J. Mol. Biol. **(1977) 114, 181-293**

Automatic Identification of Secondary Structure in Globular Proteins

MICHAEL LEVITT

Medical Research Council Laboratory of Molecular Biology
Hills Road, Cambridge, England

AND JONATHAN GREER

Department of Biological Science
Columbia University, New York, N.Y. 10027, U.S.A.

(Received 23 December 1976, and in revised form 17 March 1977)

A computer program is used to analyse automatically and objectively the atomic co-ordinates of a large number of globular proteins in order to identify the regions of a-helix, β-sheet and reverse-turn secondary structure. Several different criteria for the assignment of secondary structure are tested for accuracy, reproducibility and efficiency. The most successful criterion, which is based on patterns of peptide hydrogen bonds, inter-0 distances and inter-C^{α} torsion angles, is used to find the secondary structure of all the proteins studied. The accuracy of the derived assignments is assessed by comparing them with the secondary structure reported in the literature for each protein. The reliability of the methods is assessed by comparing the secondary structures derived from. the independently determined sets of co-ordinates available for some proteins.

We provide the first objective and consistent compilation of a-helix, β-sheet and reverse-turn secondary structure in almost all globular proteins of known tertiary structure. These data will be invaluable for analysing the relative tendencies of different amino acids to occur in different types of secondary structure, for analysing the regularity of the secondary structure itself, and for analysing how the pieces of secondary structure fit together to form the globular tertiary structure of each protein.

1. Introduction

The secondary structure of a segment of a polypeptide chain is defined formally as "the spatial arrangement of its main-chain atoms without regard to the conformation of its side-chains or to its relationship with other segments" (Kendrew *et al.*, **1970).** Three types of secondary structure are commonly found in globular protein conformations: (1) the a-helix, in which the polypeptide chain is wound into a tightly packed rod-like structure ; (2) the β-strand, in which the polypeptide chain is almost fully extended and interacts with other β-strands through hydrogen bonds to form a twisted sheet-like structure; and (3) the reverse turn, in which the polypeptide chain reverses direction, bending back on itself. The α-helix and p-strand are repeating structures in that the conformation of all the residues in the region of secondary structure are the same. Reverse turns do not have this repeating structure and only consist of a small number of adjacent residues.

Historically, both the α-helix and β-strand were first predicted by model-building studies (Pauling *et al.*, 1951; Pauling & Corey, 1951). Both structures were then found to occur in such repeating polypeptides as α-keratin (Perutz, **1951)** and silk (Marsh *et al.*, **1955).** The α-helix was first proved to exist in a globular protein in **1960,** when Kendrew and co-workers solved the atomic structure of myoglobin by X-ray crystallography (Kendrew *et al.*, **1960).** The β-strand was found five years later in the X-ray structure of lysozyme (Blake *et al.*, **1965).** The reverse turn was not initially recognized as a well-defined type of local structure. After Venkatachalam (1968) showed theoretically that turns could have precise conformations, Crawford *et al.* (1973) and Lewis *et al.* **(1971)** found many examples of reverse turns in known protein structures.

Regions of secondary structure are important in defining protein conformations. **(1)** Most of the inner core of the native structure of a globular protein is built from residues in a-helices and p-strands. The reverse turns and irregular regions of the chain lie on the outside of the molecule and link the α-helices and β-strands (Levitt & Chothia, **1976). (2)** The type of secondary structure correlates with the amino acid sequence of the chain in that region, and the secondary structure can sometimes be predicted fairly well from the sequence (Schulz *et* al., 1974*b*; Matthews, **1975** ; Argos *et al.*, **1976).**

In the past, the secondary structure of most globular proteins has been defined by inspecting the atomic model of the particular protein. While this approach is generally very satisfactory, it is somewhat subjective, as different groups of workers can use different criteria to identify secondary structure. The criteria used most often by crystallographers involves inspection of both the local conformation of a particular residue in relation to that of neighbouring residues and the pattern of hydrogen bonds involving these residues. The greatest difficulty occurs for short irregular pieces of secondary structure and at the ends of longer pieces of secondary structure, where the conformation of neighbouring residues is not much help and it is sometimes difficult to be sure that there is a possible hydrogen bond. For lysozyme, one of the first proteins solved by X-ray crystallography, the (ϕ,ψ) backbone torsion angles were used to locate the end-points of α-helices more precisely (Phillips, **1967).**

More recently, (ϕ,ψ) angles calculated from atomic co-ordinates have been used to assign up to five local conformations for each residue (Burgess *et al.*, **1974** ; Robson & Pain, **1974;** Tanaka & Scheraga, 1976*a*,*b*; Maxfield & Scheraga, **1976).** Although this assignment of secondary structure is in full accord with the formal definition, it does not always correspond to the assignment made by crystallographers, who pay more attention to hydrogen bonds than (ϕ,ψ) angles: small errors in (ϕ,ψ) values can lead to incorrect assignments. The distance between pairs of C^{α} atoms has been used to identify reverse-turns (Lewis *et al.*, **1971;** Kuntz, **1972;** Crawford *et al.*, **1973).** Other workers have suggested that inter-C^{α} torsion angles could identify secondary structure more accurately than (ϕ,ψ) angles (Srinivasan *et al.*, **1975).**

At present there is no precise rule for characterizing secondary structure in proteins. In this paper we make such precise rules, use automatic objective methods to derive secondary structure assignments for many proteins, and then test the results against the reported assignments made by X-ray crystallographers. We hope these rules, which are shown to work very well, will serve as general criteria to be used to identify secondary structure in the future. As a result of the present lack of *a* precise rule and an objective method, there is no consistent assignment of secondary structure for

many globular proteins. The results given in this paper remedy this and provide the first objective and precise compilation of secondary structure for most proteins of known conformation.

2. Methods

The **atomic** co-ordinates of 62 protein conformations as determined by X-ray crystallography form the raw material for this study. These co-ordinates were obtained through Dr J. Richard Feldmann, who has provided a most valuable service by collecting, organizing and distributing the available sets of protein co-ordinates. We are grateful for being able to use these co-ordinates and thank all the many protein X-ray crystallographers concerned. Only the C^{α} co-ordinates were used here, as they are more readily available, less subject to experimental error (an incorrectly positioned C^{α} will disturb the positioning of both the backbone and side-chain), and may soon be determined automatically from the electron density map without any manual model-building or density fitting (Greer, 1974,1975,1976*a*,*b*). These co-ordinates were first checked for errors by calculating the distance between the C^{α} atoms of adjacent residues along the amino acid sequence. For most of the proteins the mean value of this separation was between 3·759 Å and 3·873 Å, and the standard deviations were between 0·0002 Å and 0·01 Å, indicating reasonable geometry and chain continuity. For a few proteins, the mean and standard deviation of the distance between adjacent α-carbons were abnormal, as the co-ordinates had not been specified in an orthogonal Cartesian co-ordinates system in angstrom units. In these cases the method of least-squares was used to find a 3×3 co-ordinate transformation matrix that minimized the deviation of the separation of adjacent C^{α} atoms from the expected value of 3·8 Å.

Sometimes crystallographers have not been able to provide co-ordinates for every residue in the amino acid sequence. Often this is due to disordered regions **of the chain** or to proteolytic cleavage of the protein before crystallization. We have chosen to number the residues according to their positions in the actual complete amino acid sequence, irrespective of whether co-ordinates exist for all the residues. In the few cases where no amino acid sequence is available, we use the crystallographic numbering of residues. **Both trypsin** and chymotrypsin are numbered according to their zymogen sequences, but **the numbering of** elastase is not changed to make it align with these proteins. **In the proteins** analysed here, there are no available co-ordinates for the following residues : **1 to 9** in prealbumin ; **1** to 7 in trypsin; 9 to **15, 148** to **149** in chymotrypsin; **71** to 77, 144 **to 153 in** chymotrypsinogen ; **1** to 2 in ferricytochrome *b5* ; **1** in ribonuclease **A;** and 82 to 89 in carboxypeptidase B. In 2 cases the relation between the X-ray determined and the chemically determined amino acid sequence was complicated by several insertions and deletions, and we used the X-ray numbering scheme (immunoglobulin fragment **Fab and apo-lactate** dehydrogenase) .

(a) *Secondary structure from a torsion angles*

It is well-known that the (ϕ,ψ) torsion angles (dihedral angles) of amino acid residues in globular proteins tend to cluster on a plot of ϕ against ψ: a-helical conformations occur near $(-60°,$ -40") and β-sheet conformations occur near $(-90°, 120°)$. In this paper we do not use these torsion angles, as the atomic co-ordinates of the peptide group needed to calculate (ϕ,ψ) values are not always known accurately. Instead, a single torsion angle is used: the torsion angle α_i defined by the path of the 4 consecutive C^{α} atoms belonging to residues i -2, $i - 1$, i, and $i + 1$. This definition of α_i, which is different from that used previously (Levitt, 1976; where α_i depended on residues $i - 1$, i, $i + 1$, $i + 2$), is more convenient as the nature of residue i has most influence on α_i. The inter-C^{α} angle has been used to define the conformation of polypeptide random coils (Flory, 1969), to define the conformation of dipeptides (Nishikawa *et* al., 1974), to simulate the folding of a simplified protein chain (Levitt & Warshel, **1975;** Levitt, **1976;** Warshel & Levitt, **1976)** and to describe the local conformation of protein residues (Srinivasan *et al.,* **1975).**

First, the continuous range of α_i values is converted into a code defining the local conformation of each residue. This code was found by examining the histogram of α_i values of all the proteins considered here (see Fig. 1). Residues with α_i values between

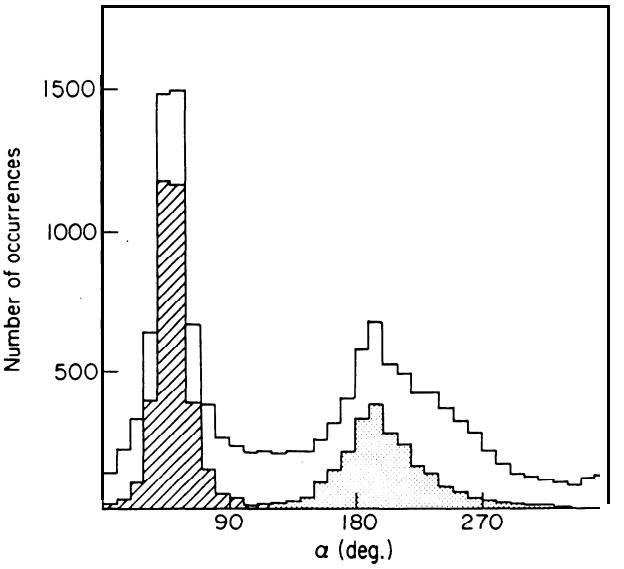


FIG. 1. Histograms of the number of occurrences of different a-angle values for all the proteins analysed here. The clear histogram shows the distribution of all the a values. The cross-hatched histogram shows the a-angles of only those residues found to be in a-helices. The shaded histogram shows the a-angles of only those residues found to be in β-sheet (Tables 3 to 10). The a-angles for residues in a-helices and β-sheets were only calculated when all 4 C^{α} atoms defining the particular a-angle were in the helix or sheet. It is clear that the 2 peaks of the histogram of all a-angles correspond to these 2 types of secondary structure. Although the residues in a-helix and β-sheet do have different local conformations as measured by the a-angle, the spread in values is quite broad. a-helical conformations occur at $0° < a < 130°$ and β-sheets occur at $130° < a < 330°$.

$10°$ and $120°$ are given code 1 for the right-handed a-helix; residues with α_i values between $120°$ and $270°$ are given code 2 for the β-sheet; and residues with α_i values between $-90°$ and $0°$ are given code 3 for fhe left-handed a-helix. Because fhere is some overlap of fhe ranges of α_i values that correspond to a-helices and β-strands, the right-handed helix code is also assigned to a residue with α_i between $120°$ and $140°$, provided that the neighbouring residues ($i - 2$, $i - 1$, $i + 1$ and $i + 2$) have been assigned as right-handed helix.

Next, 4 or more consecutive residues with the same local structure code are defined as a-helices or /I-strands, depending on the particular code value. As the torsion angle α_i is defined by the positions of the 4 adjacent C^{α} atoms of residues $i - 2$, $i - 1$, i, $i + 1$, residues i and $i - 1$ can both be equally well-associated with α_i. We resolve fhis difficulty by first assigning the state of residue i by the value of α_i, and then extending any regions of secondary structure by 1 residue towards the N-terminal (provided fhis does nof cause overlap with another region). As an α_i value cannot be calculated for the first 2 residues and fhe last residue, regions of secondary strucfure are also extended to include any adjacent residues for which no α_i value was calculated. The shortest a-helix or /I-strand defined by this method contains 4 residues.

Finally, residues that have not been assigned to either a-helix or β-sheet are tested for reverse-turn secondary strucfure. Residue i is assigned to be in a left or right-handed turn, depending on whether $-90° < \alpha_i < 0°$ or $0° < \alpha_i < 90°$. After assigning furns in this way, any residue i fhaf has nof been assigned to a-helix, β-sheet or turn secondary structure is then assigned to be in a turn if residue $i + 1$ is in a turn. Tripeptide energy calculations (Levitt, 1976) support fhe idea fhaf fhe nature of residue i has more influence on fhe preferred value of α_i than residue $i - 1$ has.

The C^{α} co-ordinates can also be used to define the bond angle τ_i between residues $i - 1$, i, $i + 1$. Because the τ_i values correlate partially with the α_i values, occur in a limited range ($80°$ to $150°$), and fluctuate subsfanfially within a region of secondary strucfure (Levitt, 1976), we do not feel that they would improve the accuracy of the a-angle method.

(b) *Secondary structure from hydrogen bonds*

Using hydrogen bonds to identify secondary structure may seem in conflict with the basic definition of secondary structure given above, for these bonds depend on the conformations of a pair of residues separated along the sequence and not on the local conformation of a particular residue. Nevertheless, the importance of the peptide group hydrogen bonds in defining the geometry of the a-helix and β-sheet has been appreciated for some time. β-Sheets, in particular, are defined more by these hydrogen bonds than by the actual local conformation of each residue in the sheet.

In this paper we have chosen to restrict our attention to the C^α co-ordinates, and consequently have no peptide group with which to form hydrogen bonds. Fortunately, it is possible to generate a good approximation to the peptide group from the C^α co-ordinates (see Levitt, 1976). In the simplified peptide group so generated there is an effective peptide nitrogen N'_i midway between adjacent a-carbons of residue $i-1$ and i, and an effective peptide oxygen O'_i, 1 A from N'_i and perpendicular to the plane through the C^α atoms of residues $i-1$, i and $i+1$. The vector equation giving $\mathbf{r}(N'_i)$ and $\mathbf{r}(O'_i)$ are

$$\mathbf{r}(N'_i) = \{\mathbf{r}(C^\alpha_{i-1}) + \mathbf{r}(C^\alpha_i)\}/2,$$

$$\mathbf{r}(O'_i) = \mathbf{r}(N'_i) + \mathbf{U}_i/|\mathbf{U}_i|,$$

where

$$U_i = [\mathbf{r}(C^\alpha_{i+1}) - \mathbf{r}(C^\alpha_i)] \times [\mathbf{r}(C^\alpha_{i-1}) - \mathbf{r}(C^\alpha_i)].$$

Note that the ith peptide group defined in this way consists of the CO group of residue $i-1$ and the NH group of residue i, yet its conformation depends mainly on the conformational angles (ϕ,ψ) of residue i. This happens as the ϕ torsion angle of most residues in globular proteins is approx. $-90°$.

With this simplified peptide group it is now possible to find hydrogen bonds between pairs of peptides. A computer program was used to generate the co-ordinates of the simplified peptide group from the C^α co-ordinate data and then used to find acceptable hydrogen bonds between pairs of peptide groups i and j. The following criteria had to be satisfied before the hydrogen bond was acceptable. (1) $|i-j| > 2$, which eliminates spurious hydrogen bonds between close neighbours along the chain; (2) $|\mathbf{r}(N'_i) - \mathbf{r}(N'_j)|$, which is the distance separating N'_i and N'_j, must be less than 6 A; and (3) the 2 peptide dipoles, which lie along the vectors $[\mathbf{r}(O') - \mathbf{r}(N'_i)]$ and $[\mathbf{r}(O'_j) - \mathbf{r}(N'_j)]$ must point in the same diection to within 60° ($[\mathbf{r}(O'_i) - \mathbf{r}(N'_j)] \cdot [\mathbf{r}(O'_j) - \mathbf{r}(N'_i)] > 0{\cdot}5$). Energy calculations (Levitt, 1976) show that 2 peptide groups interact most strongly at $|\mathbf{r}(N'_i) - \mathbf{r}(N'_j)| = 4{\cdot}2$ A, with the 2 dipoles almost parallel. The energy increases by 3 kcal/mol from the minimum value of −5 kcal/mol when the peptide separation is increased from 4·2 Å to 6 Å or the dipoles are made to deviate from alignment by 60°. The above criteria should find all pairs of peptides whose interaction energy is lower than −2 kcal/mol.

First we tried to identify secondary structure by scanning a list of the acceptable hydrogen bonds output in a convenient form (see Fig. 2). Because this list shows the values of $j-i$ (referred to here as the hydrogen bond gap) for all hydrogen bonds between peptides i and j, the secondary structure is readily apparent. The a-helical regions have hydrogen bonds between residue i and $i+3$ (sometimes also $i+4$), so that consecutive entries with $j-i=3$ or −3 indicate helix. p-strands fall into 2 classes: for parallel p-strands, $j-i$ is constant for consecutive residues i; for antiparallel p-strands, $j-i$ increases or decreases by 2 for consecutive residues i in the strand (i + j is constant). Although it was instructive and relatively easy to assign regions of a-helix and β-strand secondary structure from such a list, there remained an element of subjectivity that could only be removed by using a completely computerized algorithm.

The algorithm used was as follows: find all pairs of residues i and $i+1$ that make an acceptable hydrogen bond to other residues j and j', respectively, and satisfy either $|j-i| = |j'-(i+1)|$ or $|j+i| = |j'+(i+1)|$. The pairs of residues satisfying the first criterion are assigned to a-helices if $|j-i| = 3$, and to parallel p-strands if $|j-i| \neq 3$. The pairs of residues satisfying the second criterion are assigned to antiparallel p-strands. With this definition, regions with as few as 2 residues can be assigned to a region of repeating secondary structure.

Arg 1								
Pro 2	53,O				cys 30	— 6,N		
Asp 3	3,N				Gln 31	— 8,O		
Phe 4	3,N				Thr 32	—10,N		
cys 5					Phe 33	—12,O		
Leu 6	— 3,0				Val 34	—14,N		
Glu 7	—3,O				Tyr 35	—16,O		
Pro 8					Gly 36	—20,N	— 18,N	
Pro 9					Gly 37	-22,0		
Tyr 10					Cys 38			
Thr 11					Arg 39			
Gly 12					Ala 40			
Pro 13					Lys 41			
cys 14					'Arg 42			
Lys 15	22,N				Asn 43			
Ala 16	20,O				Asn 44	—21,N		
Arg 17					Phe 45	-23,0		
Ile 18	18,O				Lys 46	-25,N		
Ile 19	16,N				Ser 47	-26,N		
Arg 20	14,0				Ala 48	3,N		
Tyr 21	12,N	25,0	26,0		Glu 49	3,N	4,N	
Phe 22	10,O	23,N			Asp 50	3,N		
Tyr 23	8,N	21,0			cys 51	—3,O	3,N	
Asn 24	6,O				Met 52	—3,O	—3,N	4,N
Ala 25					Arg 53	- 4 0	—3,O	
Lys 26					Thr 54	—3,O		
Ala 27					cys 55	— 53,N	— 3,0	
Gly 28					Gly 56	—4,O		
Leu 29					Gly 57			
					Ala 58			

FIG. 2. List of peptide hydrogen bonds produced by the program for the small protein trypsin inhibitor (bovine). For each peptide i that makes an acceptable hydrogen bond to another peptide j, the Figure shows the hydrogen bond gap (j — i) and also which part of the peptide bond of peptide j is involved in the bond (the CO group or the NH group). For example, the 2nd peptide group (immediately preceding the residue Pro2) hydrogen bonds to the CO group of the 55th peptide (immediately preceding residue Cys55). Most of the other hydrogen bonds in the Figure are more regular than this one. They indicate α-helix for peptides 3 to 7 and 48 to 56, and antiparallel β-sheet for peptides 15 to 16, 18 to 24, 30 to 37, and 44 to 46.

Adjacent pieces of secondary structure of the same type are then merged to form longer regions. **This merging does not work for short α-helices with only 5 residues, as the hydro**gen bond gap values will be 3,3, , —3, — 3 for residues 1 to 5, respectively (there are hydrogen bonds between 1 and 4 and 2 and 5). The pairs of residues **l-2 and 4-5** are not adjacent and will not be merged into a longer helix. **This problem is overcome by defining** as helical any residue preceded by a helical residue with j — i = 3 and followed by a helical residue with j — i = 3, giving an α-helix with 5 residues. Any stretches shorter than 5 residues that are assigned as helix are deleted.

The opposite problem occurs when two separate helices adjacent in sequence are incorrectly merged into a single longer helix. If the first helix ends at peptide i, which hydrogen bonds to peptide i — 3, and the second helix starts at peptide i + 1, which hydrogen bonds to peptide i + 4, these helices will be merged together. This is prevented by recognizing that the hydrogen bond gap is — 3 for peptide i and + 3 for peptide i + 1. Consequently, the first helix is made to end at peptide i, and the second start at peptide i + 1.

In this assignment of secondary structure, the presence of a hydrogen bond to peptide i is used to define the conformational state of residue i. Because the CO group of peptide i actually belongs to residue i — 1 and the NH group actually belongs to residue i, the hydrogen bond to peptide i could define the secondary structure of either residue i — 1 or i, depending on whether the -CO or -NH group is involved in the bond. Although a simplified peptide group is used here, we can distinguish between these 2 alternatives

using the sign of the vector dot product $[\mathbf{r}(N'_j) - \mathbf{r}(N'_i)] \cdot (\mathbf{r}(O'_i) - \mathbf{r}(N'_i))$, which is positive when the CO group of peptide i accepts a hydrogen bond and negative when the NH group donates one. Once this distinction is made, pieces of secondary structure defined above can then be modified to refer to the residues that are actually involved in the hydrogen bond. If the first hydrogen bond of the segment involves the CO group, the segment should be extended by 1 residue towards the N-terminus. If the last hydrogen bond of the segment involves the CO group, the segment should be shortened by cutting off the last residue. Modifying the original assignments in this way can lead to /?-strands with a single residue if the 2 peptide hydrogen bonds involve the NH and CO groups of the same residue.

A more straightforward way to allow for the fact that peptide i is between residues $i - 1$ and i is to simply extend the segments defined above by 1 residue towards the N-terminus, provided that this does not lead to an overlap with another segment. In this way the peptide units involved in the hydrogen bonds that define the secondary structure are all in between the residues defined to have the particular secondary structure. We use this last scheme when defining the secondary structure of known globular proteins from patterns of hydrogen bonds (the method is refered to as the H-bond method).

The hydrogen bond method should be able to distinguish 3_{10} helices from α-helices, as the hydrogen bond gap $|i - j| = 2$ for 3_{10} helices and $|i - j| = 3$ for a-helices. Unfortunately, because of the generous limits used to select acceptable hydrogen bonds, 3_{10} helices have hydrogen gap values of 2 and 3, whereas α-helices have gap values of 3 and 4. Thus, the present method identifies 3_{10} helix as if it was a-helix. Improvements that will enable these 2 types of helix to be distinguished are being developed.

(c) Secondary structure from inter-C^α—C^α distances

The distances between pairs of C^α co-ordinates can be used to represent the secondary and tertiary structure of proteins in schematic form. A table of these distances, known as a contact map, is one of the best ways to represent 3-dimensional structure in 2 dimensions as the α-helices, parallel β-sheets and antiparallel β-sheets all feature very characteristically (Phillips, 1970; Nishikawa *et al.*, 1972). We have developed a set of rules, based on the inter-C^α relations of the contact map, that can be used to detect regions of secondary structure automatically. The method used here has been modified and expanded from the previous report (Greer, 197 *6a).*

(i) *α-Helices*

The computer program was first used to find a-helices. Stretches of the chain are chosen with $|\mathbf{r}(C^\alpha_i) - \mathbf{r}(C^\alpha_{i+3})| < 6$ Å and $|\mathbf{r}(C^\alpha_i) - \mathbf{r}(C^\alpha_{i+4})| < 6{\cdot}5$ Å. When residues i to j satisfy both these criteria and residue j + 1 violates them, residues i to j + 4 are set to be helical. When one residue satisfies these criteria, the minimum helix is obtained consisting of 5 residues. In a standard a-helix the corresponding values of these distances are 5·1 Å and 6·1 Å, respectively. The above rules generally select the N-terminus of the helix properly. The C-terminus is often more irregular and more difficult to define precisely. Here, we find the C-terminus of the helix by applying the rules in the reverse direction (from the C to N-terminus).

