.A. (1979b) Proc. Natl. Acad. Sci. U.S.A. 76, 3585-3589.

Karplus, M. and Weaver, D.L. (1974) Nature 260, 404-406.

Kauzmann, W. (1959) Adv. Prot. Chem. 14, 1-63.

Kendrew, J.C., Dickerson, R.E., Strandberg, B.E., Hart, R.G., Davies, D.R., Phillips, D.C. and Shore, V.C. (1960) Nature 185, 422-427.

Kraut, J., Robertus, J.D., Birktoft, J.J., Alden, R.A., Wilcox, P.E. and Powers, J.C. (1971) Cold Spring Harbor Symp. Quant. Biol. 36, 117-123.

Kretsinger, R.H. and Nockolds, C.E. (1973) J. Biol. Chem. 248, 3313-3326.

Kuntz, I.D. (1972) J. Am. Chem. Soc. 94, 4009.

Kuntz, I.D., Crippen, G.M. and Kollman, P.A. (1979) Biopolymers 18, 939-957.

Kuntz, I.D., Crippen, G.M., Kollman, P.A. and Kimelman, D. (1976) J. Mol. Biol. 106, 983-994.

Kuntz, I.D. and Kauzmann, W. (1974) Adv. Prot. Chem. 28, 239-347.

Labhardt, A.M. and Baldwin, R.L. (1979) J. Mol. Biol. 135, 231-244.

Lee, B. and Richards, F.M. (1971) J. Mol. Biol. 55, 379-400.

Lesk, A. and Chothia, C. (1980) J. Mol. Biol. 136, 225-270.

Levitt, M. (1976) J. Mol. Biol. 104, 59.

Levitt, M. (1977) Third Taniguchi International Symposium Taniguchi Foundation, Kyoto, Japan.

Levitt, M. and Chothia, C. (1976) Nature 261, 552-557.

Levitt, M. and Warshel, A. (1975) Nature 253, 694.

Lewis, P.N., Momany, F.A. and Scheraga, H.A. (1971) Proc. Natl. Acad. Sci. U.S.A. 68, 2293-2297.

Lifson, S. and Sander, C. (1980) in Protein Folding (ed. Jaenicke, R.) Elsevier/North Holland Biomedical Press, Amsterdam.

Liljas, A., Kannan, K.K., Bergsten, P.-C., Waara, I., Fridborg, K., Standberg, B., Carlbom, U., Jarup, L., Lovgren, S. and Petef, M. (1972) Nature New Biol. 235, 131-137.

Lim, V.I. (1974a) J. Mol. Biol. 88, 857-872.

Lim, V.I. (1974b) J. Mol. Biol. 88, 873-894.

Lim, V.I. (1978) FEBS Lett. 89, 10-14.

Matheson, R.R. and Scheraga, H.A. (1978) Macromolecules 11, 819-829.

Mathews, F.S., Argos, P. and Levine, M. (1971) Cold Spring Harbor Symp. Quant. Biol. 36, 387-395.

Mathews, F.S., Bethge, P.H. and Czerwinski, E.W. (1979) J. Biol. Chem. 254, 1699-1706.

Matthews, B.W. and Remington, S.J. (1974) Proc. Natl. Acad. Sci. U.S.A. 71, 4178-4182.

McCammon, J.A., Gelin, B.R. and Karplus, M. (1977) Nature 267, 585-590.

McDonald, R.C., Steitz, T.A. and Engleman, D.M. (1979) Biochemistry 18, 338-342.

McLauglin, A.D. (1971) J. Mol. Biol. 61, 409-424.

Metropolis, N., Rosenbluth, A.W., Rosenbluth, M.N., Teller, A. and Teller, E. (1953) J. Chem. Phys. 21, 1087-1092.

Momany, F.A., McGuire, R.F., Burgess, A.W. and Scheraga, H.A. (1975) J. Phys. Chem. 79, 2361-2381.

Moras, D., Olsen, K.W., Sabesan, M.N., Buehner, M., Ford, G.C. and Rossmann, M.G. (1975) J. Biol. Chem. 250, 9137-9162.

Nagano, K. (1973) J. Mol. Biol. 75, 401-420.

Nagano, K. (1974) J. Mol. Biol. 84, 337-372.

Nagano, K. and Hasegawa, K. (1975) J. Mol. Biol. 94, 257-281.

Nagano, K. (1977) J. Mol. Biol. 109, 251-274.

Nemethy, G. and Scheraga, H.A. (1977) Quart. Rev. Biophys. 10, 239-352.

Nishikawa, K., Ooi, T., Isogai, Y. and Saito, N. (1972) J. Phys. Soc. Japan 32, 1331-1337.