The other problem is to distinguish between short helices containing 5 to 7 residues and reverse turns in the chain. This is done by comparing the C^α co-ordinates of helices of 5 to 7 residues with those of an ideal helix built with $\phi = -57{\cdot}37$ and $\psi = -47{\cdot}52$ (Diamond, 1966) using a least-squares procedure. If the mean-square deviation for a helix of 7 residues exceeds 1·0 Å^2, then the last residue of the helix is dropped and the fitting procedure repeated. When this test is applied to helices with 5 or 6 residues, the corresponding acceptable mean-square deviations are 0·4 Å^2 and 0·5 Å^2, respectively. A helix which is rather distorted, or a 3_{10} helix, which does not satisfy the criteria, will not be recognized. These rules therefore serve to select helices which fit the overall geometry of the a-helix more closely than the other methods used in this work.

Sometimes a stretch of residues which satisfies the above selection procedures includes more than one helix. These helices are separated by fitting the first 5 residues of the observed helical stretch to an ideal helix and then calculating the distance of each

succeeding C^α from the helix axis of the fitted ideal helix. **When 3 consecutive C^α atoms, j, $j+1$ and $j+2$, are all found** to be more than 3·31 A from **the projected helix axis, the helix is terminated** at residue j — 1. When the C^α of residues j — 2 also differs from the expected C^α to axis distance of 2·31 A by more than 1 Å, **then the previous helix is terminated** at $j - 3$. A new ideal helix is then fitted to the 5 **residues immediately following the helix just terminated.** Each of the helical sections produced **in this way is treated as an** independent helix.

(ii) *β-sheets*

Next the computer program is used to search for the parallel and antiparallel β-sheets. Firstly, those pairs of strands are found that satisfy **either of the following criteria:** $|\mathbf{r}(C^\alpha_{i+k}) - \mathbf{r}(C^\alpha_{j-k})| < 6{\cdot}5$ Å for $k = 0,1,2,\ldots n$ **and** $|i - j| > n + 1$, indicating a 2-stranded parallel β-sheet with n residues in each strand; $|\mathbf{r}(C^\alpha_{i+k}) - \mathbf{r}(C^\alpha_{j-k})| < 6{\cdot}5$ Å for $k = 0,1,2,3,\ldots n$ and $|i - j| > 3$, indicating a 2-stranded antiparallel β-sheet with n residues in each strand. This distance is 5·5 A and 5·6 Å **in regular parallel and antiparallel** β-sheets, **respectively.** Secondly, all pairs of strands that satisfy **1 of these criteria and** have at least 2 residues in each strand are subjected to a more stringent criterion:

$$|\mathbf{r}(C^\alpha_i) - (\mathbf{r}C^\alpha_{j+1})| > |\mathbf{r}(C^\alpha_i) - \mathbf{r}(C^\alpha_j)|$$

and

$$\mathbf{r}(C^\alpha_i) - \mathbf{r}(C^\alpha_{j-1})| > |\mathbf{r}(C^\alpha_i) - \mathbf{r}(C^\alpha_j)|.$$

This condition increases the likelihood that the C^α atoms of i and j on opposite strands are in proper register. Those p-sheets that satisfy both **the above criteria are then expanded** in both directions, adding residues that only satisfy one of the first criteria without opposite C^α **atoms necessarily being closest.** The **minimum number of residues in each** strand must be 3, giving a total of at least 6 residues for the pair of β-strands that form a possible sheet. Not all of these, however, are really β-sheets. Some just represent chains **running approximately** parallel (or antiparallel) **to each other.** Those **stretches most** likely to have the local residue conformation and hydrogen bonding pattern characteristic **of true p-sheets must** be distinguished. Unfortunately, this cannot be done by fitting a, **standard** β-sheet as was done with the a-helix; the wide **variety** of **twist angles** and local **irregularities found in** β-sheet make it impractical.

Instead, the true β-sheets are selected by considering more carefully the expected **geometry of 2 parallel** or antiparallel p-strands. **Two** vectors are calculated normal to the **2 planes defined by the C^α co-ordinates of residues** i, $i+1$, j and $j+1$ and $i+1$, $i+2$, $j+1$ and $j+2$ (i is directly opposite j in the sheet), and the torsion angle between these vectors is defined as ρ_i (see Fig. 3).

The torsion angle ρ_i is calculated for each residue of the sheet, starting from the beginning of the strand that comes first in the sequence (residue ***k).*** When ρ_i is outside the range −10° to 40° (where eclipsed vectors are at 0°), the program checks if there are 3 or

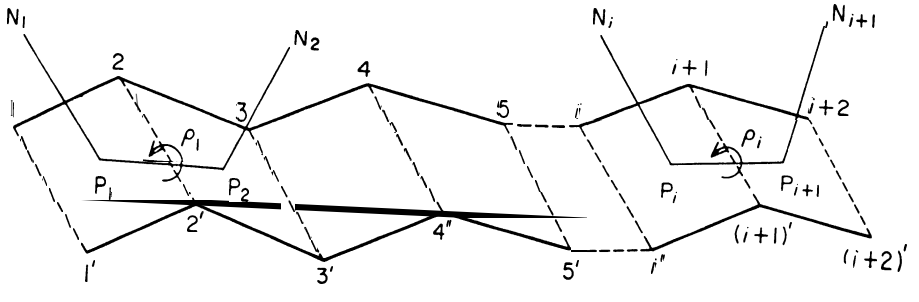


FIG. 3. Schematic representation of a β-sheet showing the vectors used to calculate the sheet twist angle ρ_i. First, 2 vectors are defined so as to be normal to the planes containing the C^α atoms of residues i,$i+1$ and j, and j, $j+1$ and i. The mean of these 2 normal vectors, called $\mathbf{n}_i$, is taken as the normal to the plane of the 4 C^α atoms i, $i+1$, j and $j+1$. The vector $\mathbf{n}_{i+1}$ is defined in the same way for the plane defined by the C^α atoms $i+1$, $i+2$, $j+1$ and $j+2$. A torsion angle ρ_i is then calculated from the 4 position vectors $\mathbf{p}_i + \mathbf{n}_i$, $\mathbf{p}_i$, $\mathbf{p}_{i+1}$, $\mathbf{p}_{i+1} + \mathbf{n}_{i+1}$, where $\mathbf{p}_i$ and $\mathbf{p}_{i+1}$ are the midpoints of the 4 atoms that define the 2 planes (i' in Fig. is j).

more residues from the beginning of the strand to the current residue i, and if there are it takes residues k to i as β-sheet. The remaining twist angles are then considered starting at residue $i + 1$. As ρ_i cannot be calculated for the last residue in the strand, sheets whose last acceptable ρ value is the jth are built to end at residue j + 1. For a β-sheet containing only 3 residues in each strand, there is only one dihedral angle ρ. In this case, the torsion angle is required to fall within a more restricted range of $0°$ to $30°$ before the β-sheet is accepted. These allowed ranges for ρ were found by examining the twist of pairs of strands reported in the literature as being β-sheet for the particular protein.

The program occasionally selected β-strands in which corresponding C^α atoms are out of step by 1 residue. This is detected by recognizing that the C^β atoms of correctly aligned residues in a β-sheet alternate together above and below the plane of the p-sheet. To test for this, the program calculates an approximate C_i^β position from the co-ordinates of C_{i-1}^α, C_i^α and C_{i+1}^α in much the same way as was done when defining the position of the O_i' atom in the simple peptide group (Lowinger & Greer, unpublished result). The β-sheet is then rejected if more than 1/3 of the C^β atoms do not alternate as expected.

3. Results

(a) *A reliable criterion for secondary structure*

Our first task is to determine how well each of the methods proposed in this paper represents the accepted definitions of secondary structure. To decide this, we have chosen six representative proteins, some small and some large, with a mixture of α-helices and β-sheet. The reported secondary structure designations, together with the derived assignments for each method, are shown in Table 1.

These six representative proteins have a total of 33 reported a-helices. Although the cc-angle method finds all these reported helices, it also finds an additional three helices (in myoglobin and carboxypeptidase). The positions of the helix termini derived by this method often differ by several residues from the reported positions, giving a total difference of 77 residues out of 380 in α-helix (20·3%). The H-bond method misses one reported helix (in cytochrome *b5*) and finds an extra helix (in carboxypeptidase), but the method does well overall with a total of 64 differences (16·8%). The C^α—C^α method misses four reported helices (in trypsin inhibitor, lysozyme and myoglobin), and finds two extra helices (in trypsin inhibitor and carboxypeptidase). Many helices found by this method have been incorrectly split into two pieces, and the overall score is worst with a total of 93 differences (24·5%).

It is clear that the definition of α-helix provided by peptide hydrogen bonds is closest to that used by protein crystallographers. The H-bond method is therefore used to assign residues to the a-helical secondary structure. Sometimes, an α-helix may be too irregular to have a recognizable pattern of hydrogen bonds. Such helices may still have enough residues in the helical local structure for the helix to be detected by the cc-angle method (for example, the second helix of cytochrome *b5*). To allow for this, residues that are not assigned to either the α-helix or β-sheet secondary structure using the preferred methods (see below) are assigned as a-helices using the cc-angle method, provided that such helices have at least five residues. Helices found in this way are distinguished from helices found by the more rigorous H-bond method in the final presentation of secondary structure that follows.

The situation for p-sheets is more complicated. A total of 25 β-strands is reported for the six proteins considered here (Table 1). The cc-angle method misses five reported β-strands and finds seven extra p-strands, giving a total difference of 101 residues out of 158 reported to be in β-sheets. The H-bond method finds all the reported

TABLE 1

Comparison of secondary structure assignment methods

Protein	α-Helices				β-Strands			
	Reported†	cc-Angles	H-bond	C^{α}—C^{α}	Reported?	a-Angles	H-bond	C^{α}—C^{α}
Trypsin inhibitor (bovine)	3-6	3-7	2-7		16-26	16-24	14-16,17-24	16-18,23-25
				24-28	27-36	30-36	29-37, 43-46	28-30,35-37
	47-58	48-56	47-55	47-56				
Cytochrome *b5* (bovine)	8-15	10-17	8-16	8-13	4-6	5-9	5-7	5-7
	33-38	33-39		35-39	21-25	20-24	21-25	21-25
	42-49	43-49	42-49	43-48	28-32	28-32	28-32	28-32
	55-62	55-61	55-61	54-61	50-54		52-54	51-53
	64-74	66-74	64-74	65-71		62-65		
	80-86	81-85	80-85	83-87	75-79		75-76, 77-79	77-79
Ribonuclease S (bovine)	3-13	4–10	3-12	3-12		13-18	13-14	12-15
	24-34	25-33	24-33	24-34			34-36	
	50-59	51-59	50-58	50-56	41-48	43-47	44-48	42-49
					61-63		61-63	
					71-75	72-75	69-76	71-75
					80-86	78-86	79-87,89-91	79-86
					96-111	94-101, 105-110	94-111	97-110
					116-124	119-122	116-121	118–124
Lysozyme (chicken)	5-15	5-15	4-15	4-16	1-3		2-3	1-3
	24-34	25-34	24-35	26-27	38-46	41-52	38-40,43-47	38-40,42-46
					50–54		49-53	50-54
	80-85	81-85	79-85		57-60	56-59	58-60	58–61
	88-96	88-100	88-101	88-101		71-80	74-76	
	109-115	109-115	108–115	108-1 15				
	119-124	120-124	120-125					

Myoglobin (whale)	3-18	4–18	3-19	3-11, 11-18		97-101		
	20-35	19-35	20-36	19-22,23-26				
	36-42	36-42	37-43					
		43-49		43-47				
	51-57	52-57	51-58	51-58				
	58-77	58-78	59-78	59-77				
	86-94	82-94	85-97	82-86, 87-95				
	LOO-118	102-l 19	100-l 19	101-115, 116-123				
	125-148	125-149	124-149	124–133, 134-150				
Carboxypeptidase A				3-7	32-36	33-37	32-41	32-41
(bovine)	14-28	15-29	14-29	15-28, 29-33	49-53	45-53	45-53	44-53
	72-88	73-93	72-90	73-83, 84-92	60-66	60-65	59-7 1	60-68
	94-103	94-103	93-102	93-103	104-109	104–108	103-112	103-111
	112-122	113-120	113-121	114-121		128-131	126-131	
		143-147	143-147				139-142	
	173-187	174-186	173-186	173-187			156-158	
	215-231	216-233	215-233	216-231			164-166	
		243-246			190-196	191–195	188-197	191-196
	254-262	254-261	253-261	254-261	ZOO-204	199-203	198-204	ZOO-204
	285–306	283-305	285-305	285–291, 292–307		209-213		
					239-241	234–239	238-240	238-242
					265-271	265-270	264-271	266-271
							272–274	
Total different	(380)	i i	64	93	(158)	101	93	48
% of total	(44·4)‡	9-o	i-5	10·9	(18·5)‡	11·8	10·9	5·6
% of x-helix or β-sheet		20·3	16·8	24·5		63·9	58·9	30·4

† References for these reported secondary structures appear in Tables 3 to 6.

‡ Value in parentheses is the percentage of total residues in secondary structure.

β-strands plus nine extra strands, six of which contain less than four residues. A few of these extra β-strands are spurious, as they are not connected to any other β-strands to form a sheet. They arise when residues at the ends of an a-helix make two hydrogen bonds with one or two distant residues (for example, residues 34 to 36 in ribonuclease and 74 to 76 in lysozyme). These spurious β-strands are readily detected when the list of hydrogen bonds is used to link the p-strands into a regular β-sheet (see Pigs 5 to 21). A few of the β-strands correctly found by this method have been wrongly split into two pieces. The total difference of 93 residues (58·9%) for the H-bond method is only slightly better than that obtained with the a-angle method; this is due mainly to the 32 residues in the nine extra pieces. The C^{α}—C^{α} method also finds all the reported p-strands, but it only finds one extra strand. In one case this method breaks a single β-sheet into two smaller β-sheets (trypsin inhibitor), but the total difference from the reported values is best at only 48 residues to (30·4%).

Most of the extra β-sheets found by the H-bond method are not found by the C^{α}—C^{α} method, in greater agreement with the reported assignments. However, β-sheets are much more difficult to define, as they can sometimes be quite short and irregular. Consequently, investigators usually tend to report the more pronounced multistranded β-sheet regions, often neglecting to mention the shorter, more irregular regions because of uncertainty as to whether these regions are truly β-strand interactions. While the C^{α}—C^{α} method is particularly good at delineating the clear β-regions, the H-bond method is better able to detect these shorter and more irregular regions. Detailed examination of a number of structures and their hydrogen bonding patterns suggested to us that a systematic and objective scheme of β-sheet detection should include these short, irregular regions. Consequently, for the final derived β-sheet assignments we have chosen to simply combine the results of the C^{α}—C^{α} method and the H-bond method. In the final presentation of the derived secondary structure of a large number of globular proteins we distinguished β-strands that are found mainly by each method alone and by both methods together.

Once residues have been assigned to a-helices and β-strands by this combination of methods, left-handed and right-handed turns are assigned using the a-angle method.

(b) *Secondary structure of globular proteins*

The preferred combination of methods described above is now used to derive secondary structure assignments for a large number of globular proteins of known three-dimensional structure. As a further assessment of the performance of our method we first compared many of the derived assignments with the reported values available in the literature (see Table 2). The objective criteria generally find more extra residues than they miss from the reported assignments. While the number of extra a-helical residues derived is only slightly greater than the number missed, considerably more extra β-sheet residues are found than missed. Overall there are 1259 extra residues and 610 missed residues, giving a total difference of 1869 out of 9213 residues (20·3%). When these differences are calculated separately for a-helices and β-sheets, we get 740 residues different out of 3627 a-helical residues (20%), and 1129 residues different out of 2136 b-sheet residues (53%). Clearly, the reported and derived assignments agree much more closely for the a-helices. Very few complete α-helices or β-strands are missed, but the derived assignments do contain a number of additional pieces of secondary structure. Most of the extra β-strands contain three

or less residues, and when such short P-strands are omitted from the derived assignments, the total difference drops to 1602 residues (17·4%).

Figure 4 shows how well the ends match for those α-helices and β-strands both reported in the literature and derived by the present methods. Although most ends match to within two residues, some systematic difference is apparent. The reported positions of the N and C-termini of the α-helices and the N-termini of the β-sheets occur, on average, about one residue after the derived positions. This trend is not very marked, and adding one residue to the derived positions of the N and C-termini of all the a-helices and the N-termini of all the p-strands would actually give a worse total difference of 1882 (compared with 1869).

The complete derived secondary structure assignments for the proteins analysed here are given in Tables 3 to 6. The proteins have been divided into four classes (Levitt & Chothia, 1976) according to their relative amounts of oc-helix and β-sheet and the arrangement of these secondary structure segments along the sequence. These derived assignments are also presented schematically in Tables 7 to 10, which also show the amino acid sequences and the residues that are in left-handed and right-handed turns.

Together with the recognition of the β-strand regions, the programs also derive information about which residues are interacting through hydrogen bonds to form the β-sheet. This information is not easily presented in succinct form in a table, nor would a table provide a complete indication of the actual inter-residue hydrogen bonding patterns. Accordingly, the β-sheet interactions are presented in a series of schematic drawings (Figs 5 to 21), showing both the two-dimensional arrangement of the residues in a β-sheet and the hydrogen bonds formed between them. The arrangement of β-strands in these diagrams was done automatically by the computer program, but some manual rearrangement was sometimes needed to make a more pleasing layout.

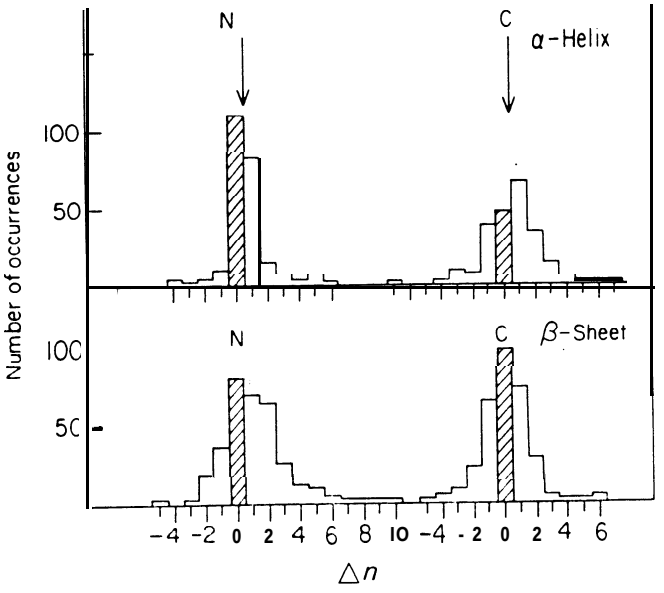


FIG. 4. This Figure shows how often the reported positions of the N and C-termini of α-helices and β-sheets differ from the derived positions, by 0,1,2, etc. residues (measured by the Δn value). The shaded columns of the histograms show $\Delta n = 0$, i.e. identical reported and derived positions. The mean relative position of the reported assignments and their standard deviations are as follows: mean, 0·6, 0·5, 1·0, 0·0; standard deviation, 145, 1·9, 2·2, 1·6 for N-terminal a-helix, C-terminal a-helix, N-terminal β-sheet, and C-terminal /?-sheet, respectively.

TABLE 2

Comparison of reported secondary structure with the final assignments derived in Tables 3 to 6

Protein	No. of residues	Resolution (Å)	a-Helix				β-Sheet				Total % difference
			Residues		Segments		Residues		Segments		
			Extra	Missed	Extra,	Missed	Extra,	Missed	Extra	Missed	
Calcium-binding protein B	108	1·85	4	4	0	0	0	3	—	—	10
Hemerythrin	113	5·5	12	4	1(5)	0	—	—	—	—	14
Carboxyhemoglobin (*Glycera*)	147	2·5	5	5	0	0	—	—	—	—	7
Cyanomethemoglobin (lamprey)	148	2·0	4	5	0	0	—	—	—	—	6
Myoglobin	153	1·4	16	0	0	0	—	—	—	—	10
Aquomethemoglobin (horse)	287	2·8	24	7	0	0	—	—	—	—	12
Deoxyhemoglobin (horse)	287	2·8	25	8	0	0	—	—	—	—	11
Deoxyhemoglobin (human)	287	2·8	24	7	0	0	—	—		—	11
Bence-Jones dimer REI	214	2·0	—			—	27	20	3(2 x 3,5)	0	22
Superoxide dismutase	151	3·0	—	—		—	10	8	2(2x 3)	0	12
Concanavalin A (Argonne)	237	2·4	6	0	1(6)	0	42	8	0	0	24
Concanavalin A (Rockerfeller)	237	2·0	6	0	1(6)	0	35	2	2(2 x 3)	0	18
Chymotrypsin (MRC)	236	2·0	1	8	0	0	42	13	5(4 x 3,4)	0	27
Chymotrypsin (Michigan)	236	2·0	7	6	1(6)	0	31	12	4(3 x 3,4)	0	24
Insulin	102	2·8	13	3	0	0	3	4	1(2)	0	22
Trypsin inhibitor	58	1·5	0	3	0	0	6	3	0	1(1)	16
Cytochrome *b5*	85	2·0	1	5	0	0	2	4	0	0	14
High potential iron protein	85	2·0	2	4	0	1(4)	2	6	0	1(3)	10
Cytochrome c "outer"	103	2·0	2	5	0	0	10	0	2(2 x 5)	0	17

Cytochrome c "inner"	103	2·0	3	6	0	0	13	0	4(2,3,2 x 4)	0	21
Ribonuclease A	124	2·8	0	4	0	0	14	3	2(2,3)	0	17
Ribonuclease S	124	2·0	0	3	0	0	21	1	3(3 x 3)	0	20
Lysozyme	129	2·0	3	1	0	0	5	1	0	0	12
Nuclease (Staphylococcus aureus)	142	2·0	4	2	0	0	12	5	2(2 x 3)	0	16
Papain	212	2·8	5	6	1(5)	0	33	4	5(5 x 3)	2(1,3)	23
Thermolysin	316	2·3	12	3	0	0	38	0	1(4)	0	17
Thioredoxin	108	2·8	8	6	1(7)	1(5)	7	4	0	0	23
Flavodoxin	138	1·9	9	4	1(6)	0	10	5	1(3)	0	20
Adenylate kinase	194	2·8	20	6	0	0	12	0	0	0	20
Phosphoglycerate kinase	218	3·5	29	1	2(6,7)	0	21	8	4(4 x 3)	0	27
Triose phosphate isomerase no. 1	247	2·5	5	13	0	0	13	2	0	0	13
Triose phosphate isomerase no. 2	247	2·5	5	19	0	0	10	3	0	0	15
Subtilisin BPN	275	2·5	13	10	1(5)	1(7)	69	0	9(6 x 3,4,7,8)	0	33
Sub tilisin novo	275	2·8	9	12	1(5)	1(7)	62	3	11(8 x 3,6,2 x 7)	0	31
Carboxypeptidase A	307	2·0	13	6	1(5)	0	47	0	4(2 x 3,4,6)	0	21
Carboxypeptidase B	307	2·8	29	23	1(6)	0	69	6	9(5 x 3,2 x 4,5,6)	0	39
Malate dehydrogenase	325	2·5	7	14	1(5)	0	17	20	3(3 x 3)	1(6)	18
Lactate dehydrogenase	329	2·0	27	20	1(13)	0	30	23	2(2 x 3)	0	30
Lactate dehydrogenase NAD	329	2·8	28	17	1(13)	0	28	25	25(3)	0	30
D-Glyceraldehyde-3-phosphate Dehydrogenase "green"	333	2·9	10	16	0	1(6)	28	25	2(2,3)	(16)	21
Dehydrogenase "red"	333	2·9	9	18	0	1(6)	23	27	1(3)	1(6)	23
Alcohol dehydrogenase	374	2·4	14	9	1(5)	0	39	8	1(3)	0	19
Hexokinase	450	2·7	11	22	1(5)	0	3	39	1(3)	2(6,9)	17
Totals	9213		425	315	116	35	834	295	291	41	
Percentage of total residues	100		4·6	3·4	1·3	0·4	9·0	3·1	3·2	0·4	

TABLE 3

α-Helix and β-sheet secondary structure as derived for nine all-α proteins

Protein	cc-Helix	β-Sheet	Reference(s)
Calcium-binding Protein B (carp)	7-19; **25-33; 41-51; 60-64; 65-70;** 78-88; **99-106;**	57 : 54 **96 : 98**	Kretsinger & Nockholds (1973)
Azomyohcmerythrin *(Thermiste pyroides)*	**10/14;** 18-38; **40-62;** 69-84; **87-105; 106-111;**		Ward *et al.* **(1975)**
Hemerythrin (*Phascolopsus gouldii*)	**10/14; 18-38; 40-62; 69-87; 92-110; 111-116;**		Ward *et al.* **(1975)**
Carboxyhemoglobin *(Glycera dibranchiate)*	**3-15; 21-37; 40-45; 52-63; 64-72; 75-90; 99-104; 105-118; 123-145;**		Padlam & Love **(1974)**
Cyanmethemoglobin (sea lamprey)	**12-19; 20-29; 30-45; 46-52; 60-66; 67-87; 91-105; 111-126; 131-146;**		Hendrickson *et al.* **(1973)**
Metmyoglobin (sperm whale)	**3-19; 20-36; 37-43; 44/49; 51-58; 59-78; 85-97; 100-119; 124-149;**		Kendrew *et al.* **(1960)** Watson (1969)
Aquomethemoglobin (horse)	**3-17; 20-36; 37-41; 42/46; 52-72; 75-80; 81-91; 94-113; 118-138 ;** **4-18; 19-35; 36-42; 50-56; 57-77; 80-86;** 87-96; **99-118; 123-143;**		Perutz *et al.* **(1960)**
Deoxyhemoglobin (horse)	**3-17; 19-36; 37-43;** 52-72; **75-81; 82-91; 94-113; 118-138;** **4-18; 20-35; 36/45; 50-56; 57-77; 80-86; 87-96;** 99-118; 123-143;		Bolton *et al.* **(1968)**
Deoxyhemoglobin (human)	**3-17; 20-36; 37-41; 42/46; 52-72;** 75-81; 82-91; 94-113; **118-138;** **4-18;** 19-35; 36-42; **50-56; 57-77; 80-86; 87-96; 99-118; 123-143;**		Muirhead & Perutz (1963) Fermi (1975)

The assignments derived for the α-helices using the or-angle method are distinguished from those derived by the H-bond method by using a / symbol rather than a – symbol between the first and last residue **in the particular helix. The** assignments derived **for β-strands are divided into 3 classes distinguished by the symbol between the first and** last **residue in the particular strand. This symbol is = if most residues in the strand are assigned by both the** C^{α}—C^{α} and H-bond methods; it is = if most residues in the strand **are assigned by the** C^{α}—C^{α} **method alone; and it is : if most residues are assigned by the H-bond method alone.**

TABLE 4

a-Helix and β-sheet secondary structure as derived for fifteen all-β proteins

Protein	α-Helix	β-Sheet	Reference(s)
Rubredoxin at 2·0 Å (*Clostridium pasteurianum*)		**3=7; 10=14; 17:20; 23:26; 48-51;**	**Herriott et** *al.* **(1970)**
Rubredoxin at 1·5 A (*Clostridium pasteurianum*)		3-7; **10-14; 17:20; 23:26;** 48-52;	**Watenpaugh** *et al. (1973)*
Variable part of Bence-Jones dimer **(human myeloma RFI)**		3-7; **9-13;** 15-26; 32-39; **44-50; 53:55;** 61-66; 69-79; 83-91; **96:99; 101-106;** 3-7; 9-13; **15-26;** 32-39; **43-50; 53:55; 61-66;** 69-79; 83-91; **96:99;** 101-106;	**Epp** *et al.* **(1975)**
Immunoglobulin G Fab' (human myeloma NEW)	**117-122;**	**S-12;** 13-25; 27-30; 33-40; **46:48;** 55-61; 63-73; 78-86; 89-94; **95-101;** 107-115; **124-135; 139-145; 148:150; 153-157;** 160-162; **166-176;** 184-192; **195-203;**	**Poljak** ***et al, (1974)***
	191/195;	**1-7;** 9-12; 16-25; 31-40; **42:44;** 45-52, **54:59; 65-72;** 74-82; **89-99; 101-107;** 109-115; 123-127; **140-149; 153-158; 171:175;** 177-186; 197-204; **207-214;**	
Bence-Jones dimer Mcg (human)	**124-130;** 185-191;	**3-12;** 15-22; **33-40;** 45-52; **54:56;** 62-68; **71-78;** 84-92; 98-109; **115:123;** 132-143; 147-154; **156:158; 161-170; 174-184; 193-200; 203-210;**	**Edmundson et** *al.* (1974)
	124-130; 184-191;	**2-12;** 13-23; 33-40; 45-51; **54:56;** 62-67; **70-80;** 85-94; **96-109; 115:117;** 118-123; 134-142; 147-154; **156:158;** 161-170; 174-182; 193-200; 203-210;	
Prealbumin dimer (human)	76-82;	**11-19; 26-35; 41-50; 53:56;** 66-75; **87-98; 104-112; 115-123;**	**Blake et** *al.* (1974)
	76-82;	**11-19; 22=24; 26-36; 40-50; 53-56; 67-75; 87-97; 104-112;** 114-123;	
Cu,Zn superoxide dismutase (bovine)		**2=9; 14-22; 26-35; 39-46; 80-89; 91-100; 102:104; 108:110; 113-119; 140-149;**	**Richardson et** al. (1975)

TABLE *Acontinued*

Protein	α-Helix	β-Sheet	Reference(s)
Concanavalin A (Argonne) (jack bean)	80-85;	4-14; 20-30; 38:40; 46:55; 59-67; 72-79; 87-98; 100-118; 123-131; 140-148; 153-160; 164-175; 178-181; 185-201; 207-216;	Hardman & Ainsworth (1972)
Concanavalin A (Rockefeller) (jack bean)	80-85;	4-12; 23-30; 33-40; 45-56; 59-68; 70-79; 87-98; 102-118; 123-131; 140-144; 146-150; 153-156; 158-160; 164-166; 168-175; 178-181; 185-201; 207-216;	Reeke *et al. (1975)*
Alkaline serine protease B *(Streptomyces grisseus)*	30-37; 138/142; 173-180;	3=9; 12=21; 25-29; 51-59; 63-71; 78:80; 83-89; 93-101; 104-112; 114:117; 127-134; 145:147; 149-152; 153-162; 164-172;	Delbaere *et al.* (1975)
Trypsin-DIP (bovine)	151-159; 219-228;	8=14; 21-27; 29-37: 40-45; 53-59; 68-80; 83=85; 88=90; 92-99; 119-131; 137:139; 142-150; 167-173; 177=182; 184-190; 193-201; 208-216;	Kreiger *et al. (* 1974)
Alpha chymotrypsin A (MRC) (bovine)	164-170; 234-243;	16=18; 20:22; 29-35; 39-47; 50-55; 64-69; 80-91; 93-96; 99-109; 121:123; 132-144; 150:152; 155-163; 180-185; 188=193; 195-203; 206-216; 224-231;	Birktoft & Blow (1972)
Alpha chymotrypsin A (Michigan) (bovine)	70/75; 164-172; 234-243;	1:3; 19=22; 29-35; 39:41; 42-48; 50-57; 64-69; 80-91; 93:95; 100-109; 121:123; 132-141; 155-163; 180-185; 195-203; 206-216; 224-231;	Tulinsky *et al. (1973)*
Chymotrypsinogen A (bovine)	11-16; 164-172; X2-243:	29-35; 39-47; 50-55; 64-69; 80-91; 93-96; 99-109; 121:123; 132-141; 155-163; 180-186; 197-203; 206-217; 223-231;	Freer *et al.* (1970)
Elastase (porcine)	44-48; 154-160; 227-238:	1=3; 5:7; 14-22; 25-35; 38-43; 52-59; 68-79; 94-100; 123-132; 145-153; 172-175; 181=183; 188-196; 199-209; 221-226;	Watson *et al.* (1970) Sawyer *et al.* (1973)

See footnote to Table 3.

TABLE 5

a-Helix and β-sheet secondary structure as derived for seventeen a + β proteins

Protein	a-Helix	β-Sheet	Reference(s)
Insulin dimer (porcine)	4 - 9 ; 12-20; 24-33; 34-40; 1-9; 12-19; 28-40;	44-49; 20:21; 42-48;	Rlundell *et al.*, (1972)
Trypsin inhibitor (bovine)	2-7; 47-55;	14:25; 28:37; 43:46;	Huber et *al.* (1971) Deisenhofer & S teigemann (1975)
Ferredoxin *(Peptococcus aerogenes)*	13-18; 39-45;	2-5; 22=25; 28=31; 49-52;	Adman *et al.* (1976)
Ferricytochrome *b5* (bovine)	8-16; 33/39; 42-49; 55-61; 64-74; 80-85;	5-7; 21-25; 28-32; 51=54; 75:79;	Mathews *et al.*, (1972)
Oxidized high potential iron Protein *(Chromatium vinosum* D)	11-17;	18=22; 48-51; 59-64; 69=78;	Carter *et al.* (1974)
Ferrocytochrome c (tuna)	1 - 15 ; 50/55; 60-68; 71-75; 87-101;	19=21; 31=33; 69:70; 83:85;	Dickerson & Timkovich (1975)
Ferricytochrome c 'outer' (tuna)	2-12; 49-55; 60-69; 70-75; 87-101;	17:21; 29:33;	Takano *et al.* (1973)
Ferricytochrome c 'inner' (tuna)	2-15; 49-54; 60-68; 71-75; 87-101;	17-20; 29-32; 69:70; 83:85;	
	1-12; 13/17; 49/55; 60-69; 87-101;	18:20; 30:32;	Tanaka *et al.* (1975)
Ferricytochrome *c2* *(Rhondospirillum rubrum)*	2-15; 49-54; 63-72; 73-82; 96-106;	16:21; 27:29; 30:33;	Salemme *et al.* (1973)
Cytochrome *c550* *(Paracoccus denitrifians)*	4-16; 40/44; 55-65; 71-80; 81-90; 106-118;	17-23; 26-31; 33-39;	Timkovich & Dickerson (1973)
Ribonuclease A (bovine)	4-12; 24-33; 50-58;	13-14; 34:36; 43-49; 60:64; 71:76; 77-87; 96-112; 118=123;	Carlisle *et al. (1974)*
Ribonuclease S (bovine)	3-12; 24-33; 50-58;	13-15; 34:36; 42-49; 61:63; 68-76; 78-87; 89:91; 94-111; 116-124;	Wyckoff *et al.* (1970)

TABLE 5—*continued*

Protein	a-Helix	β-Sheet	Reference(s)
Lysozyme (chicken)	4-15; **24-35;** 79-85; **88-101; 108-115; 120-125;**	1-3; **38-40;** 42-47; **49-54; 58-61; 74:76;**	Blake et *al.* (1965,1967)
Nuclease *(Staphylococcus aureus)*	**54-63; 64-69; 100-106; 121-135;**	8-19; 21-27; **30-36; 38=41;** 71-76; **81:83;** 87-95; **97:99; 109-112;**	**Arnone et *al.* (1971)**
Papain **(papaya)**	**24-40; 49-57;** 67-77; **96/100;** 117-127; 138-143;	**17:19;** 79-81; **92:94;** 104-106; **108-113; 129-135; 147:149;** 158-167; **169-175; 177:179; 185-191; 205-211;**	**Drenth et cd. (197 1 (l)**
Thermolysin (***Bacilus*** *thermoproteolyticus)*	66-88; **136-153; 159-180;** 229-245; 261-275; **280-295;** 300-313;	4-12; **16-25;** 27-33; **36-45; 52-58;** 60-62; **95-106; 109-116;** 119-126; **130:133; 186-190; 201=205;** 246-250; **253-258;**	**Colman** *et al.* (1972)

See footnote to Table 3.

TABLE 6

a-Helix and β-sheet secondary structure as derived for 21 α/β proteins

Protein	a-Helix	β-Sheet	Reference(s)
Oxidized thioredoxin (*Escherichia coli*)	11-18; 34-49; 64/70; 94-106;	4-7; 21-28; 52-59; 77-82; 85-91;	Holmgren *et al.* (1975)
Oxidized flavodoxin (*Clostridium* MP)	10-26; 39-44; 64-75; 93-106; 124-136;	1-7; 30-34; 47-57; 60:62; 78-88; 107-111; 114=118;	Burnett *et al.* (1974)
Semiquinone f lavodoxin (*Clostridium* MP)	10-26; 39-46; 64-74; 93-106; 124-136;	l-7; 30-34; 47-57; 60:62; 78-88; 107-111; 114-118;	Mayhew & Ludwig (1975
Adenylate kinase (porcine)	l-7; 21-31; 39-49; 52-61; 68-83; 100-109; 121-136; 141-157; 159-166; 179-192;	8-15; 34-38; 88-95; 113-120; 169-175;	Schulz *et al.* (1974a)
Phosphoglycerate mutasc (yeast)	10-15; 26-42; 57-71; 97-103; 104-113; 128/134; 145-153; 154-165; 172-184;	l-7; 16:18; 50-55; 76-81; 94:96; 117:119; 142:144; 167-171; 196-203; 207:209; 212-214;	Campbell *et al.* (1974)
Triose phosphate isomerase dimer (chicken)	16-30; 43-54; 78-85; 95-101; 104-118; 130-135; 137-153; 176-195; 196-203; 213-222; 232-244;	4-11; 35-42; 58-65; 88-94; 120-129; 158-167; 204-208; 226-231;	Banner *et al.* (1975)
	16-30; 45-54; 78-85; 94-101; 104-118; 130-135; 136-153; 176-195; 196-202; -215-222; 236-245;	4-11; 35-42; 58-64; 88-93; 120-129; 158-166; 204-208; 227-231;	
Carbonic anhydrase B (human)	10/16; 127/133; 151/155; 156-161; 176-181; 215-224;	1:3; 6:9; 24-29; 34:36; 46:48; 52-58; 61-67; 72:74; 83-93; 102-105; 112-121; 136-146; 168-172; 186-195; 199-209; 210-214; 226=228; 234=236; 243:245; 252-255;	Kannan *et al.* (1976)
Carbonic anhydrase C (human)	10/14; 126-132; 150-161; 177/181; 215-225;	2:4; 7:9; 24-31; 35:38; 43=48; 53-59; 62-68; 74-79; 84-94; 103-107; 112-122; 136-146; 169-173; 186-194; 202-209; 211-214; 252-256;	Liljas et *al.* (19'72)
Subtilisin BPN' (*Bacillus amyloliquifaciens*)	5-10; 12-19; 63-71; 103-117; 132-145; 167/171; 219-238; 242-253;	1=4; 22:24; 26-32; 45=51; 72:74; 79=82; 83-96; 120-125; 147-153; 155:157; 174-181; 184:186; 190:192; 196:199; 200:203; 204-209; 213-218; 254:256; 263:269;	Wright *et al.* (1969) Poulos *et al.* (1977)

TABLE &-continued

Protein	α-Helix	β-Sheet	Reference(s)
Subtilisin novo *(Bacillus amyloliquifuciens)*	12-19; 63-71; 103-117; 132-145; **167/171;** 219-238; 242-253;	**2:4; 8:10; 26:31; 35:37; 43:45; 47:50; 58:60; 72:74; 80:86; 88:97; 100:102; 120:125; 147:153; 174:181; 184:186; 196:199; 200:202; 204:206; 207:210; 212:215; 216:218; 254:256; 263:268;**	Drenth et al. (1971*b*,1972)
Carboxpeptidase A (bovine)	14-29; **72-90; 93-102;** 113-121; 143-147; 173-186; 215-233; 253-261; **285-305;**	32-41: 44-53; **59-71; 103-112; 126:131; 139:142; 156:158; 164:166;** 188-197; **198-204;** 238-242; 264-271; **272:274;**	Hartsuck & Lipscomb (1971) Ludwig et *al.* (1967)
Carboxpeptidase B (bovine)	**11-26; 69-79; 90-100;** 109-120; 170-183: 213-217; 218-229; **240-245;** 252-259; 284-294; **299-304;**	**28-38;** 41-51; 57-65; **66:68; 101:108;** 123-128; 136-139; **144:146; 153:155; 161:164;** 186-195; **197-203; 230:232;** 236-239; **249:251; 262-269; 270:272; 279:283; 296:298;**	Schmid & Herriott (1976)
Malate dehydrogenase (porcine)	15-24; 43-51; 52-58; **97-102; 103-107;** 123-136; 147-162; **201/205; 207-212;** 217-225; 234-251; 297-323;	**5-11; 25:27;** 32-39; **62-69; 78-84; 115-121; 143=146; 169-173; 174-177; 183=186; 190=193; 198:200; 214:216;** 256-261; **262:264; 275:277;** 2787282;	Webb et al. (1973)
Lactate dehydrogenase (dogfish)	2-8; **35-42; 54-70; 105-121;** 122-128; 139-151; 163-178; 224-241; 247-261; **306-324;**	21-29; **32:34;** 43-52; **75-81; 90-96;** 131-137; 157-161; **185=187;** 188-192: **198:200; 204-206;** 299-211; **265:273; 283:291; 299:301;**	Adams et *al.* (1970)
Lactate dehydrogenase-NAD (dogfish)	2-8; 28-42; 54-69; **105-115;** 116-127; **139-151;** 164-178; 224-242; **248-261;** 338-326;	21-27; 46-52; 75-81; **87-96; 128:129; 130-137; 157-160; 188:190; 198:200; 204:206; 209:211; 265:268; 269:273; 283:292; 295:297; 299-301;**	White *et al.* (1976)
D-Glyceraldehyde-3-phosphate dehydrogenase (lobster) "green"	9-23; 35-44; **100-110; 147-158;** 159-164; **208-217; 218/222; 250-264; 313-331;**	1-7; 25-32; **45:46; 50:52; 56:59; 62=65; 68-75; 88-94; 112-119;** 124-128; **141-146; 165-176; 202-207;** 223-231; 236-246; 268-274; **287-293; 297-300; 302-311;**	
D-Glyceraldehyde-3-phosphate dehydrogenase (lobster) "red"	9-23; **35-45; 100-110; 147-164; 209-214; 218/222;** 250-264; 3X-331;	2-7; **27-32; 56-59; 62-65;** 68-75; **81:83;** 89-94; **111-119; 124-127; 141-146; 167-177; 202-206;** 223-231; **235-245;** 268-274; 287-293: **297-300; 302-312:**	Moras *et al.* (1975) Adams *et al.* (1973)

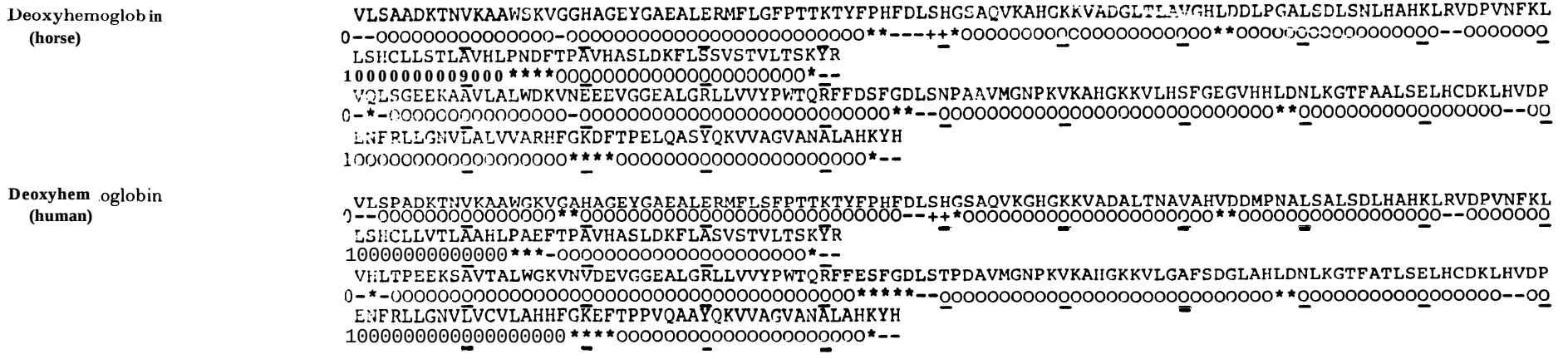
Alchohol dehydrogenase (horse)	46-53; 101/105; 166/172; 173-182; 183-187; 201-214; 228-235; 249-258; 271-282; 306/310; 323-337; 354-363;	4=7; 8:15; 21-31; 32-45; 62-79; 87-92; 126-132; 135:139; 143:145; 148-153; 156=162; 191-199; 215-222; 238-241; 263-270; 286-294; 311-316; 338:340; 345-351; 368-374;	Eklund et *al.* (1976)
Phosphoglycerate kinase (yeast)	21-36; 52-63; 73-77; 91-102; 124-129; 147-163; 172-178; 195-202; 248-262; 307-317; 348-363;	2=4; 10=12; 14-19; 40-49; 64-72; 103-108; 111:113; 117:119; 131-138; 167-171; 187-191; 230-236; 237:239; 265-270; 272:274; 297-303; 318:319; 320-326; 364-371;	Bryant et *al.* (1974)
Phosphoglycerate kinase (horse)	34-50; 70/74; 75-86; 99-108; 141-155; 164-170; 189-203; 216-228; 238-246; 257-271; 313-318; 319-326; 345-360; 368-374; 385-394; 398-403;	13-21; 25:27; 29:31; 51-61; 90-93; 111-119; 127-132; 135-140; 156-163; 183-188; 205-211; 212:214; 230-237; 249:251; 273-276; 279-285; 293-299; 302:304; 307:312; 327-332; 336:338; 342:344; 362-367; 380-384;	Blake & Evans (1974)
Hexokinase B III (yeast)	4-10; 14-18; 20-25; 26-39; 105-114; 115-120; 165-171; 172-176; 187-192; 193-197; 255-264; 270-276; 280-286; 287-295; 311-320; 338-361; 362-370; 383/387; 390-401; 422-433; 434-441;	60-69; 72-82; 85-95; 129-135; 180=186; 203-208; 213-219; 237-242; 296:298; 377:381;	Steitz et cd. (1976)

See footnote to Table 3.