Nozaki, Y. and Tanford, C. (1971) J. Biol. Chem. 246, 2211.

Orr, H.T., Lancet, D., Robb, R.J., Lopez de Castro, J.A. and Strominger, J.L. (1979) Nature 282, 266-270.

Pace, C.N. (1975) Crit. Rev. Biochem. 3, 1-43.

Palau, J. and Puigdomenech, P. (1974) J. Mol. Biol. 88, 457-469.

Pauling, L. (1960) The Nature of the Chemical Bond. Cornell University Press, Ithaca.

Pauling, L., Corey, R.B. and Branson, H.R. (1951) Proc. Natl. Acad. Sci. U.S.A. 37, 205-211.

Pereti, P.F. (1974) Boll. Chim. Farm. 113, 187-218.

Perutz, M.F., Kendrew, J.C. and Watson, H.C. (1965) J. Mol. Biol. 13, 669-678.

Perutz, M.F., Muirhead, H., Cox, J.M., Gorman, L.C.G., Mathews, F.S., McGanoy, E.L. and Webb, L.E. (1968) Nature 219, 29-32.

Phillips, D.C. (1967) Lysozyme and the Development of Protein Crystal Chemistry.

Pickover, C.A., McKay, D.B., Engleman, D.M. and Steitz, T.A. (1979) J. Biol. Chem. 254, 11323-11329.

Privalov, P.L. and Khechinashuili, N.N. (1974) J. Mol. Biol. 86, 665.

Ptitsyn, O.B. and Finkelstein, A.V. (1970a) Biophysics 15, 785-796.

Ptitsyn, O.B. and Finkelstein, A.V. (1970b) Dokl. Biochem. 195, 322-325.

Ptitsyn, O.B., Finkelstein, A.V. and Falk, P. (1979) FEBS Lett. 101, 1-5.

Ptitsyn, O.B. and Rashin, A.A. (1975) Biophys. Chem. 3, 1-20.

Pullman, B. and Pullman, A. (1974) Adv. Prot. Chem. 28, 347-526.

Quiocho, F.A. and Lipscomb, W.N. (1971) Adv. Prot. Chem. 25, 1.

Ramachandran, G.N. and Sasisekharan, V. (1968) Adv. Prot. Chem. 23, 283-437.

Richards, F.M. (1974) J. Mol. Biol. 82, 1-14.

Richards, F.M. (1977) Ann. Rev. Biophys. Bioeng. 6, 151-176.

Richards, F.M. (1979) Carlsberg Res. Commun. 44, 47-63.

Richards, F.M. and Richmond, T.J. (1977) Molecular Interactions and Activity in Proteins. Ciba Foundation, Amsterdam, 23-45.

Richards, F.M. and Wyckoff, H.W. (1971) in The Enzymes (Boyer, P.D., ed.) Vol. 4, 3rd edit. pp. 647-806, Academic Press, New York.

Richards, F.M., Richmond, T.J., Sternberg, M.J.E. and Cohen, F.E. (1980) in Protein Folding (ed. Jaenicke, R.) Elsevier/North Holland Biomedical Press, Amsterdam.

Richardson, J.S. (1977) Nature 268, 495-500.

Richardson, J.S. (1979) The Protein Structure Colouring Book Little River Institute, Bahama, North Carolina.

Richardson, J.S. (1980) Adv. Prot. Chem. 39, in press.

Richardson, J.S., Getzoff, E.D. and Richardson, D.C. (1978) Proc. Natl. Acad. Sci. U.S.A. 75, 2574-2578.

Richardson, J.S., Richardson, D.C., Thomas, K.A., Śilverton, E. and Davies, D.R. (1976) J. Mol. Biol. 102, 221-235.

Richmond, T.J. (1980) unpublished.

Richmond, T.J. and Richards, F.M. (1978) J. Mol. Biol. 119, 537-555.

Roberts, J.D. (1959) Nuclear Magnetic Resonance, McGraw-Hill, New York.

Robson, B. and Osguthorpe, D.J. (1979) J. Mol. Biol. 132, 19-51.

Robson, B. and Pain, R.H. (1974) Biochem. J. 141, 883-897.

Robson, B. and Suzuki, E. (1977) J. Mol. Biol. 107, 357-367.

Romero-Herrera, A.C., Lehmann, H., Jaysey, K.A. and Friday, A.E. (1978) Phil. Trans. Roy. Soc. B. 283, 61-163.

Rose, G. (1978) Nature 272, 586-590.

Rose, G.D. (1979) J. Mol. Biol. 134, 447-470.

Rossmann, M.G., Liljas, A., Brandon, C.I. and Banaszak, L.J. (1975) in The Enzymes 11, 61-102.