TABLE 7

Symbolic secondary structure as derived for nine *all-α* ***proteins***

Protein	Symbolic secondary structure†
Calcium-binding protein B (carp)	AFAGVLNDADIAAALEACKAADSFNHKAFFAKVGLTSKSADDVKKAFAIIDQDKSGFIEEDELKLFLQNFKADARALTDGETKTFLKAGDSDGDGKIGVD 0--*+--OOOOOOOOOOOOO*++--OOOOOOOOO-***---OOOOOOOOOOO**---///OOOOOOOOOOO***----OOOOOOOOOO*****---///OO EFTALVKA 1OOOOOO*-
Azomyohemerythrin (*Thermiste pyroides*)	GFPIPDPYVWDPSFRTFYSIIDDEHKTLFNGIFHLAIDDNADNLGELRRCTGKHFLNQEVLMEASQYQFYDEHKKEHDGFINALDNWKGDVKWAKAWLVN 0--*-++*--OOOOO---OOOOOOOOOOOOOOOOOOOO-OOOOOOOOOOOOOOOOOOOOOO++*+++OOOOOOOOOOOOOOOOO*-OOOOOOOOOOOOOO HIKTIDFKYKGKI 100000000OO*-
Hemerythrin (*Phascolopsus gouldii*)	GWEIPEPYVWDESFRVFYEQLDEEHKKIFKGIFCDIRDNSAPNLATLVKVTTNHFTHEEAMMDAAKYSEVVPHKKMHKDFLEKIGGLSAPVDAKNVDYCK 0--*-++*--OOOOO---OOOOOOOOOOOOOOOOOOOO-OOOOOOOOOOOOOOOOOOOOOO++*+++OOOOOOOOOOOOOOOOOOO++*-OOOOOOOOO EWLVNHIKGTDFKYKGKL - 1000000000OOOOOO*-
Carboxyhemoglobin (*Glycera dibranchiata*)	GLSAAQRQVIAATWKDIAGNDNGAGVGKDCLIKHLSAHPEMAAVFGFSGASDPAVADLGAKVLAEIGVAVSHLGDEGKMVAEMKAVGVRHKGYGDKHIKG O--OOOOOOOOOOOOO *****OOOOOOOOOOOOOOOO+-OOOOOO**++ *-OOOOOOOOOOOOOOOOOOOO--OOOOOOOOOOOOOOOO**++***-OO EYFEPLGASLLSAMEHRIGGKMDAAAKDAWAAAYADISGALISGLES 1000000000OOOOOOOOO *-***OOOOOOOOOOOOOOOOOOOOOOO*-
Cyanmethemoglobin (sea lamprey)	PIVDTGSVAPLSAAEKTKIRSAWAPVYSDYETSGVDILVKFFTSTPAAEEFFPKFKGLTTADELKKSADVRWHAERIIDAVDDAVASMDDTEKMSSMKDL 0--****+--++OOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOO***** --OOOOOOOOOOOOOOOOOOOOOOOOOOOO***OOOOOOOOOO SGKHAKSFEVDPEYFKVLAAVIADTVAAGDAGFEKLLRMICILLRSAY 1OOOOO++*--OOOOOOOOOOOOOOOO+**+OOOOOOOOOOOOOOOO*-
Metmyoglobin (sperm whale)	VLSEGEWQLVLHVWAKVEADVAGHGQDILIRLFKSHPETLEKFDRFKHLKTEAEMKASEDLKKHGVTVLTALGAILKKKGHHEAELKPLAQSHATKHKIP 0--OOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOOO-OOOOOOOOOOOOOOOOOOOOOOOOOOOO+*--**OOOOOOOOOOOOO--O IKYLEFISEAIIHVLHSRHPGNFGADAQGAMNKALELFRKDIAAKYKELGYQG 1000000000OOOOOOOOO *****OOOOOOOOOOOOOOOOOOOOOOOOO-**-
Aquomethemoglobin (horse)	VLSAADKTNVKAAWSKVGGHAGEYGAEALERMFLGFPTTKTYFPHFDLSHGSAQVKAHGKKVGDALTLAVGHLDDLPGALSDLSNLHAHKLRVDPVNFKL 0LSHCLLSTLAVHLPNDFTPAVHASLDKFLSSVSTVLTSKYR OOOOO---**OOOOOOOOOOOOOOOOOO**OOOOOOOOOOOOOOOO--OOOOOOO 10000000000000 ****OOOOOOOOOOOOOOOOOOOO*-- VQLSGEEKAAVLALWDKVNEEEVGGEALGRLLVVYPWTQRFFDSFGDLSNPGAVMGNPKVKAHGKKVLHSFGEGVHHLDNLKGTFAALSELHCDKLHVDP 0ENFRLLGNVLVVVLARHFGKDFTPELQASYQKVVAGVANALAHKYH*--OOOOOOOOOOOOOOOOOOOOOOOOOOO**OOOOOOOOOO OOOOOO--OO 100000000000000000 ****OOOOOOOOOOOOOOOOOOOO*--



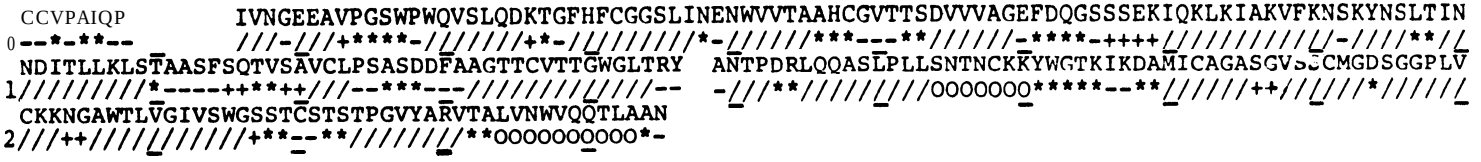
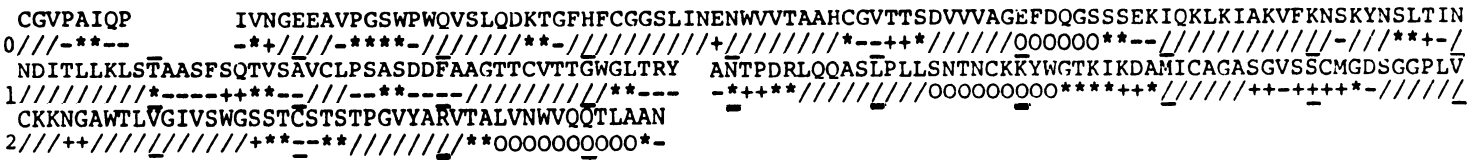
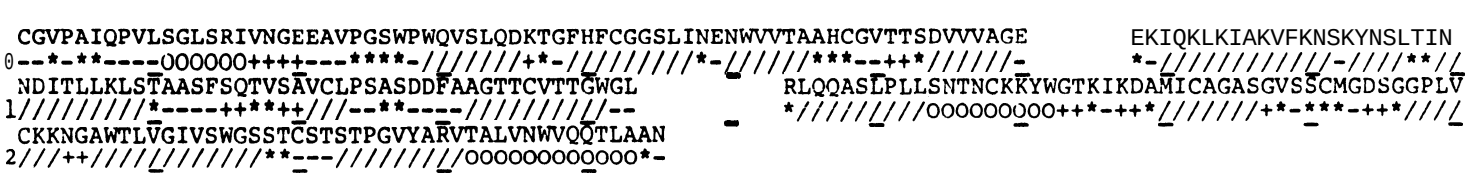
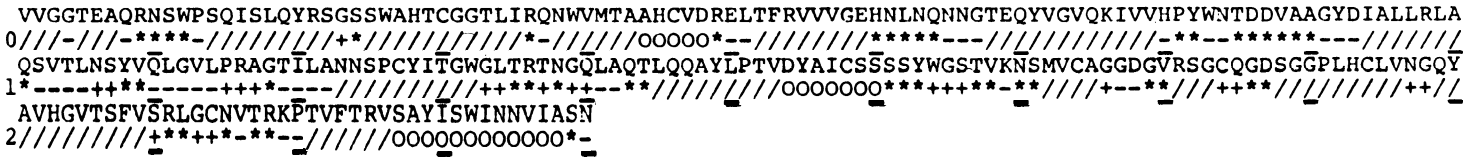
† Residues judged to be in α-helix are shown as 0, in β-sheet as /, in right-handed turns as *, in left-handed turns as +, and in undefined secondary structure as —. There are 100 residues in each row with every tenth residue's symbol underlined. The first symbol shows the number of hundreds that must be added to the residue numbers in that row. Each chain in multimeric proteins is numbered separately. The amino acid sequence is given above the secondary structure using the one letter abbreviations (see legend to Fig. 5).

TABLE 8

Symbolic secondary structure as derived for 15 all-β proteins

Protein	Symbolic secondary structure
Rubredoxin at 2·0 A *(Cloetridium pasteurianum)*	MKKYTCTVCGYDYTPDIGDPVDGIDPGVEFKDIPDDWVCPVCGIGKDEFEEVD 0--/////**/////**///*/**////*--***--**---***---**//*/--
Rubredoxin at 1·5 A *(Clostridium pasteurianum)*	MKKYTCTICGYVYDPEDGDPEDGIDPGTDFKDIPDDWVCPICGIGKDEFEEVEE 0--/////**/////**///*/**////*--***--**-++***---**//*//--
Variable part of Bence-Jones dimer (human myeloma REI)	DIQMTQSPSSLSASVGDRVTITCQASQDIIKYLNWYQQTPGKAPKLLIYEASNLQAGVPSRFSGSGSGTDYTFTISSLQPEDIATYYCQQYQSLPYTFGQ 0--/////*////-/////L////// *+-++////////++++//////*+*///++---*//////++/*/////////***////////*/**+*////- GTKLQIT 1//////- DIQMTQSPSSLSASVGDRVTITCQASQDIIKYLNWYQQTPGKAPKLLIYEASNLQAGVPSRFSGSGSGTDYTFTISSLQPEDIATYYCQQYQSLPYTFGQ 0-*/////*/L///-/////L/////// --++////////+*-////////*+*///+*---*//////++/*/////////***////////*/+++*////- GTKLQIT 1//////
Immunoglobulin G Fab' (human myeloma NEW)	SVLTQPPSVSGAPGQRVTISCTGSSSNIGAGNHVKWYQQLPGTAPKLLIFHNNARFSVSKSGSSATLAITGLQAEDEADYYCQSYDRSLRVFGGGTKLTV 0--*-/////////////////////*////--////////+***++///--++--/////*/////////***-/////////++///////// LRQPKAAPSVTLFPPSSEELQANKATLVCLISDFYPGAVTVAWKADSSPVKAGVETTTPSKQSNNKYAASSYLSLTPEQWKSHKSYSCQVTHEGSTVEKT 1/*----///*/////-OOOOOO+/////*/////***/*/////+*//*--/////--*//+**+////*/////++**---////////*/++/////* VAPTECS 2///*--- VQLEQSGPGLVRPSQTLSLTCTVSGSTFSNDYYTWVRQPPGRGLEWIGYVFYHGTSDTDTPLRSRVTMLVNTSKNQFSLRLSSVTAADTAVYYCARDLIA 0///////-///**-/////////--***/////////*/////////*/////*****///////+/////////*--***/////////+ GCIDVWGQGSLVTVSSASTKGPSVFPLAPSSKSTSGGTAALGCLVKDYFPEPVTVSWNSGALTSGVHTFPAVLQSSGLYSLSSVVTVPSSSLGTQTYICN 1///////+/*/////------///*+*-++-++*+-*/////////***//////++**-++---++/////*/////////+++*OOOOO-///* VNHKPSNTKVDKRVEPKSC 2////**///*////--++-
Bence–Jones dimer Mcg (human)	SALTQPPSASGSLGQSVTISCTGTSSDVGGYNYVSWYQQHAGKAPKVIIYEVNKRPSGVPDRFSGSKSGNTASLTVSGLQAEDEADYYCSSYEGSDNFVF 0--/////////+*/////////++*******-/////////++++/////////*///***--////////**////////--***/////////+++**/// GTGTKVTVLGQPKANPTVTLFPPSSEELQANKATLVCLISDFYPGAVTVAWKADGSPVKAGVETTKPSKQSNNKYAASSYLSLTPEQWKSHRSYSCQVTH 1/////////**+--/////*///OOOOOO*-/////////*///***///*////+///--*/////////*+*-/////*////OOOOOOO-////////* EGSTVEKTVAPTECS 2++///////*----- SALTQPPSASGSLGQSVTISCTGTSSDVGGYNYVSWYQQHAGKAPKVIIYEVNKRPSGVPDRFSGSKSGNTASLTVSGLQAEDEADYYCSSYEGSDNFVF 0-/////////////////////*///++***+***-////////+*++//////*+*///+*---*//////**/////////*/**--/////////*/////* GTGTKVTVLGQPKANPTVTLFPPSSEELQANKATLVCLISDFYPGAVTVAWKADGSPVKAGVETTKPSKQSNNKYAASSYLSLTPEQWKSHRSYSCQVTH 1/////////*----/////*///OOOOOOO*---/////////**--/////*/+///--*/////////*+*+/////////*//-OOOOOOOO-////////* EGSTVEKTVAPTECS 2++///////*+***+-

TABLE 8—*continued*

Protein	Symbolic secondary structure
Alpha chymotrypsin A (MRC) (bovine)	
Alpha chymotrypsin A (Michigan) (bovine)	
Chymotrypsinogen A (bovine)	
Elastase (porcine)	

See footnote to Table 7.


Prealbumin dimer
(human)
Cu Zn auperoxide dismutase
(bovine)
Concanavalin A (Argonne)
(jack bean)
Concanavalin A (Rockerfeller)
(jack bean)
Alkaline serine protease B
(Streptomyces grisseus)
Trypsin-DIP
(bovine)

TABLE 9

Symbolic secondary structure as derived for 15 a + β proteins

Protein	Symbolic secondary structure
Insulin dimer (porcine)	GIVEQCCTSICSLYQLENYCNFVNQHLCGSHLVEALYLVCGERGFFYTPKA 0--*OOOOOO+-OOOOOOOOO--*OOOOOOOOOOOOOOOO++*//////-- GIVEQCCTSICSLYQLENYCNFVNQHLCGSHLVEALYLVCGERGFFYTPKA OOOOOOOOOO+-OOOOOOOO//-****--OOOOOOOOOOOO-////////**-
Trypsin inhibitor (bovine)	RPDFCLEPPYTGPCKARIIRYFYNAKAGLCQTFVYGGCRAKRNNFKSAEDCMRTCGGA 0-OOOOOO++--++//////////////**//////////**---////OOOOOOOOO*--
Ferredoxin (*Peptococcus aerogenes*)	AYVINDSCIACGACKPECPVNIIQGSIYAIDADSCIDCGSCASVCPVGAPNPED 0-////+*-----OOOOOO**-////+*//////***----OOOOOOO**-/////*-
Ferricytochrome *b5* (bovine)	SKAVKYYTLEQIEKHNNSKSTWLILHYKVYDLTKFLEEHPGGEEVLREQAGGDATEDFEDVGHSTDARELSKTFIIGELHPDDRSKI 0 --///OOOOOOOOO****/////++////0000000**OOOOOOOO-////OOOOOOO--OOOOOOOOOO/////OOOOOO*-
Oxidized high potential iron protein (*Chromatium vinosum* D)	SAPANAVAADNATAIALKYNQDATKSERVAAARPGLPPEEQHCADCQFMQADAAGATDEWKGCQLFPGKLINVNGWCASWTLKAG 0--****---**OOOOOOO//////****--*****--**++****-++***--////*-**-++/////****/////////*------
Ferrocytochrome c (tuna)	GDVAKGKKTFVQKCAQCHTVENGGKHKVGPNLWGLFGRKTGQAEGYSYTDANKSKGIVWNNDTLMEYLENPKKYIPGTKMIFAGIKKKGERQDLVAYLKS OOOOOOOOOOOOOOOO***//+**---++*-///*-**-++--++--++OOOOOO**--OOOOOOOOO//OOOOO**-----///-OOOOOOOOOOOOOO ATS 10*-
Ferricytochrome c "outer" (tuna)	GDVAKGKKTFVQKCAQCHTVENGGKHKVGPNLWGLFGRKTGQAEGYSYTDANKSKGIVWNNDTLMEYLENPKKYIPGTKMIFAGIKKKGERQDLVAYLKS 0-OOOOOOOOOOO****///**---++/////*-**--**--**---OOOOOOO----OOOOOOOOOOOOOOOO**---------OOOOOOOOOOOOO ATS 10*-
Ferricytochrome c "inner" (tuna)	GDVAKGKKTFVQKCAQCHTVENGGKHKVGPNLWGLFGRKTGQAEGYSYTDANKSKGIVWNNDTLMEYLENPKKYIPGTKMIFAGIKKKGERQDLVAYLKS 0-OOOOOOOOOOOOOO*////++++++-++////**-++--**--++---OOOOOO-----OOOOOOOOO//OOOOO**-----///-OOOOOOOOOOOOO ATS 10*-
Ferrocytochrome c (Bonito)	GDVAKGKKTFVQKCAQCHTVENGGKHKVGPNLWGLFGRKTGQAQGYSYTDANKSKGIVWNQNTLMEYLENPKKYIPGTKMIFAGIKKKGERQDLVAYLKS OOOOOOOOOOOOOOOO//*++-++++*//++-**-++*++++-**OOOOOO----OOOOOOOOOO+****+**-----++--OOOOOOOOOOOOO ATS 10*-
Ferricytochrome *c2* (*Rhodospirillum rubrum*)	EGDAAAGEKVSKKCLACHTFDQGGANKVGPNLFGVFENTAAHKDNYAYSESYTEMKAKGLTWTEANLAAYVKNPKAFVLEKSGDPKAKSKMTFKLTKDDE 0-OOOOOOOOOOOOO/////**++-////////*-++++*******--OOOOOO****----OOOOOOOOOOOOOOOOOOOO+-++---**----OOOOO IENVIAYLKTLK 1000000*++**-

TABLE 9—*continued*

Protein	Symbolic secondary structure
Cytochrome *c*550 (*Paracoccus denitrificans*)	NEGDAAKGEKEFNKCKACHMIQAPDGTDIKGGKTGPNLYGVVGRKIASEEGFKYGEGILEVAEKNPDLTWTEANLIEYVTDPKPLVKKMTDDKGAKTKMT 0--*OOOOOOOOOOOOO/////////+*//////+///////OOOOO+++--**---OOOOOOOOOOO**---OOOOOOOOOOOOOOOOOOOO--**---**- FKMGKNQADVVAFLAQDDPDAGQGQAAGAGSNSQ 1---**OOOOOOOOOOOOO--------++++++-++**--
Ribonuclease A (bovine)	ETAAAKFERQHMDSSTSAASSSNYCNEMMKSRNLTKDRCKPVNTFVHESLADVQAVCSQKNVACKNGQTNCYQSYSTMSITDCRETGSSKYPNCAYKTT 0 --OOOOOOOOO//++-++++++OOOOOOOOOO///++*---///////OOOOOOOOO*/////-*****//////////////++*--***-///// QANKHIIVACEGNPYVPVHFDASV 1//////////////+*-++/////-/
Ribonuclease S (bovine)	KETAAAKFERQHMDSSTSAASSSNYCNQMMKSRNLTKDRCKPVNTFVHESLADVQAVCSQKNVACKNGQTNCYQSYSTMSITDCRETGSSKYPNCAYKTT 0--OOOOOOOOOO///-++---*-OOOOOOOOOO///**++-////////OOOOOOOOO**///+-**//////////*//////////+//+*/////// QANKHIIVACEGNPYVPVHFDASV 1//////////////*+*-////////////
Lysozyme (chicken)	KVFGRCELAAAMKRHGLDNYRGYSLGNWVCAAKFESNFNTQATNRNTNGSTDYGILQINSRWWCDNGRTPGSRNLCNIPCSALLSSDITASVNCAKKIVS 0///OOOOOOOOOOO--+++++-OOOOOOOOOOO+-///*//////*//////+*-///*****--*****--///--OOOOOOO**OOOOOOOOOOOO DGDGMNAWVAWRNRCKGTDVQAWIRGCRL 1O+*+++*OOOOOOOO**--OOOOOO*---
Nuclease (*Staphylococcus aureus*)	ATSTKKLHKEPATLIKAIDGDTVKLMYKGQPMTFRLLLVDTPETKHPKKGVEKYGPEASAFTKKMVENAKKIEVEFNKGQRTDKYGRGLAYIYADGKMVN 0--****+*//////////+///////++///////-///+-***-++***+OOOOOOOOOOOOOOO-//////-***///**-////////+///O EALVRQGLAKVAYVYKPNNTHEQHLRKSEAQAKKEKLNIWSENDADSGQ 1OOOOOO--///*--****-OOOOOOOOOOOOOOO-++***-
Papain (Papaya)	IPEYVDWRQKGAVTPVKNQGSCGSCWAFSAVVTIEGIIKIRTGNLNQYSEQELLDCDRRSYGCNGGYPWSALQLVAQYGIHYRNTYPYEGVQRYCRSREK 0--**--****-**++-///++--OOOOOOOOOOOOOOOO**------OOOOOOOO***-+++*-OOOOOOOOOO+//+****---**///-OOOOO GPYAAKTDGVRQVQPYNQGALLYSIANQPVSVVLQAAGKDFQLYRGGIFVGPCGNKVDHAVAAVGYGPNYILIKNSWGTGWGENGYIRIKRGTGNSYGVC 1**-///-//////+**OOOOOOOOOO+//////**OOOOO-++///--------//////////+//////-///**+++//////*---**++ GLYTSSFYPVKN 2+**-///////-
Thermolysin (*Bacillus thermoproteolyticus*)	ITGTSTVGVGRGVLGDQKNINTTYSTYYYLQDNTRGDGIFTYDAKYRTTLPGSLWADADNQFFASYDAPAVDAHYYAGVTYDYYKNVHNRLSYDGNNAAI 0--*//////////**-////////////+///////**////////////+---**/////////*///---OOOOOOOOOOOOOOOOOOOOOOO--++++*////// RSSVHYSQGYNNAFWNGSEMVYGDGDGQTFIPLSGGIDVVAHELTHAVTDYTAGLIYQNESGAINEAISDIFGTLVEFYANKNPDWEIGEDVYTPGISGD 1//////++/////////**/////////***////**OOOOOOOOOOOOOOOOOO---**OOOOOOOOOOOOOOOOOOOOOO-++--/////**++******- SLRSMSDPAKYGDPDHYSKRYTGTQDNGGVHINSGIINKAAYLISQGGTHYGVSVVGIGRDKLGKIFYRALTQYLTPTSNFSQLRAAAVQSATDLYGSTS 2/////******+--*****-----***++OOOOOOOOOOOOOOOO/////++//////--OOOOOOOOOOOOOOO-**-OOOOOOOOOOOOOOO*+++O QEVASVKQAFDAVGVK 3OOOOOOOOOOOOO---

See footnote to Table 7.

TABLE 10

Symbolic secondary structure as derived for 21 α/β proteins

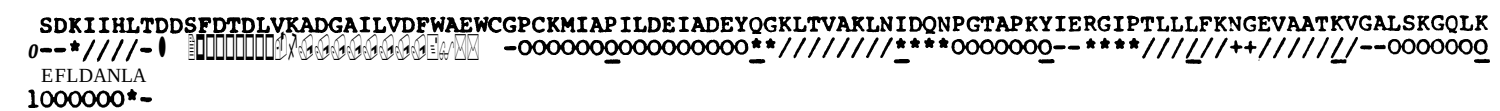
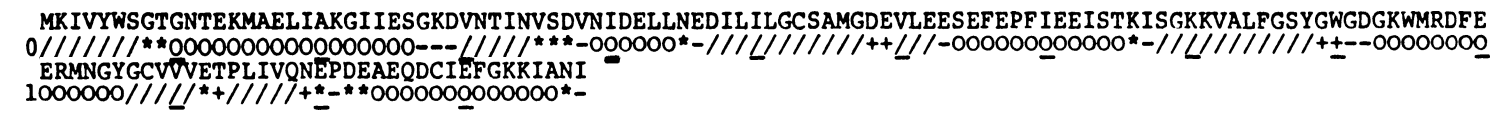
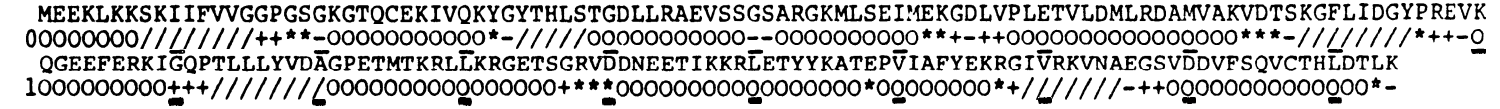
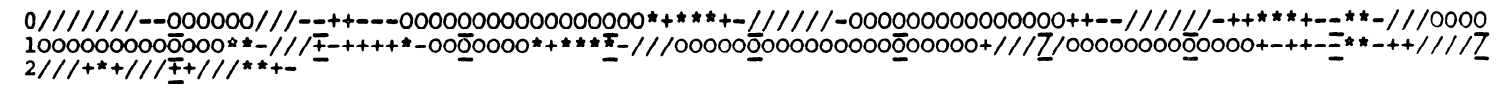
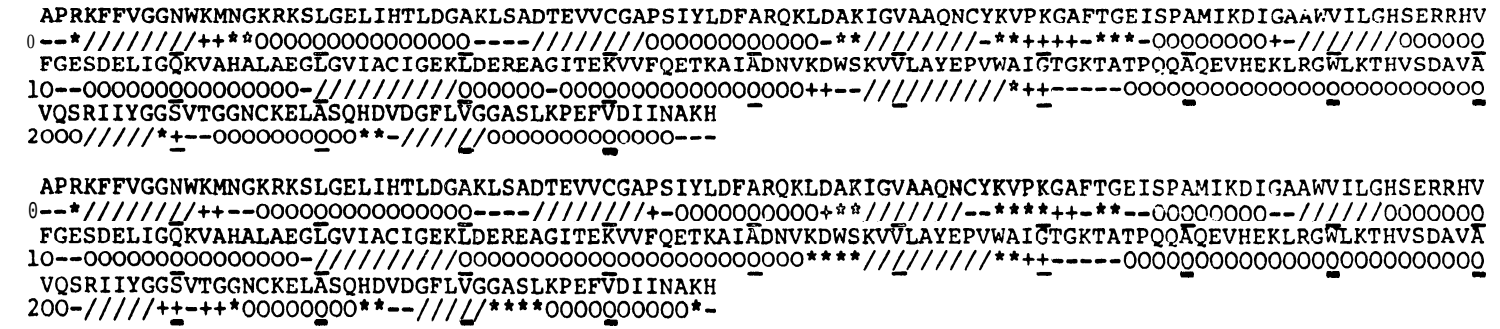
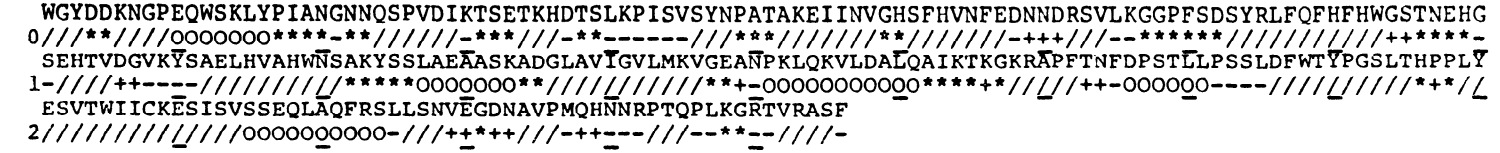
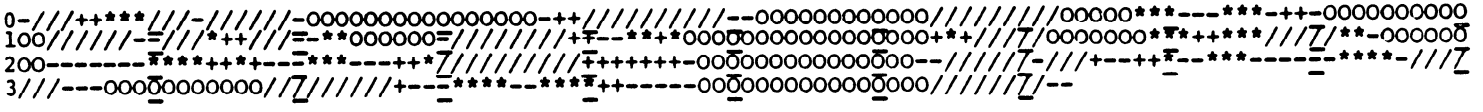
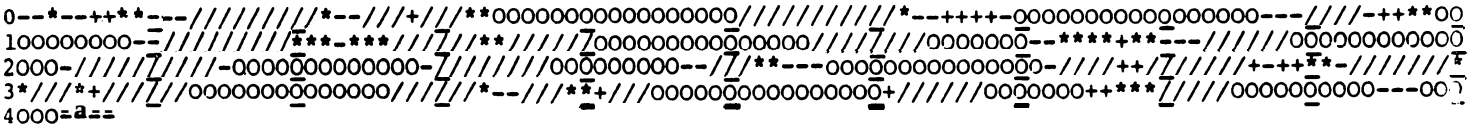
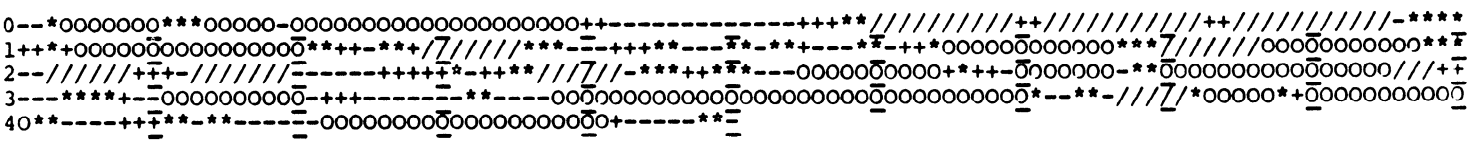
Protein	Symbolic secondary structure
Oxidized thioredoxin (*Escherichia coli*)	
Oxidized flavodoxin (*Clostridium* MP)	
Semiquinone flavodoxin (*Clostridium* MP)	
Adenylate kinase (porcine)	
Phosphoglycerate mutase (yeast)	
Triose phosphate isomerase dimer (chicken)	
Carbonic anhydrase B (human)	

TABLE 10 *continued*

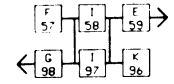
Protein	Symbolic secondary structure
Carbonic anhydrase C (human)	HWGYGKHDGPEHWHKDFPIAKGERQSPVDIDTHTAKYDPSLKPLSVSYDQATSLRILNDGHAFNVEFDDSEDKAVLKGGPLDGTYRLIQFHFHWGSLDGE GSQHTVDKKKYAAELHLVHWNTKYGDFGKAVQQPDGLAVLGIFLKVGSAKPGLQKVVDVLDSIKTKGKSADFTNFDPRGLLPESLDYWTYPGSLTTPPLL ECVTWIVLKEPISVSSEQVLKFRKLNFDGEGEPEELMVDNWRPAQPLKNRQIKASF
Subtilisin BPN' (*Bacillus amyloliquifaciens*)	AQSVPYGVSQIKAPALHSQGYTGSNVKVAVIDSGIDSSHPDLKVAGGASMVPSETPNFQDDNSHGTHVAGTVAALNNSIGVLGVAPSSALYAVKVLGDAG SGQYSWIINGIEWAIANNMDVINMSLGGPSGSAALKAAVDKAVASGVVVVAAAGNEGSTGSSSTVGYPGKYPSVIAVGAVDSSNQRASFSSVGPELDVMA PGVSIQSTLPGNKYGAYNGTSMASPHVAGAAALILSKHPNWTNTQVRSSLQNTTTKLGDSFYYGKGLINVQAAAQ
Subtilisin novo (*Bacillus amyloliquifaciens*)	AQSVPYGVSQIKAPALHSQGYTGSNVKVAVIDSGIDSSHPDLKVAGGASMVPSETPNFQDDNSHGTHVAGTVAALNNSIGVLGVAPSSALYAVKVLGDAG SGQYSWIINGIEWAIANNMDVINMSLGGPSGSAALKAAVDKAVASGVVVVAAAGNEGSTGSSSTVGYPGKYPSVIAVGAVDSSNQRASFSSVGPELDVMA PGVSIQSTLPGNKYGAYNGTSMASPHVAGAAALILSKHPNWTNTQVRSSLQNTTTKLGDSFYYGKGLINVQAAAQ
Carboxpeptidase A (bovine)	ARSTNTFNYATYHTLDEIYDFMDLLVAQHPELVSKLQIGRSYEGRPIYVLKFSTGGSNRPAIWIDLGIHSRQWITQATGVWFAKKFTENYGQNPSFTAIL DSMDIFLEIVTNPNGFAFTHSENRLWRKTRSVTSSSLCVGVDANRNWDAGFGKAGASSSPCSETYHGKYANSEVEVKSIVDFVKNHGNFKAFLSIHSYSQ LLLYPYGYTTQSIPDKTELNQVAKSAVAALKSLYGTSYKYGSIITTIYQASGGSIDWSYNQGIKYSFTFELRDTGRYGFLLPASQIIPTAQETWLGVLTI MEHTVNN
Carboxpeptidase B (bovine)	TTGHSYEKYNNWETIEAWTEQVASENPDLISRSAIGTTFLGNTIYLLKVGKPGSNKPAVFMDCGFHAREWISPAFCQWFVRE EIHMTEFLDKL DFYVLPVVNIDGYIYTWTTNRMWRKTRSTRAGSSCTGTDLNRNFDAGWCSIGASNNPCSETYCGSAAESEKESKAVADFIRNHLSSIKAYLTIHSYSQMM LYPYSYDYKLPKNNVELNTLAKGAVKKLASLHGTTYSYGPGATTIYPASGGSDDWAYDQGIKYSFTFELRDKGRYGFVLPESQIQPTCEETMLAIKYVTS YVLEHL
Malate dehydrogenase (porcine)	

TABLE 10—(*continued*)

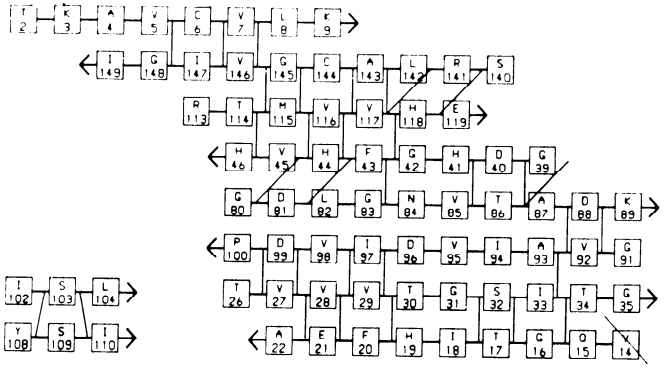
Protein	Symbolic secondary structure
Phosophoglycerate kinase (yeast)	
Phosphoglycerate kinase (horse)	
Hexokinase B III (yeast)	

See footnote to Table 7.

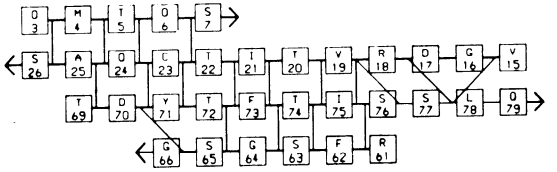

Lactate dehydrogenase (dogfish)
Lactate dehydrogenase-NAD (dogfish)
D-Glyceraldehyde-3-phosphate dehydrogenase (lobster) "green"
D-Glyceraldehyde-3-phosphate dehydrogenase (lobster) "red"
Alchohol dehydrogenase (horse)

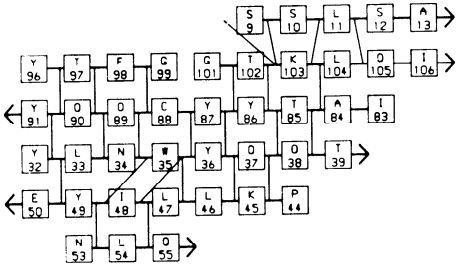


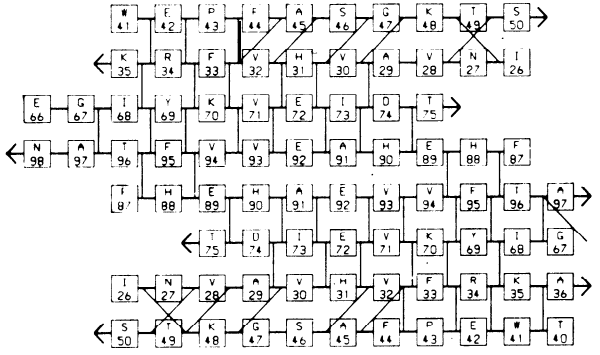
Calcium binding protein

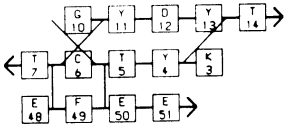


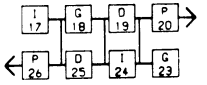
Superoxide dismutase



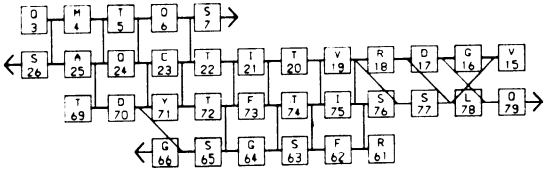


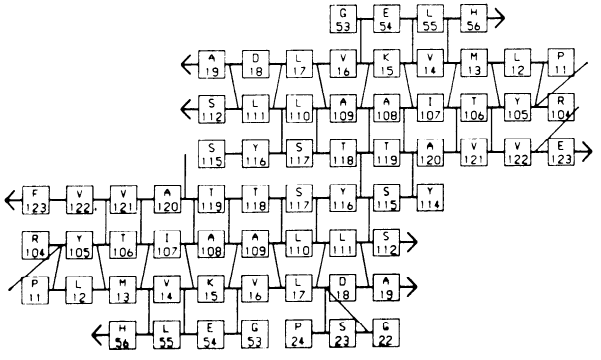




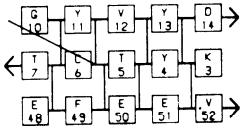


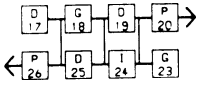
Rubredoxin (2Å)



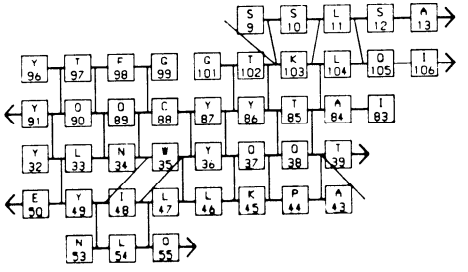


Prealbumin (dimer)





Rubredoxin (15Å)



Bence - Jones REI

The availability of several sets of independently determined co-ordinates for some of these proteins allows us to assess the reliability of these assignments. Table **11** shows the comparison of the derived secondary structure for proteins that are independent halves of a dimer in the same crystal, related forms determined by the same group, and the same or similar forms determined by different groups. The overall differences between the pairs of structures compared is 497 residues out of a total of 4069 residues (12·2%). If the short β-strands (less than 4 residues) are omitted from the derived assignments, the overall difference drops to 431 residues (10·6%). The difference for each section of the Table expressed as a percentage of the total number of residues in that section is 8·2%, 5·0% and 20·1%, respectively (6·7 %, 5·6% and 16·8% when the short β-strands are omitted). Much of the difference in the third section arises from the single comparison of carboxypeptidase A with carboxypeptidase B. When this comparison is omitted, the difference for the third section drops to 14·8% (12·2% without the short p-strands), and the overall difference drops to 9·6% (8·4% without the short p-strands). Although the proteins considered contain almost equal numbers of residues *in a-helix* and β-sheet (1437 and **1422** residues, respectively), there are almost twice as many differences for the β-sheet (324 residues compared with 173 for a-helix). When the short β-strands are omitted there are only 257 differences for β-sheets.

(c) *Computing requirements*

The computer programs used in this work are simple, compact and efficient. They assign the secondary structure of all the proteins considered here, output the tables of results, and plot the p-sheet diagrams in less than six minutes of central processor time on an IBM 370/165 computer (a typical core access and multiplication takes 2 μs). The programs are written in FORTRAN IV, run in 200K bytes of core storage, and are available from the authors.

4. Discussion

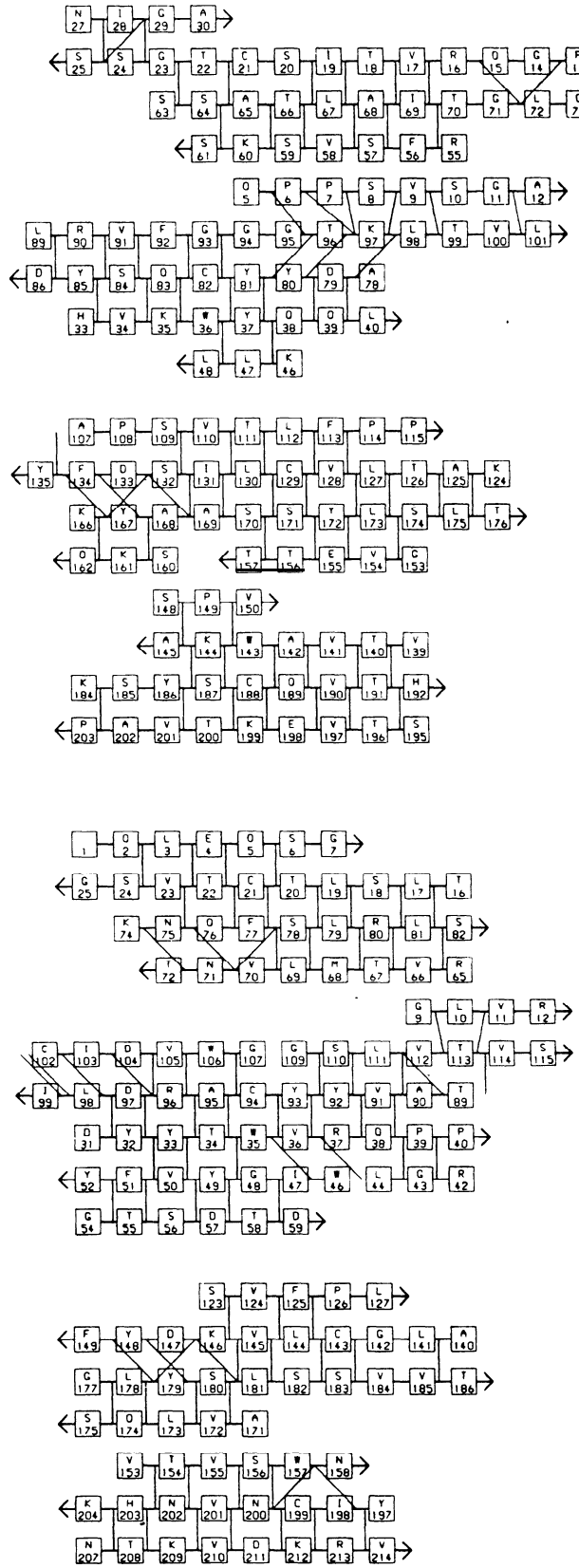
(a) *Definition of secondary structure*

There are basically two classes of definitions for these secondary structure elements : local and global. Local definitions of secondary structure involve the specification of a particular range of local residue conformations that are indicative of a-helix or

---

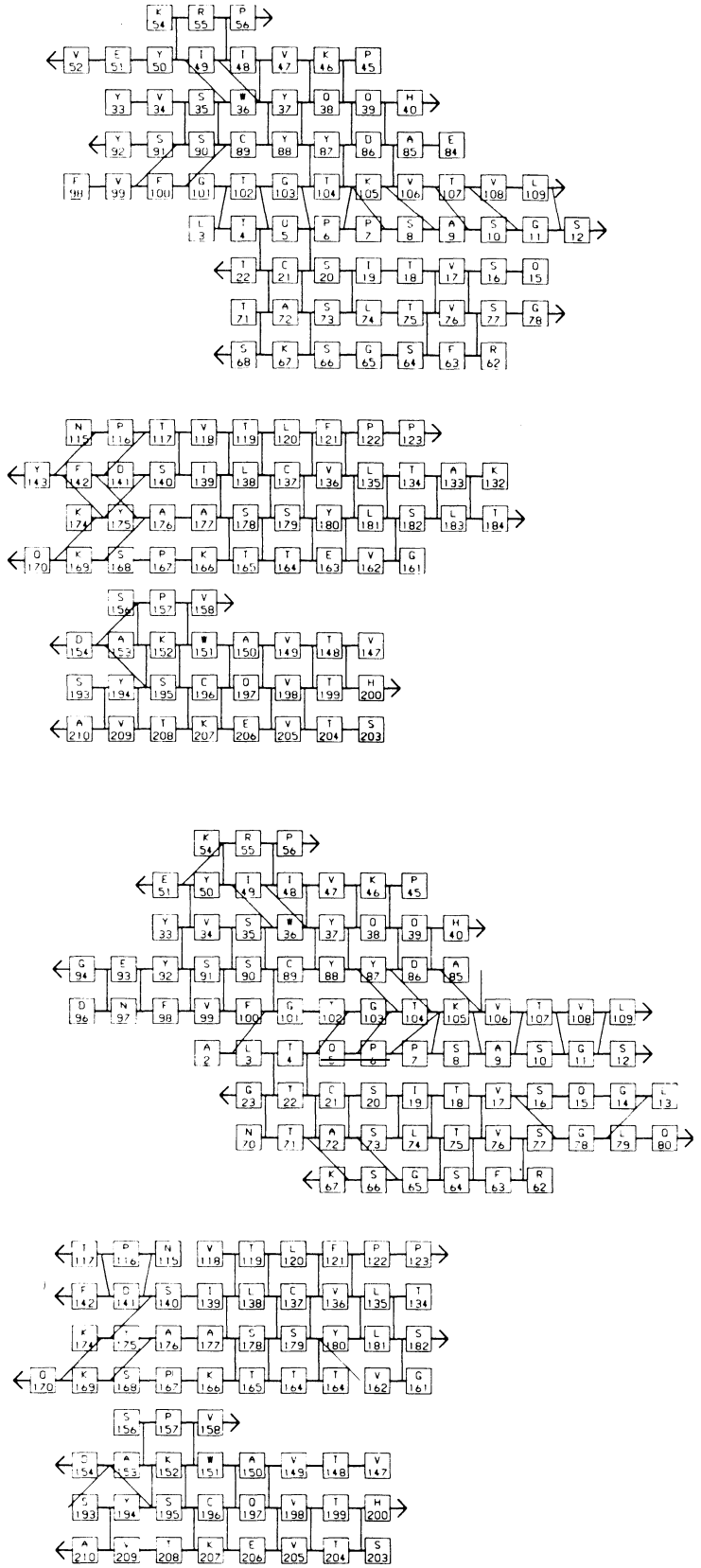
FIG. 5. This and the following Figures are schematic presentations of the β-sheet interactions in the proteins studied here. Each residue is marked as a square box that contains the residue number and the amino acid type according to the single letter code (A = Ala, C = Cys, D = Asp, E = Glu . F = Phe, G = Gly, H = His, I = Ile, K = Lys, L = Leu, M = Met, N = Asn, P = Pro, Q = Gln, R = Arg, S = Ser, T = Thr, V = Val, W = Trp, Y = Tyr). Adjacent residues along the sequence are connected by a heavy line with an arrow-head showing the chain direction. The inter-peptide hydrogen bonds are shown connecting the NH and CO groups involved. The NH group immediately precedes the residue, while the CO group immediately follows it along the chain.

The following /&strands found by the program and given in Tables 3 to 10 are spurious, in that the strands cannot be connected into the β-sheet. 177 = 182 in trypsin; 16 = 18, 188 = 193 in chymotrypsin (MRC) ; 34: 36 in ribonuclease A; 34: 36 in ribonuclease S; 74: 76 in lysozyme; 92:94 in papain; 25:27, 214:216 in malate dehydrogenase; 185 = 189 in lactate dehydrogenase; 338:340 in alcohol dehydrogenase ; and 296-298 in hexokinase. In each case the symbol between the first and last residue of the strand is = if derived by the C^{α}—C^{α} method alone, : if derived by the H-bond method alone, and - if derived by both methods.



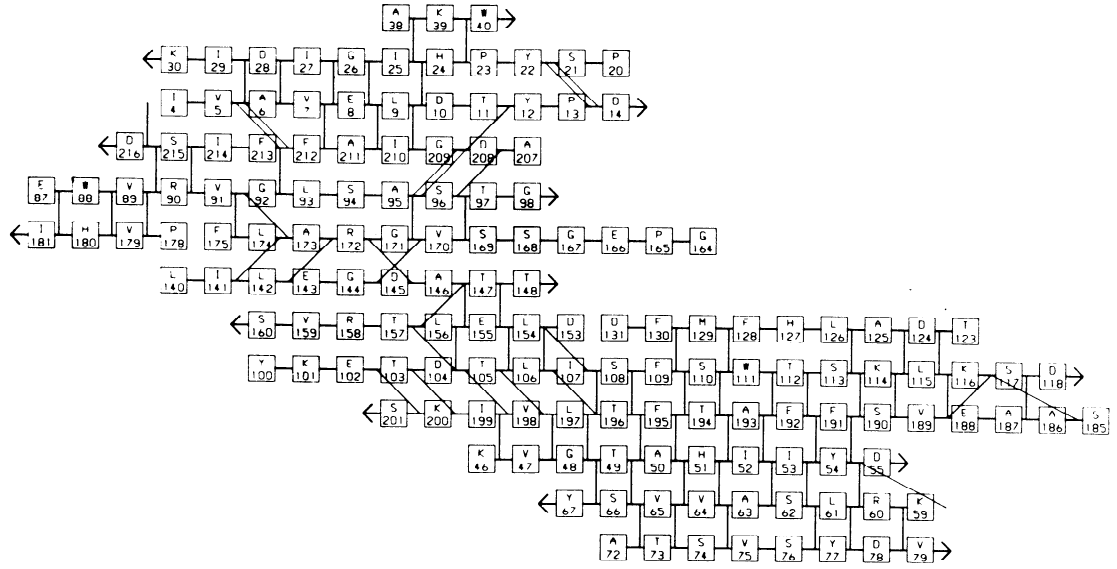
Immunoglobin G **Fab'**

FIG. 6

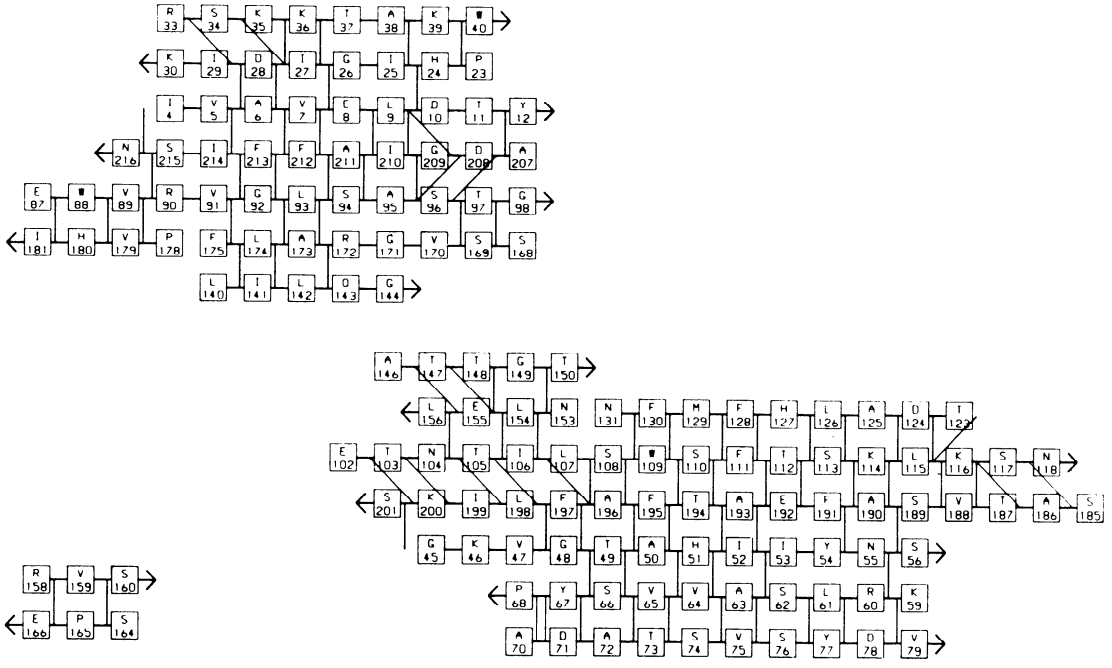


Bence – Jones dimer Mcg

FIG. 7



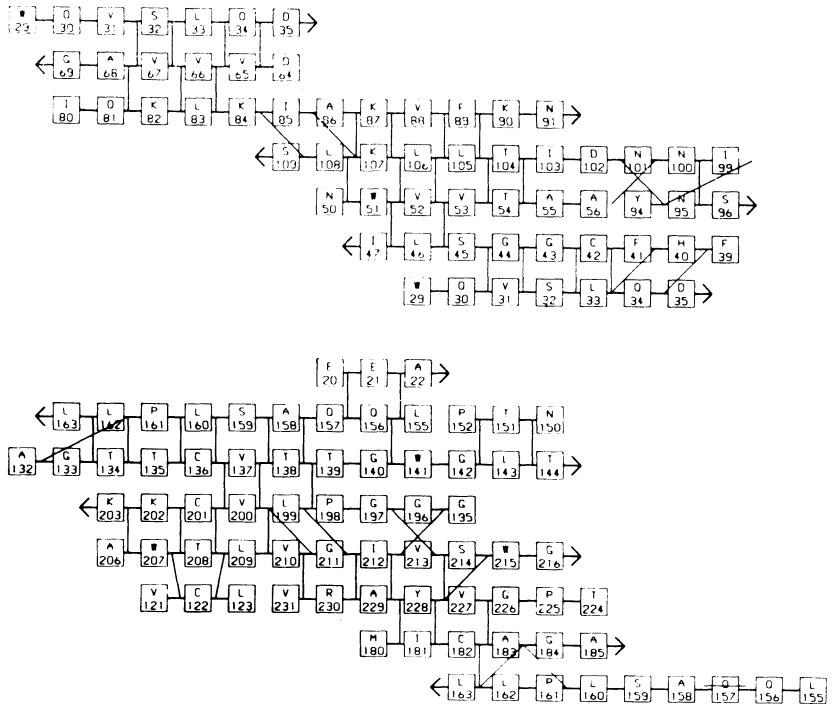
Concanavalin A (Argonne)