Saul, F.A., Amzel, L.M. and Poljak, R.J. (1978) J. Biol. Chem. 253, 585-597.

Sawyer, L., Shotton, D.M. and Watson, H.C. (1973) Biochem. Biophys. Res. Comm. 53, 944-951.

Scheraga, H.A. (1968) Adv. Phys. Org. Chem. 6, 103-184.

Scheraga, H.A., Scott, R.A., Vanderkooi, G., Leach, S.J., Gibson, K.D., Ooi, T. and Nemethy, G. (1967) Conformations in Biopolymers, Vol.I. Academic Press, New York.

Schiffer, M. and Edmundson, A.B. (1967) Biophys. J. 7, 121-135.

Schiffer, M. and Edmundson, A.B. (1968) Biophys. J. 8, 29-39.

Schmid, F.X. and Baldwin, R.L. (1979) J. Mol. Biol. 135, 199-215.

Schulz, G.E. (1980) J. Mol. Biol. 138, 335-347.

Schulz, G.E., Elzinga, M., Marx, F. and Schirmer, R.H. (1974a) Nature 250, 120-123.

Schulz, G.E., Barry, C.D., Friedman, I., 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. (1974b) Nature 250, 140-142.

Schulz, G.E. and Schimer, R.H. (1979) Principles of Protein Structure. Springer-Verlag, New York.

Schwarz, G. (1965) J. Mol. Biol. 11, 69-77.

Shen, L.L. and Hermans, J. (1972) Biochemistry 11, 1836-1841.

Shrake, A. and Rupley, J.A. (1973) J. Mol. Biol. 79, 351-372.

Steiner, D.F. and Clark, J.L. (1968) Proc. Natl. Acad. Sci. U.S.A. 60, 622-629.

Steitz, T.A., Fletterick, R.J., Anderson, W.F. and Anderson, C.M. (1976) J. Mol. Biol. 104, 197-222.

Sternberg, M.J.E., Grace, D.E.P. and Phillips, D.C. (1979) J. Mol. Biol. 130, 231-253.

Sternberg, M.J.E. and Thornton, J.M. (1976) J. Mol. Biol. 105, 367-382.

Sternberg, M.J.E. and Thornton, J.M. (1977a) J. Mol. Biol. 110, 269-283.

Sternberg, M.J.E. and Thornton, J.M. (1977) J. Mol. Biol. 110, 285-296.

Sternberg, M.J.E. and Thornton, J.M. (1978) Nature 271, 15-20.

Sternberg, M.J.E., Cohen, F.E., Richmond, T.J. and Richards, F.M. (1979) unpublished data.

Stubbs, G., Warren, S. and Holmes, K. (1977) Nature 267, 216-221.

Sturtevant, J.M. (1977) Proc. Natl. Acad. Sci. U.S.A. 74, 2236-2240.

Takano, T. (1977) J. Mol. Biol. 110, 569-584.

Tanford, C. (1973) The Hydrophobic Effect, Wiley, New York.

Venkatachalam, C.M. (1968) Biopolymers 6, 1425-1436.

Vuk-Pavlovic, S. and Siderer, Y. (1977) Biochem. and Biophys. Res. Comm. 79, 885-889.

Wagner, G. and Wutrich, K. (1979) J. Mol. Biol. 134, 75-94.

Warshel, A. and Levitt, M. (1976) J. Mol. Biol. 106, 421-437.

Watson, H.D. (1969) Progress in Stereochemistry 4, 299-333.

Weatherford, D.W. and Salemme, F.R. (1979) Proc. Natl. Acad. Sci. U.S.A. 76, 19-23.

Wetlaufer, D. (1973) Proc. Natl. Acad. Sci. U.S.A. 70, 697-701.

Wu, T.T. and Kabat, E.A. (1971) Proc. Natl. Acad. Sci. U.S.A. 68, 1501-1506.

Wu, T.T. and Kabat, E.A. (1973) J. Mol. Biol. 75, 13-31.

Zav'yalov, V.P. (1977) Biophys. Biophys. Acta 490, 506-514.

Zimm, B.H. and Bragg, J.K. (1959) J. Chem. Phys. 31, 526.

APPENDIX 1

R.M.S. DEVIATION PROGRAMS

APPENDIX 2

BUILD, FOLD, AND REALSPACE

AND HEME CONSTRAINT PROGRAMS

APPENDIX 3

KINETICS AND PGK HINGE PROGRAMS

APPENDIX 4

BETA-SANDWICH PROGRAMS

APPENDIX 5

PURE-PARALLEL AND MIXED SHEET PROGRAMS