Concanavalin A (Rockerfeller)

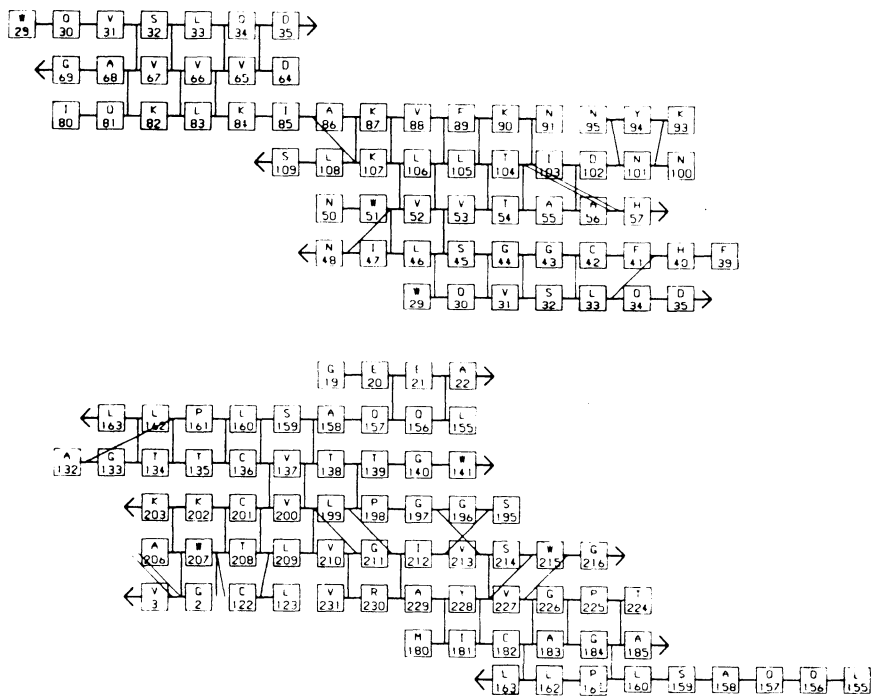
FIG. 8

β-sheet. Such a definition turns out to be quite poor, however, as the conformation of a residue (specified, for example, by the (ϕ, ψ) torsion angles) may vary considerably within a region of secondary structure that is not perfectly regular. This is particularly true of β-sheets, which show a greater variability in local residue conformation than do a-helices. In addition, the (ϕ, ψ) angles can be subject to considerable experimental error, as they depend very sensitively on the precise orientation of the peptide group, which cannot always be accurately positioned in the electron density. Some improvement can be achieved by examining the local conformations



α - Chymotrypsin (MRC)

(a)



α- Chymotrypsinogen (Michigan)

FIG. 9

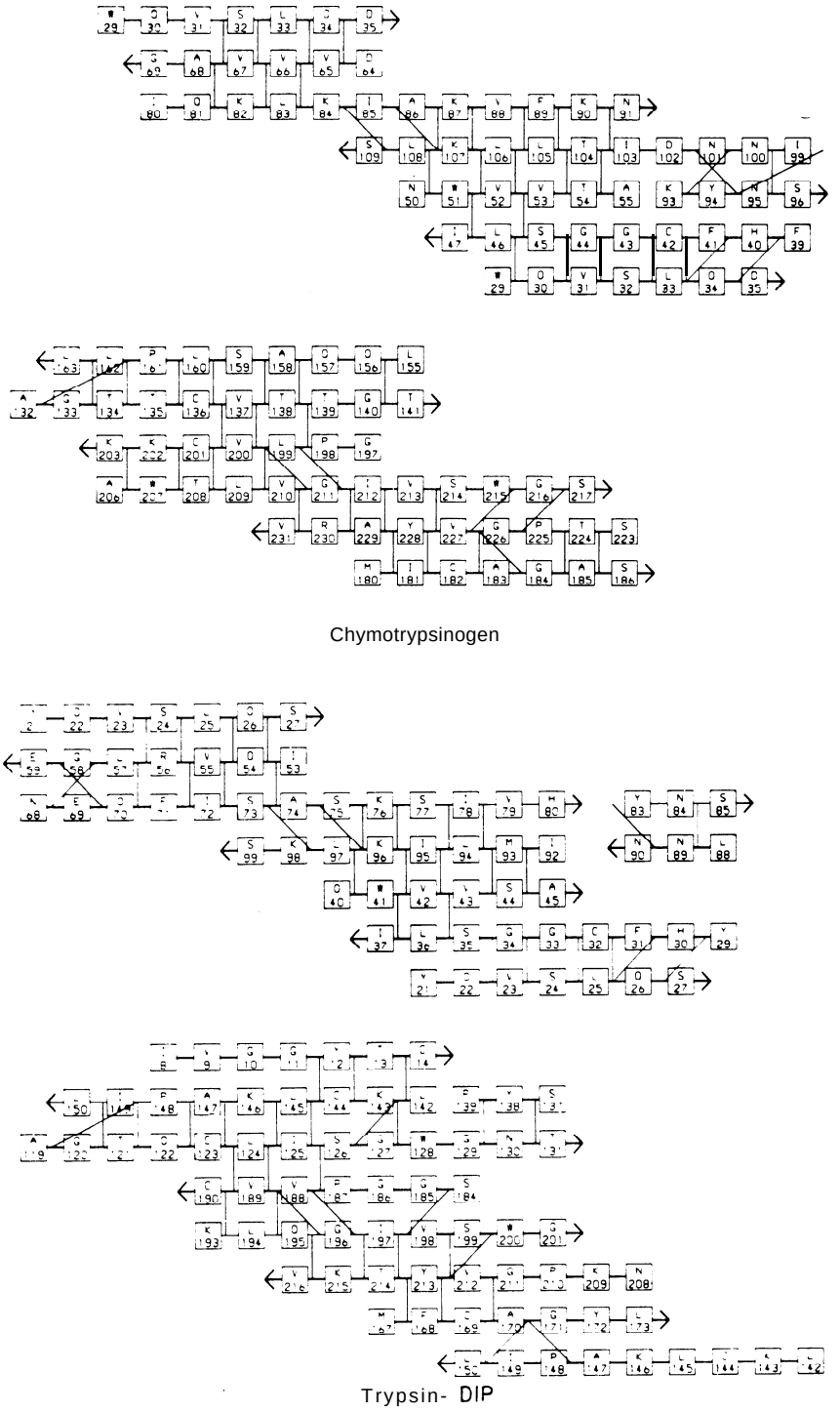


FIG. 9

of several consecutive residues together, in order to "average out" irregularities and errors.

Global definitions of secondary structure involve the specification of the patterns of interactions indicative of different types of secondary structure. The most obvious interactions are the hydrogen bonds between pairs of residues often well-separated in sequence. The basic requirement of such a definition is an agreement on the acceptable geometry of these hydrogen bonds and the number of hydrogen bonds necessary to specify a particular secondary structure region. Difficulties arise when

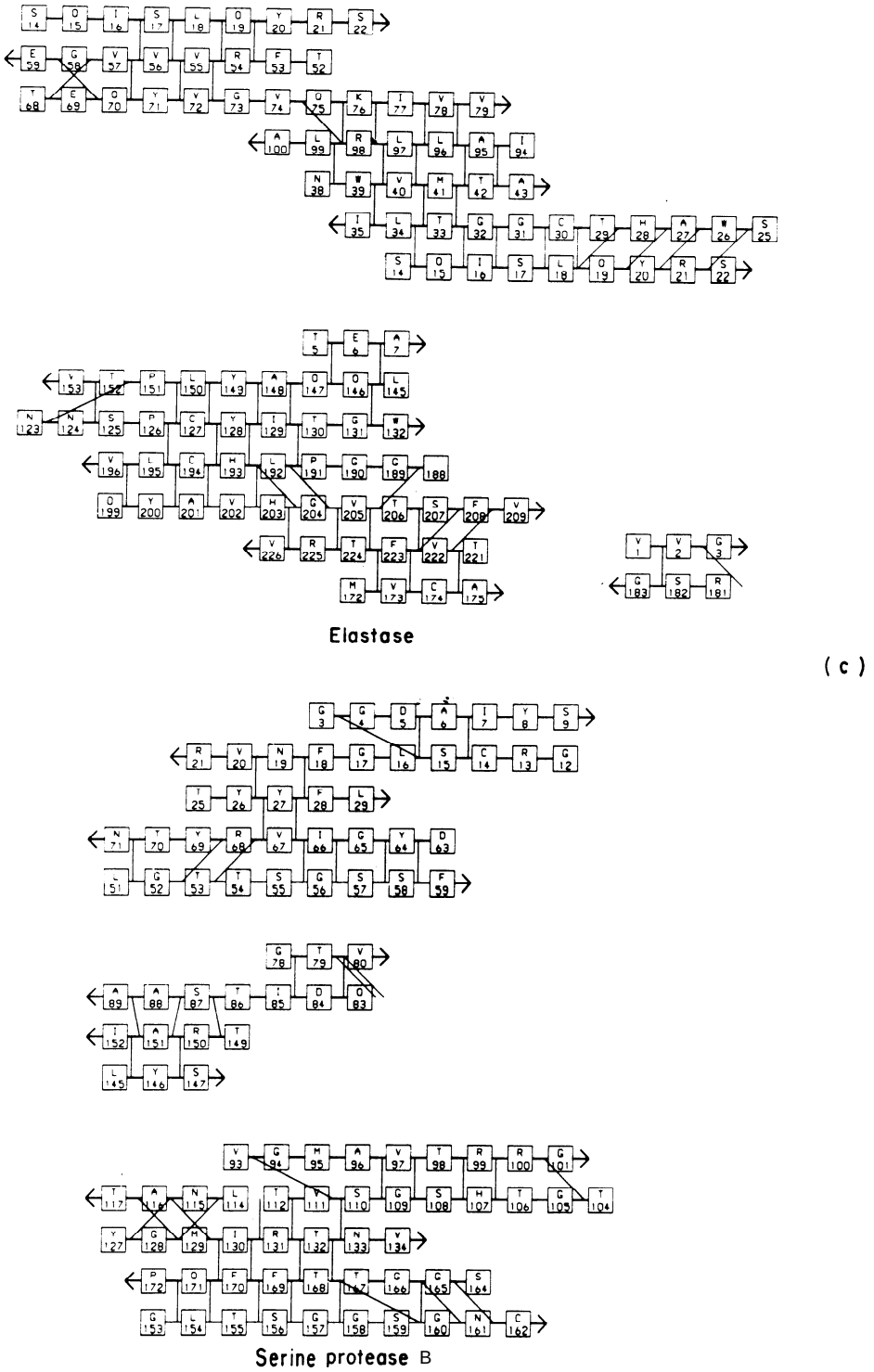
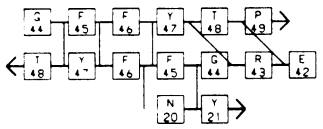


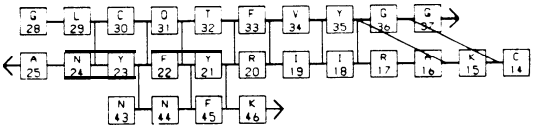
FIG. 9

trying to find criteria for acceptable hydrogen bonds that are broad enough to mitigate co-ordinate errors, yet do not lead to extra spurious bonds (see Levitt, 1974). Nevertheless, as one is searching for patterns formed by several hydrogen bonds, an occasional spurious bond is not a great problem. Lists of hydrogen bonds are usually only available once the co-ordinates have been refined.

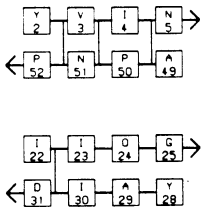
The a-angle method proposed in this work corresponds to the local definition of secondary structure, whereas both the H-bond and C^{α}—C^{α} methods are specific forms of global definitions. All the methods operate on a-carbon co-ordinates, which



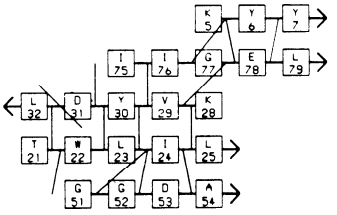
Insulin



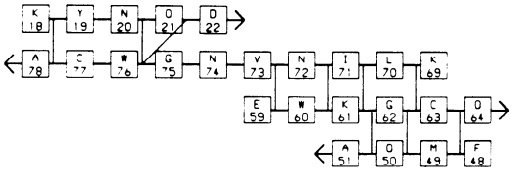
Trypsin inhibitor



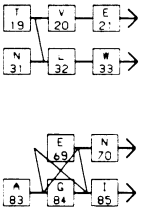
Ferredoxin



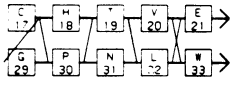
Ferricytochrome *b* 5



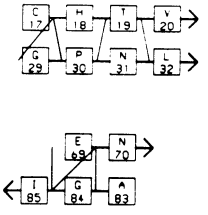
High potential iron yrotein



Ferrocytochrome *c* (tuna)



Ferricytochrome *c* (outer)



Ferricytochrome *c* (inner)

FIG. 10

are usually the most accurate and generally available at an early stage of electron density map interpretation.

Each of the above methods has its own particular advantages and disadvantages. The a-angle depends on the positions of four adjacent C^{α} atoms, and it is much less subject to experimental error than the conventional (ϕ, ψ) angles. Here, the main use of the a-angle method is to locate left-handed and right-handed reverse turns. The method also usefully recognizes those a-helices too distorted to have a clear pattern of hydrogen bonds, yet for which each of several consecutive residues has the helical local conformation. This method is the only one of the three tested here that can identify extended chain not involved in β-sheet interactions, or be sensitive to the handedness of the data and recognize left-handed a-helices. The major weakness of the method stems from its local nature, which means that it cannot recognize short β-strands, find the ends of a-helices and β-strands precisely, or distinguish between #I-strands and extended chain not involved in β-sheet interactions.

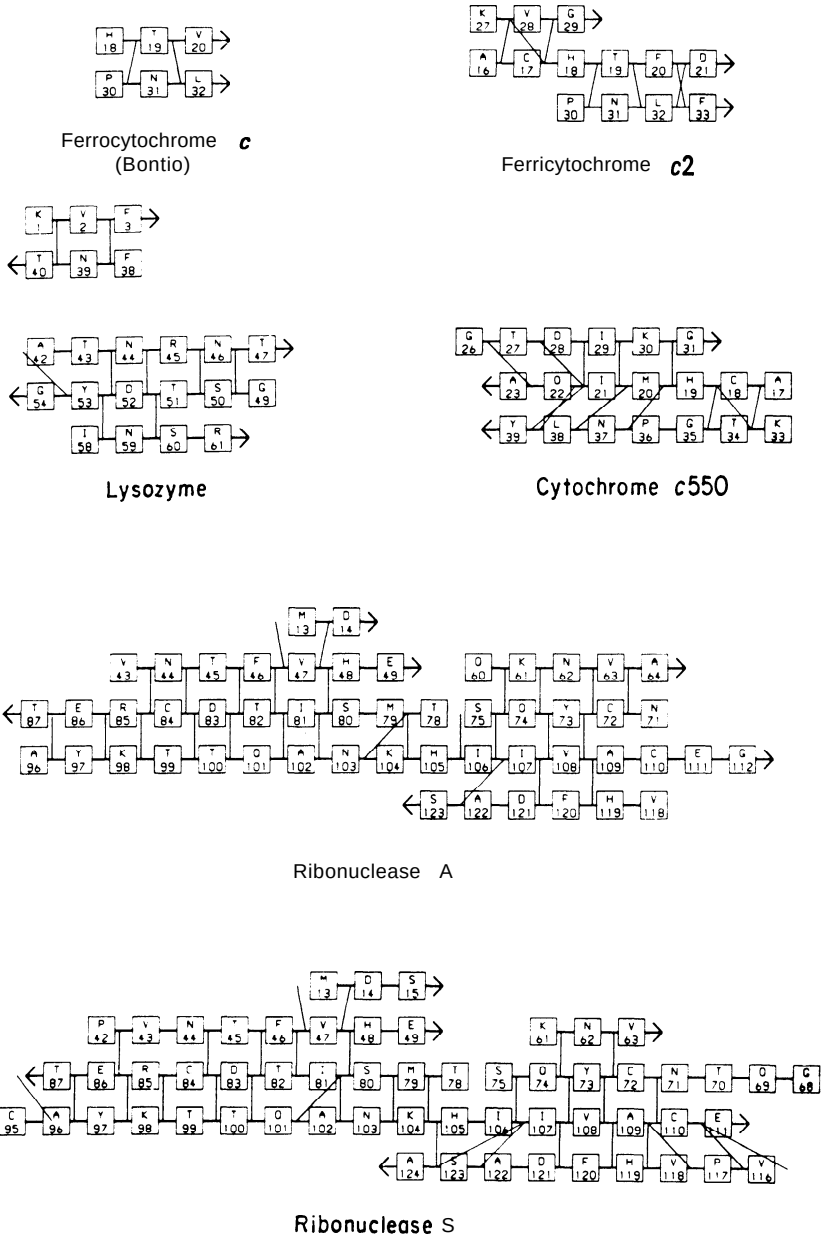
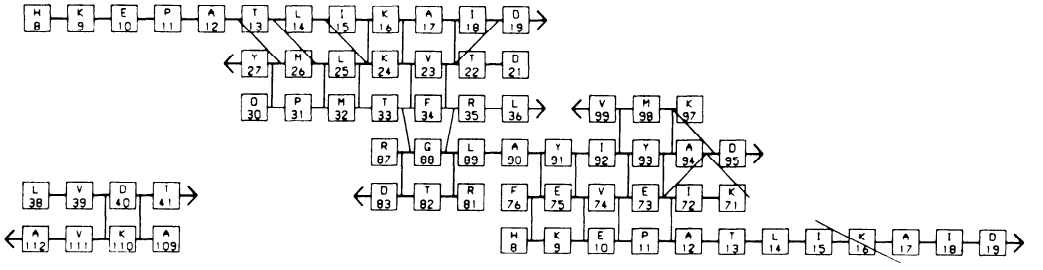
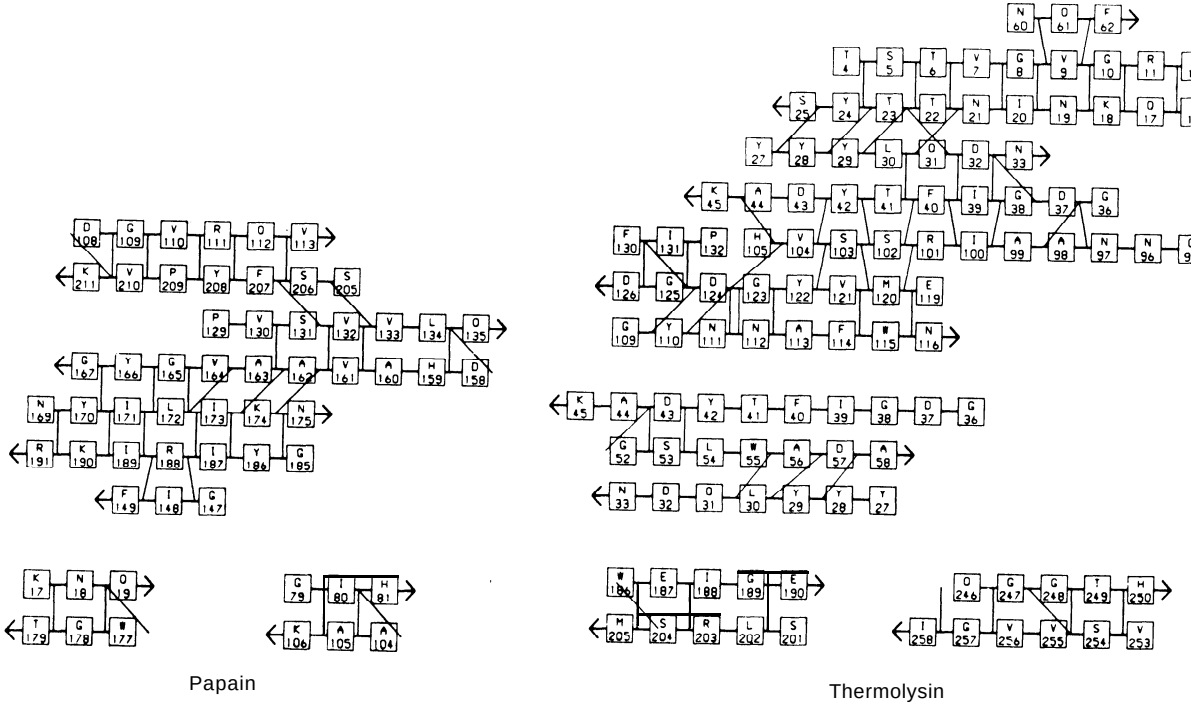


FIG. 11

The C^{α}—C^{α} technique uses inter-C^{α} distances just as does a contact map. It recognizes α-helices less well than the other methods, because the criterion of inter-C^{α} distance is too rigid to include either 3_{10} or distorted helices. Attempts to relax these rules, for example by increasing the' permitted distances beyond 6·0 and 6·5 A, have not succeeded, because then the criteria are insufficiently selective and will wrongly identify a highly twisted region of the chains, such as a series of turns one after the other, as a-helix. Problems also exist for very short helices, which are difficult to distinguish from turns. The introduction of a test to fit the derived short helices to an ideal a-helix does improve detection of such short α-helices. When the C^{α}—C^{α} method is applied to β-sheet detection, it selects those parts of the protein structure where the chains run approximately parallel (or antiparallel). True β-strands are distinguished by selecting for the expected, reasonable regular geometry between a-carbons in the pair of p-strands. The parameter most reflective of β-sheet structure,



Nuclease (S. *aureus*)



Papain

Thermolysin

FIG. 12

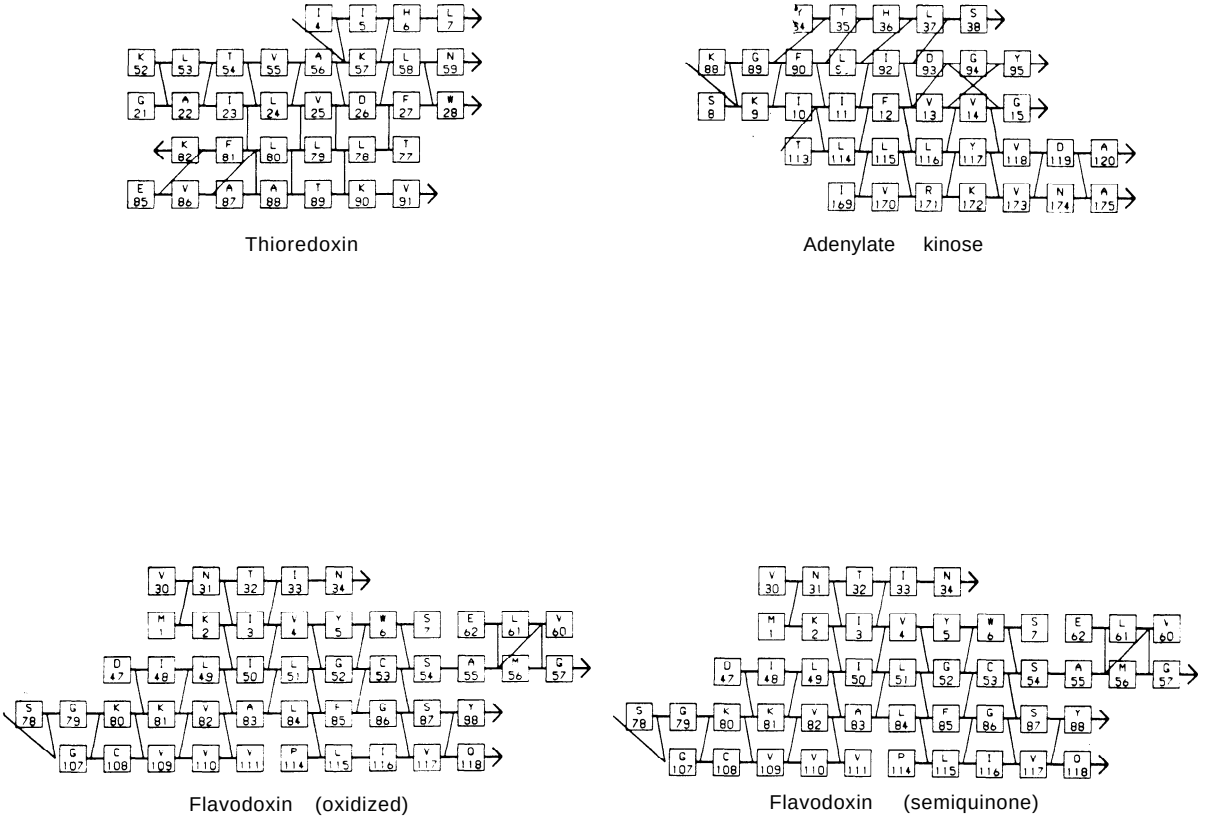
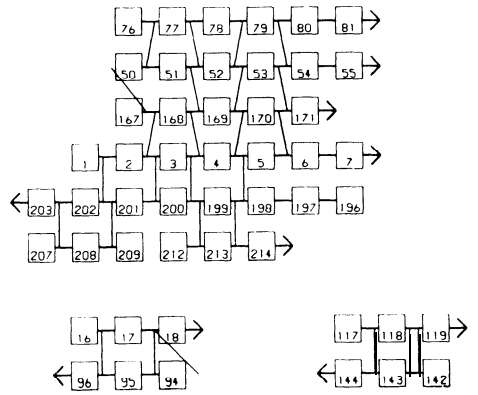


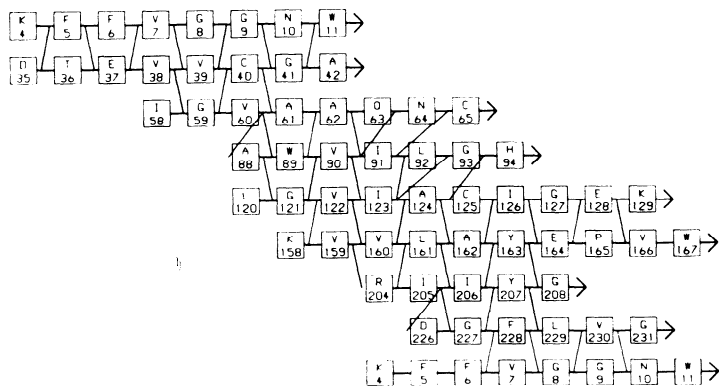
FIG. 13

at least as compared to the usual reported values, turns out to be the dihedral angle ρ, which describes the twist of the sheet out of the plane formed by the two strands. When this angle was restricted to the range $-10°$ to $40°$, excellent agreement with reported β-sheet values was achieved. However, the demand for this regularity in twist angle causes some short and irregular strands to be rejected. This method therefore recognizes well-formed β-sheet and is somewhat conservative in its treatment of shorter, more irregular strands.

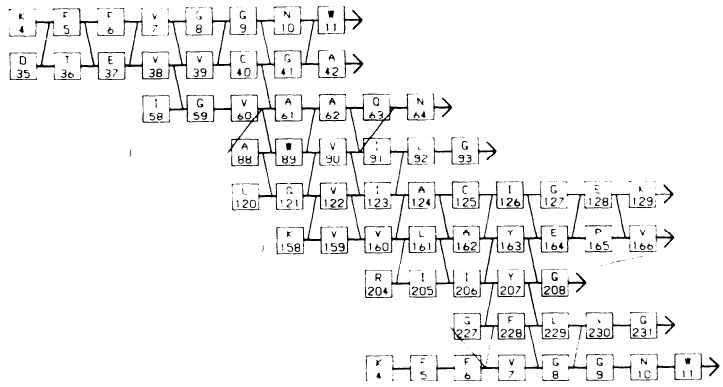
The H-bond method is based on the detection of the stabilizing electrostatic interactions between pairs of residues in globular proteins. These interactions, the hydrogen bonds between peptide groups, define the geometry of the α-helix and p-sheet secondary structure very precisely. As the criteria used here require that both the peptide–peptide distance and relative orientation be acceptable, it is possible to allow generous distance and orientation limits, yet still find relatively few good interactions. The subsequent analysis of the pattern formed by these hydrogen bonds further improves the selectivity of the method only considering as real, occurrences of two or more hydrogen bonds in the same region of chain. Thus the H-bond method is both very sensitive and yet strongly selective. It can detect an a-helix held together by a pair of hydrogen bonds and distinguish. a short, irregular β-sheet with a pair of hydrogen bonds from strands that are merely parallel or antiparallel. Such interactions could easily be missed by observation of a model alone. The method does have two small weaknesses, both of which are nicely corrected by each of the other two methods. Helices too irregular to have good hydrogen bonds will be missed by the H-bond method but these helices are often detected by the or-angle method. Sheets with an internal irregularity will be broken into smaller pieces by the H-bond



Phosphoglycerate **mutase**

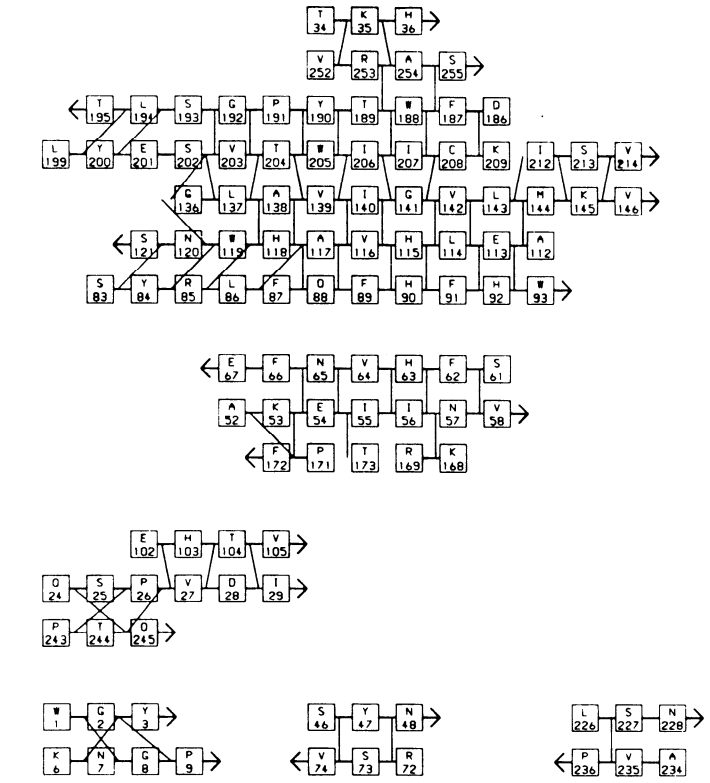


Triose phosphate isomerase (no. I)

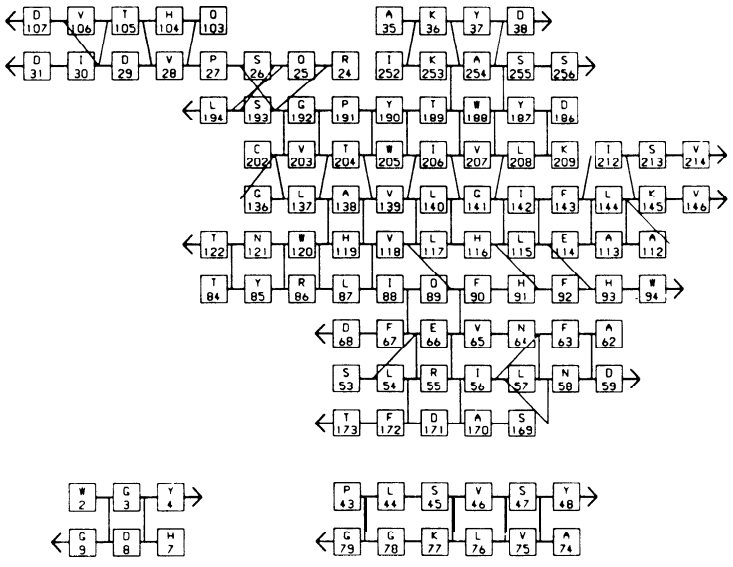


Triose phosphate **isomerase (no. 2)**

FIG. 14



Carbonic anhydrase B



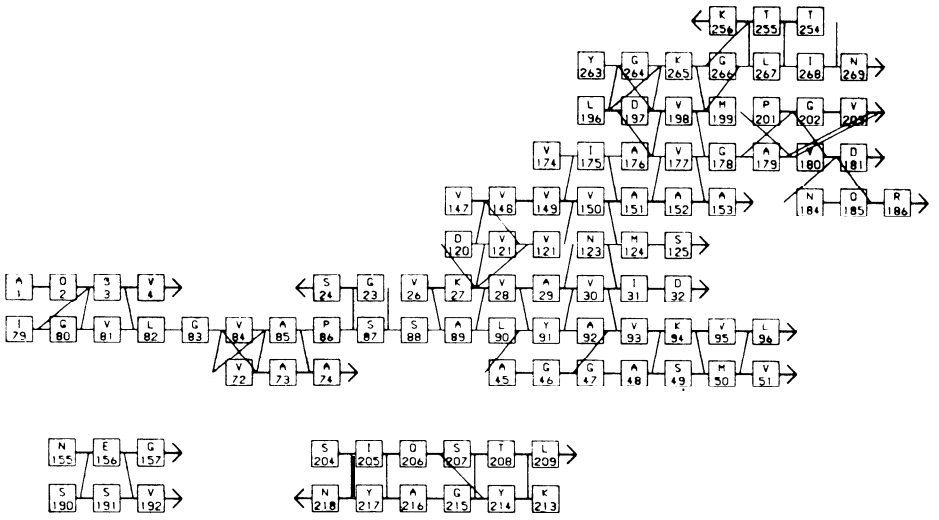
Carbonic anhydrase C

FIG. 15

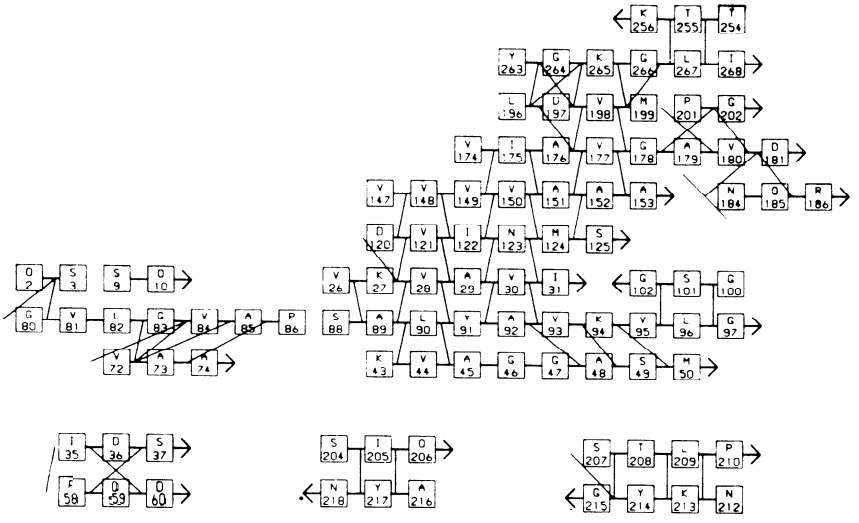
methods but these sheets are often joined by the C^{α}—C^{α} method, allowing the program to produce clear p-sheet diagrams automatically.

(b) *Comparison of reported and derived assignments*

Most of the reported α-helices and β-strands (Table 2) are found by the automatic procedure used here. With just a few exceptions, all the α-helices and p-strands added or missed by the procedure are very short. We have checked all these cases by



Subtilisin BPN'



Subtilisin novo

FIG. 16

inspection of atomic models and stereo drawings of the particular protein, and conclude that the assignments derived by the method are a reasonable reflection of the conformation described by the C^{α} co-ordinates we have used. We do not feel that a detailed comparison of the derived results with the reported values would be fair to the crystallographers who have so kindly provided us with their co-ordinate data. Often the secondary structure is reported at an early stage before the co-ordinates have been accurately refined. In other cases, the crystallographers do not actually delineate the secondary structure precisely, but give a list of hydrogen bonds. In these cases, such lists must be interpreted by the reader to give regions of secondary structure. This uncertainty is reflected in the fact that the reported secondary structure quoted as the "X-ray data" by groups. studying the correlation of sequence and secondary structure are often different (Crawford et *al.,* 1973 ; Chou & Fasman, 1974; Lim, 1974): for several proteins the percentage of amino acids with different

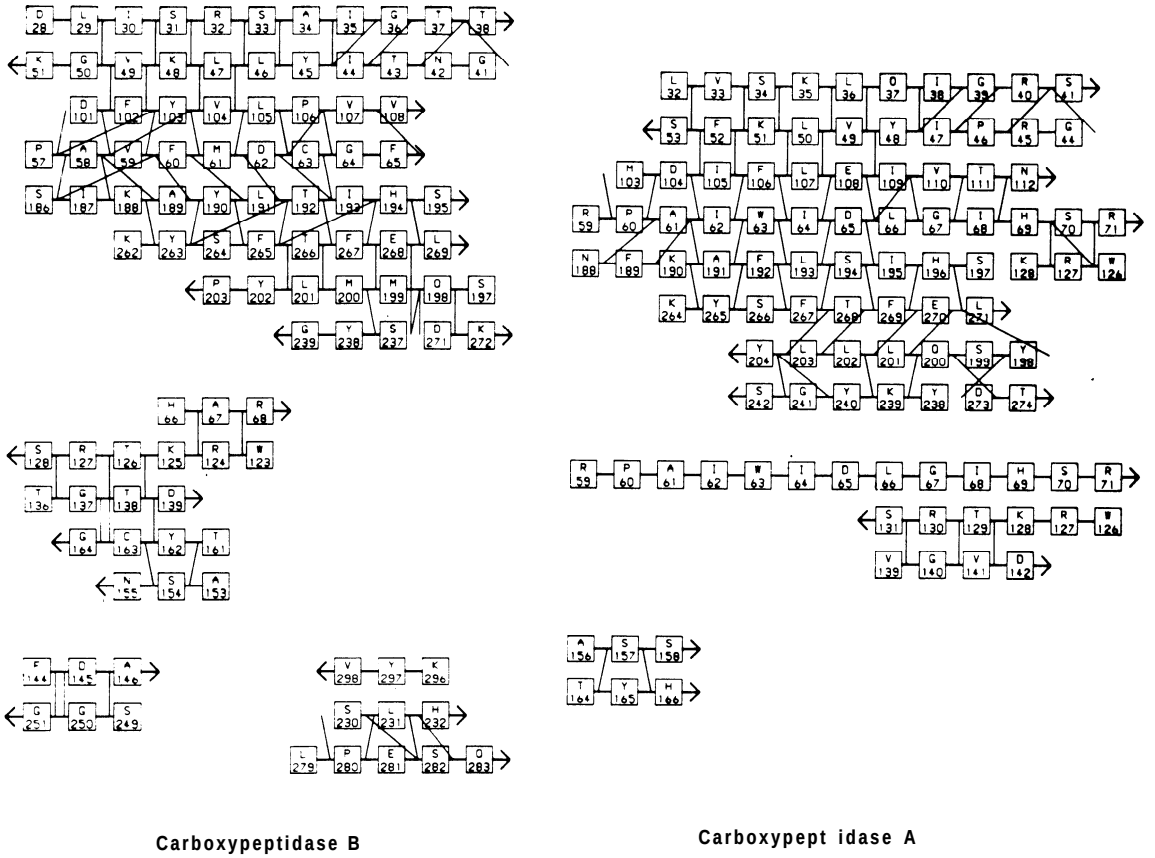


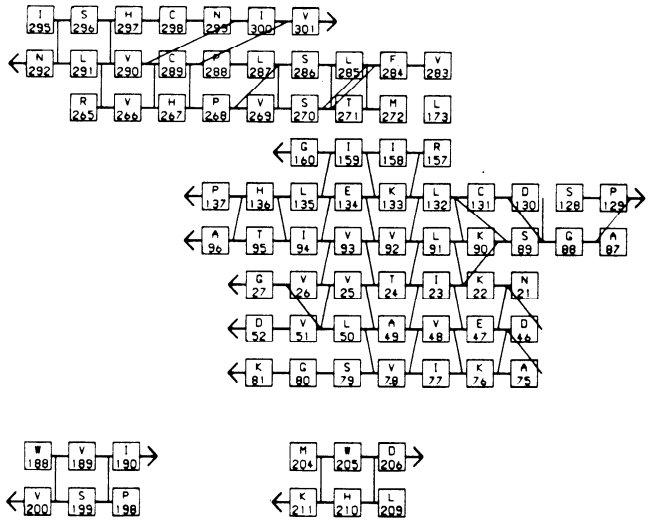
FIG. 17

reported assignments is as high as 15%, while for others all three groups use the same secondary structure assignments.

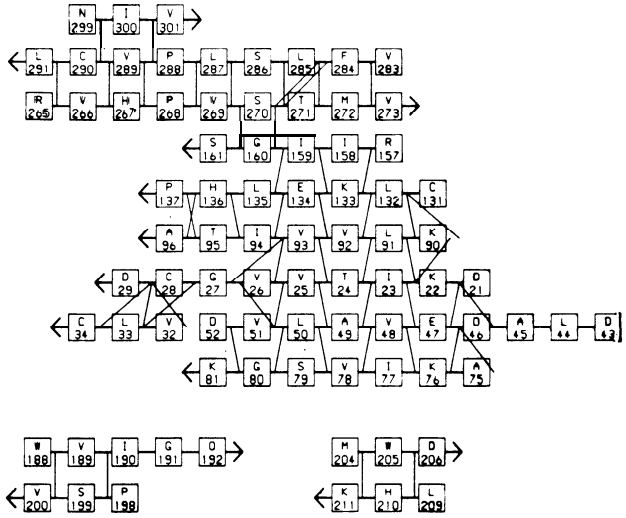
The overall difference of 1869 residues between the derived and reported secondary structure of the proteins shown in Table 2 amounts to an average difference of only 1·5 residues at both ends of each piece of secondary structure (there are 649 reported pieces). This difference is very small when one considers that at present protein crystallographers do not have a generally accepted definition of what constitutes secondary structure. In particular, there is no definition of the range of local residue conformations acceptable in an a-helix or β-strand, of what constitutes acceptable pcptide hydrogen bond geometry, of the minimum number of hydrogen bonds needed to define an a-helix,or β-sheet, or of the precise position of the ends of the regions of secondary structure in relation to the acceptable local conformations and hydrogen bonds.

(c) Reliability of the derived secondary structure assignments

We have examined how sensitive the derived assignments are to co-ordinate errors using the different sets of co-ordinates available for quite a few of the proteins analysed here (see Table 11). The accuracy of the derived assignments is calculated as half the total number of differences between the two assignments for each protein, expressed as a percentage of the number of residues found in *a-helix* and β-sheet (half the total difference is the difference of each assignment from the average assignment). The average accuracy of assignments made on similar proteins determined by the same group of workers whether as dimers in the same crystal or as independent

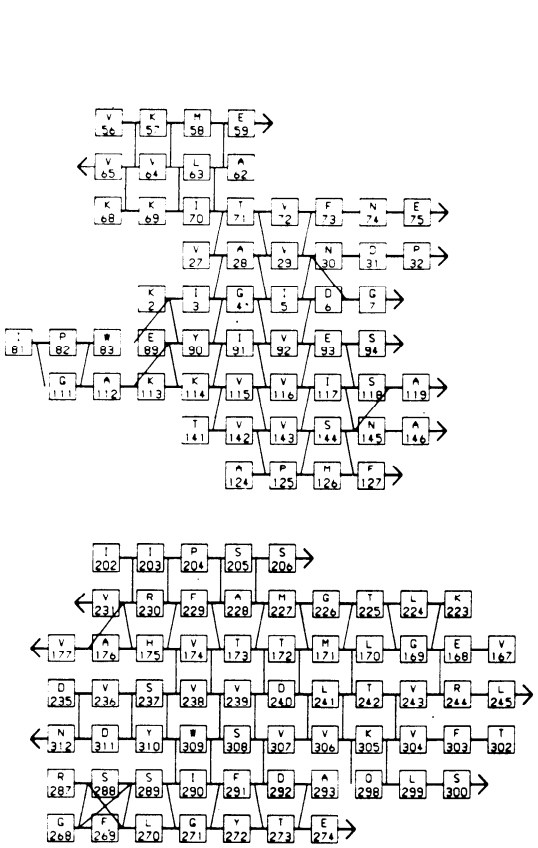


Lactate dehydiogenase-NAD

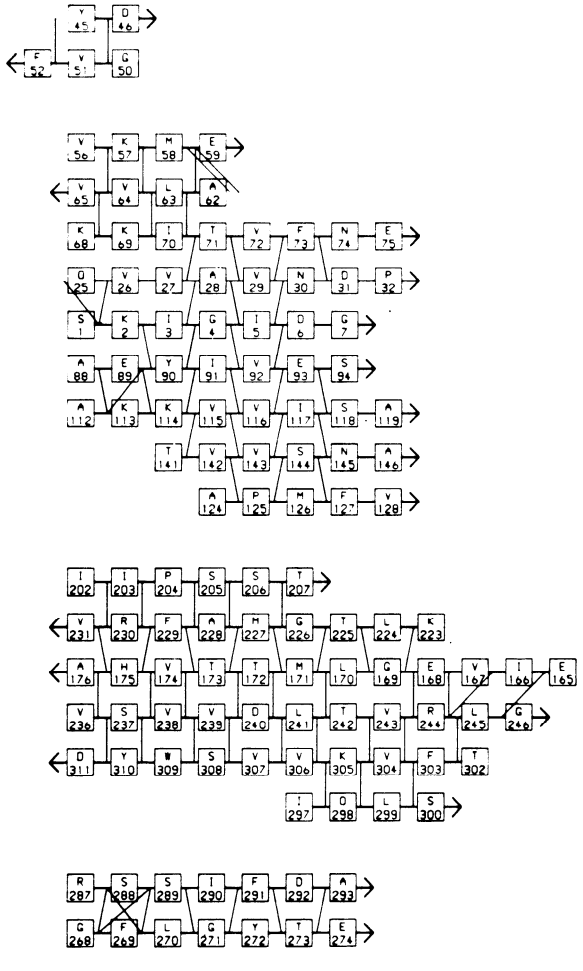


Lactate dehydrogenase

(a)



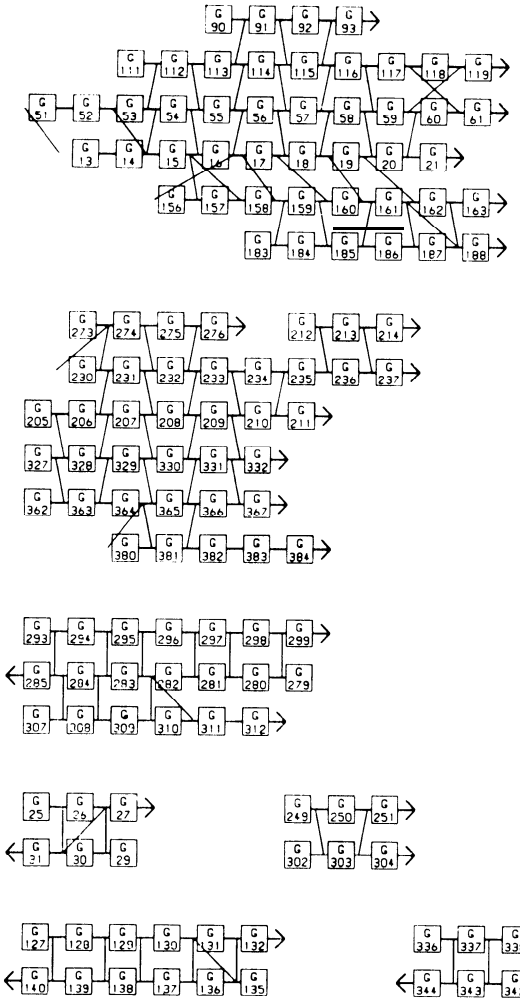
D-Glyceraldehyde-3-phosphate dehydrogenose "green"



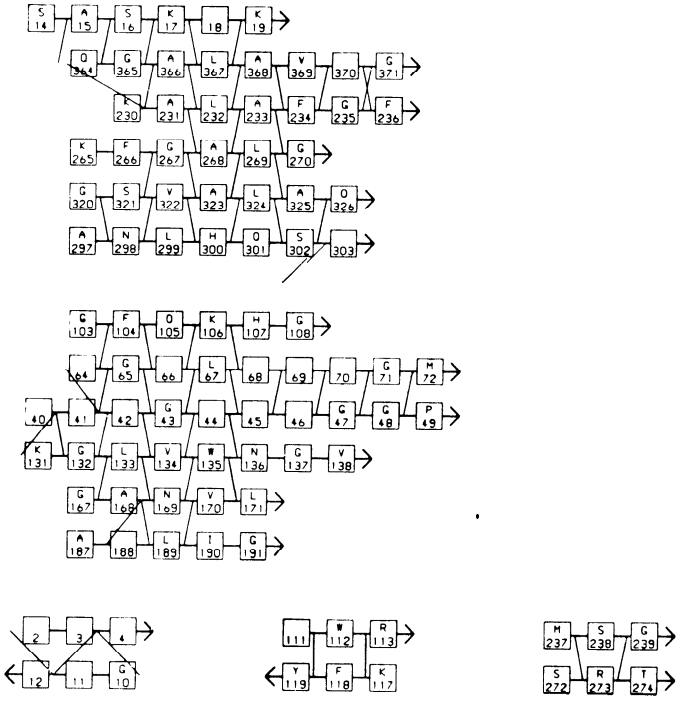
"red"

(b)

FIG. 18

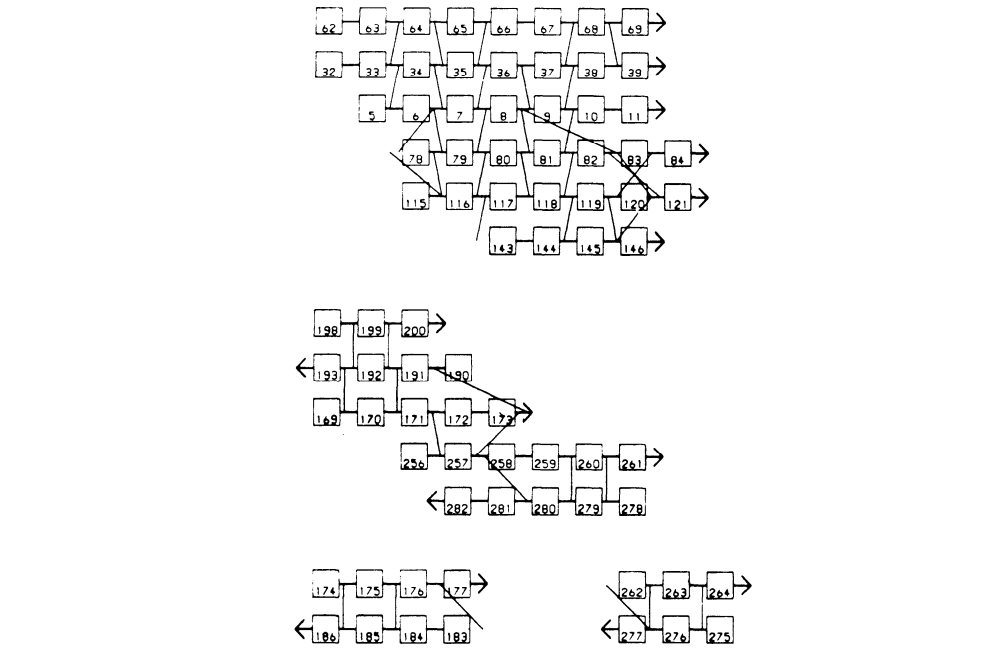


Phosphoglycerate kinase (horse)

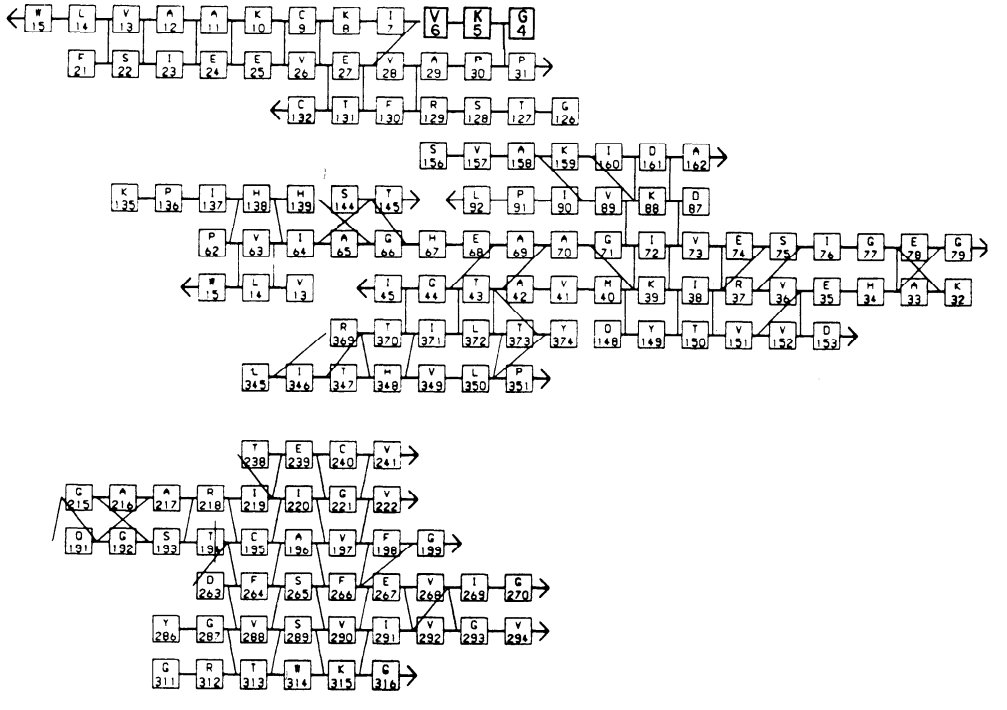


Phosphoglycerate kinase (yeast)

FIG. 19



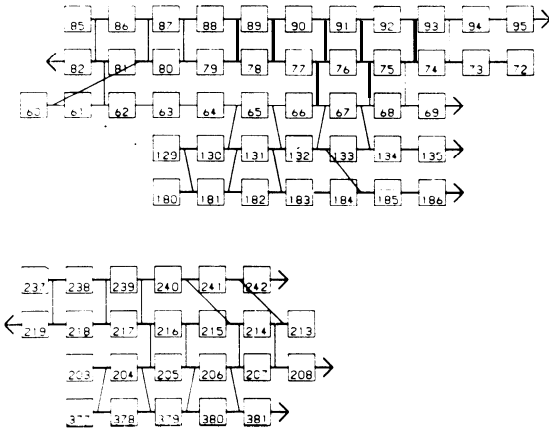
Malate dehydrogenase



Alcohol dehydrogenase

FIG. 20

crystal structures, is 3·5% for α-helices and 6·2% for p-strands. The accuracy of assignments made on similar or related proteins determined by different groups is lower (14·8% and 15·8% for α-helix and /I-sheet, respectively), indicating both the real differences in structure and differences due to interpretation of the density map. Closer analysis reveals that the α-helices and long regular p-strands are quite well-reproduced in these related data sets, and that it is mainly the irregular strands that cause the increased differences between the β-sheet assignments. This is entirely



Hexokinase

FIG. 21

TABLE 11

Comparison of the derived secondary structure in related proteins

Proteins compared	Total residues	α-Helix Total	Extra	Missed	β-Sheet Total	Extra	Missed
Dimers in crystal							
Bence-Jones REI	107	0	0	0	76	1	0
Bence-Jones Mcg	215	14	0	0	143	19	0
Prealbumin	114	6	0	0	75	8	1(3)
Insulin	51	31	8	0	7	5	1(2)
Ferricytochrome c	103	49	6	0	11	7	2(2,3)
Triose phosphate isomerase	247	128	12	0	60	4	0
D-glyceraldehyde-3-phosphate dehydrogenase	333	100	5	0	124	21	2(2 x 3)
Total for section	1170	328	31	0	496	65	16
Different forms from the same group							
Hemerythrin/azomyo-hemerythrin	15	92	11	0	0	0	0
Horse hemoglobin, aquomet/deoxy	287	246	8	0	0	0	0
Deoxyhemoglobin, horse/human	287	246	8	0	0	0	0
Rubredoxin, 2·0 Å/1·5 Å	54	0	0	0	22	1	0
Flavodoxin, oxidized/semiquinone	138	62	3	0	47	0	0
Lactate dehydrogenase Apo/NAD complex	329	139	16	0	84	14	2(2×3)
Total for section	1210	785	46	0	153	15	6
Same or similar protein by different groups							
Concanavalin A Argonne/Rockerfeller	237	6	0	0	153	22	0
Chymotrypsin, MRC/Michigan	245	21	8	1(6)	124	34	4(3 x 3,6)
Chymotrypsin, Michigan/ogen	245	26	14	2(2 x 6)	119	17	2(3,4)
Ribonuclease, S/A	124	28	1	0	61	17	1(3)
Carbonic anhydrase B/C	256	40	7	0	114	45	3(3 x 3)
Subtilisin BPN′/novo	275	86	6	1(6)	102	34	7(7 x 3)
Carboxypeptidase, A/B	307	117	60	2(5,6)	100	75	7(6 x 3,5)
Total for section	1689	324	96	35	773	244	78

expected, as minor changes in the co-ordinates may cause the program to accept or reject somewhat different regions of the irregular β-structure. Once specific rules for choosing β-sheet are developed, borderline cases will always exist.

When we tested the methods on co-ordinates that had been energy refined (Levitt, 1974), the derived secondary structures differed only slightly from those derived from unrefined co-ordinates. It is of interest that for lysozyme the number of hydrogen bonds found by a detailed all-atom criterion changed considerably after energy refinement (see Levitt, 1974), indicating that the present criteria for simplified hydrogen bonds are much less sensitive to small atomic shifts (about 0·3 Å).

5. Conclusion

We have made precise rules for assigning secondary structure in proteins and then used an objective automatic procedure to analyse the secondary structure of very many different globular proteins. Our derived assignments differ significantly from those reported in the literature. These differences are not a result of measurement errors in the co-ordinates we have used, as the method is insensitive to small atomic shifts, and also gives very similar assignments for proteins for which several sets of co-ordinates are available. These differences do not seem to be a result of intrinsic limitations of our definitions, simplifications or computer programs. No simple change of the definitions (like moving the end-points of regions of secondary structure by one residue) improves the agreement with the reported assignments. The simplifications are based on careful analysis of polypeptide geometry; they offer several advantages over the detailed all-atom representation, as the C^{α} co-ordinates are most accurate, the *α-angle* is less sensitive to errors than are the conventional (ϕ, ψ) angles, and the number of simplified hydrogen bonds is less sensitive to co-ordinate changes than the number of detailed hydrogen bonds. The computer programs are sophisticated, have been tested extensively, and agree well with manual assignments made on the same data. In all *cases* where a whole a-helix or β-strand has been added or missed by the method, we have checked the co-ordinates and find the automatic assignments to be reasonable.

The reasons for the differences between the reported and derived assignments seem to be the use of different definitions of secondary structure, the justifiable reluctance to report the less certain regions of secondary structure, and the fact that the co-ordinates we have obtained from the crystallographers are often more refined than the model used to make the original reported assignments. The rules used by our procedure work well and could form the basis for a standard recognition of secondary structure.

Although the rules of secondary structure assignment used by the different procedures have been given above (Methods), we will summarize the main features of the rules used by the preferred combination of procedures. α-*Helix* is assigned to residues n to m inclusive when every peptide group between these residues (peptide groups $n + 1$ to m) makes a good hydrogen bond to another peptide group separated from it by three residues. The shortest a-helix is required to have two hydrogen bonds and, therefore, consists of five residues. Distorted a-helices that are not recognized by this criterion are assigned for residues n to m, when $m - n \geq 5$ and α_i is in the range 0° to 120° for all values of i from $n + 1$ to m, inclusive. α-Helices discovered by the latter criterion are not assigned until residues have been assigned to

α-helix by the first criterion and to β-sheet by the following criterion. *β-Sheet* is assigned to residues n to m inclusive when every peptide group between these residues (peptide groups $n+1$ to m) either makes a good hydrogen bond to another strand parallel or antiparallel to the first strand or has C^{α} geometry that gives the strand separation and twist expected of β-sheet. The hydrogen bonds to the other strand are counted only when at least two adjacent hydrogen bonds are in correct register for parallel or antiparallel sheets. The shortest β-strand consists of three residues connected by two hydrogen bonds to another β-strand. *Reverse turns* are assigned to residues m to n inclusive when they are not part of a-helix or β-sheet secondary structure assigned previously, and the C^{α} backbone bends sharply at each of these residues. More precisely, α_i must be in the range $-90°$ to $90°$ for all values of i from $n+1$ to m, inclusive. The shortest turn can consist of a single residue that is between two regions of secondary structure.

Much effort has been devoted to the prediction of secondary structure from amino acid sequence (cf. Schulz *et al*., 1974*b*). Preparing probability tables for helical or sheet tendencies of a particular amino acid depends on secondary structure assignments used on the sample sets of proteins. Even though different assignment criteria for different proteins will average out if a large enough array of proteins is used, systematic differences will affect the probabilities as a whole. Even more critical is the importance of secondary structure assignments in a protein for which predictions are being made. If subjective criteria are used to assign secondary structure, then the reliability of the comparison of predicted to reported values must also be subjective. The introduction of objective criteria for assignment of secondary structure should allow the predictors to evaluate objectively the accuracy of their prediction schemes. Work on this project is already underway (M. Levitt, manuscript in preparation).

An automated system for selecting secondary structure introduces the possibility of producing schematic diagrams of protein secondary structure automatically and reproducibly. These can be used to represent secondary structure trends in molecules and to compare secondary structures in related molecules or in related molecular domains. Such work is in progress (J. Greer, manuscript in preparation).

Most fascinating to us is the ability to process objectively and reproducibly the secondary structure of a large number of proteins. Thus, intelligent analysis can now be performed on secondary structure environments and inter-secondary structure relations on a large sample of proteins automatically. We hope such studies will begin to elucidate the secrets of how protein structure forms with such exquisite precision.

We thank Dr Bruce Bush for his suggestion of the molecular representation used in Tables 7 to 10. This research was supported by the Medical Research Council, National Institutes of Health grant HL 16601, the Columbia University Computer Center, and the American Philosphical Society.

REFERENCES

Adams, M. J., Ford, G. C., Koehoek, R., Lentz, P. J. Jr, McPherson, A. Jr, Rossmann, M. G., Smiley, I. E., Schewitz, R. W. & Wonacott, A. J. (1970). *Nature* (*London*), 227, **10984103.**

Adams, M. J., Buehner, M., Chandrasekhar, K., Ford, G. C., Hackert, M. L., Liljas, A., Rossmann, M. G., Smiley, I. E., Allison, W. S., Everse, J., Kaplan, N. O. & Taylor, S. S. (1973). *Proc. Nat. Acad. Sci., U.S.A.* 70, **1968-1972.**

Adman, E. T., Sieker, L. C. & Jensen, L. H. (1976). *J. Biol. Chem.* 251, 3801-3806.

Argos, P., Schwarz, J. & Schwarz, J. (1976). *Biochim. Biophys. Acta, 439,* **261-273.**

Arnone, A., Bier, C. J., Cotton, F. A., Day, V. W., Hazen, E. E. Jr, Richardson, D. C., Richardson, J. C. & Yonath, A. (1971). *J. Biol. Chem. 246,* **2302-2316.**

Banner, D. W., Bloomer, A. C., Petsko, C. A., Phillips, D. C., Pogson, C. I., Wilson, I. A., Corran, P. H., Furth, A. J., Milman, J. D., Offord, R. E., Priddle, J. D. & Waley, S. G. (1975). *Nature (London), 255,* **609-614.**

Birktoft, J. J. & Blow, D. M. (1972). *J. Mol. Biol. 68,* **187-240.**

Blake, C. C. F. & Evans, P. R. (1974). *J. Mol. Biol. 84,* 585–601.

Blake, C. C. F., Koenig, D. F., Mair, G. A., North, A. C. T., Phillips, D. C. & Sarma, V. R. (1965). *Nature (London), 206,* **757-761.**

Blake, C. C. F., Mair, G. A., North, A. C. T., Phillips, D. C. & Sarma, V. R. (1967). *Proc. Roy. Soc. ser.* **B167, 365-377.**

Blake, C. C. F., Geisow, M. J., Swan, I. D. A., Rerat, C. & Rerat, B. (1974). *J. Mol. Biol.* **88, l-12.**

Blundell, T. L., Dodson, G., Hodgkin, D. & Mercola, D. (1972). *Advan. Protein Chem.* **26, 279-402.**

Bolton, **W., Cox, J. M. & Perutz, M. F. (1968).** *J. Mol. Biol. 33, 283-297.*

Bryant, T. N., Watson, H. C. & Wendell, P. L. (1974). *Nature (London), 247,* 14–17.

Burgess, A. W., Ponnuswamy, P. K. & Scheraga, H. A. (1974). *Isr. J. Chem.* **12, 239-286.**

Burnett, R. M., Darling, G. D., Kendall, D. S., Le Quesne, M. E., Mayhew, S. G., Smith, W. W. & Ludwig, M. L. (1974). *J. Biol. Chem. 249, 4383-4392.*

Campbell, J. W., Watson, H. C. & Hodgson, G. I. (1974). *Nature (London), 250,* **301-303.**

Carlisle, C. H., Palmer, R. A., Mazumdar, S. K., Gorinsky, B. A. & Yeates, D. G. R. (1974). *J. Mol. Biol. 85,* **l-18.**

Carter, C. W. Jr, Kraut, J., Freer, S. T., Xuong, N. H., Alden, R. A. & Bartsch, R. G. (1974). *J. Biol. Chem.* **249, 4212-4225.**

Chou, P. Y. & Fasman, G. D. (1974). *Biochemistry,* **13, 222-245.**

Colman, P. M., Jansonius, J. N. & Matthews, B. W. (1972). *J. Mol. Biol.* **70, 701-724.**

Crawford, J. L., Lipscomb, W. N. & Schellman, *C. G.* **(1973).** *Proc. Nat. Acad. Sci., U.S.A.* **70, 538-542.**

Deisenhofer, J. & Steigemann, W. (1975). *Acta Crystallogr. ser. B,* **31, 238-250.**

Delbaere, L. T. J., Hutcheon, W. L. B., James, M. N. G. & Thiessen, W. R. (1975). *Nature (London), 257, 758-763.*

Diamond, R. D. (1966). *Acta Crystallogr.* **21, 253-266.**

Dickerson, R. E. & Timkovich, R. (1975). In *The Enzymes* **(Boyer, P., ed.), vol. 11, pp. 397-547, Academic Press, New York.**

Drenth, J., Jansonius, J. N., Koekoek, R. & Wolthen, B. G. (1971*a*). *Advan. Protein Chem.* **25, 79-115.**

Drenth, J., Hol, W. G. T., Jansonius, J. N. & Koekoek, R. (1971*b*). *Cold Spring Harbor Symp. Quant. Biol. 36,* **107-116.**

Drenth, J., Hol, W. G. T., Jansonius, J. N. & Koekoek, R. (1972). *Eur. J. Biochem. 26,* **177-181.**

Edmundson, A. B., Ely, K. R., Girling, R. L., Abola, E. E., Schiffer, M. & Westholm, F. A. (1974). In *Progress in Immunology II* (Brent, L. & Holbrow, J., eds), vol. 1, pp. 103-l 13, North-Holland, Amsterdam.

Eklund, H., Nordstrom, B., Zeppezauer, E., Söderland, G., Ohlsson, I., Biowe, T., Söderberg, B. O., Tapia, O., Bränden, C. I. & Akeson, A. (1976). *J. Mol. Biol.* **102, 27-59.**

Epp, O., Lattman, E. E., Schiffer, M., Huber, R. & Palm, W. (1975). *Biochemistry,* **14, 4943-4952.**

Fermi, G. (1975). *J. Mol. Biol. 97, 237-256.*

Flory, P. J. (1969). In *Statistical Mechanics of Chain Molecules,* **pp. 248-306, Wiley, New York.**

Freer, S. T., Kraut, J., Robertus, J. D., Wright, H. T. & Xuong, *N.* **H.** *(1970). Biochemistry,* **9, 1997-2009.**

Greer, J. (1974). *J. Mol. Biol. 82,* **279-301.**

Greer, J. (1975). *J. Mol. Biol. 98, 649-653.*

Greer, J. (1976a). *J. Mol. Biol.* 100, 427-458.
Greer, J. (1976b). *J. Mol. Biol.* **104, 371-386.**
Hardman, K. D. & Ainsworth, C. F. (1972). *Biochemistry,* 11, 4910-4919.
Hartsuck, J. A. & Lipscomb, W. N. (1971). In *The Enzymes* (Boyer, P., ed.), **vol. 3, pp. l-56, Academic Press, New York.**
Hendrickson, W. A., Love, W. E. & Karle, **J. (1973).** *J. Mol. Biol. 74,* **331-361.**
Herriott, J. R., Sieker, L. C., **Jensen, L. H. & Lovenberg, W. (1970).** *J. Mol. Biol.* **50, 391-406.**
Holmgren, A., Söderberg, **B. 0 ., Eklund, H. &** Bränden, **C. Z. (1975).** *Proc.* **Nut.** *Acad. Sci., U.S.A.* **72, 2305-2309.**
Huber, R., Kukla, D., Rühlmann, **A. & Steigemann, W. (197 1). In** *Proc. Int. Res. Conf. on Protease Inhibitors* **(Fritz, H. & Tschesche, H., eds), pp. 56-65, Walter de** Gruyter, **Berlin.**
Kannan, K. K., Notstrand, B., Fridborg, K., Lövgren, **S., Ohlsson, A. & Petef, M. (1975).** *Proc. Nat. Acad. Sci., U.S.A. 72,* **51-55.**
Kendrew, **J. C., Dickerson, R. E., Strandberg, B. E., Hart, R. G., Davies, D. R., Phillips, D.** *C.* **& Shore, V. C. (1960).** *Nature (London),* **185, 422-427.**
Kendrew, **J. C.,** Klyne, **W., Lifson, S., Miyazawa, T., Nemethy, G., Phillips, D. C., Ramachandran, G. N. & Scheraga, H. A. (1970).** *Biochemistry, 9,* **3471-3479.**
Kretsinger, R. H. & Nockholds, **C. E. (1973).** *J. Biol. Chem.* **248, 3313-3326.**
Kreiger, M., Kay, L. M. & Stroud, R. M. (1974). *J. Mol. Biol. 83, 209-230.*
Kuntz, I. D. (1972). *J. Amer. Chem. Soc.* **94, 4009-4012.**
Levitt, M. (1974). *J. Mol. Biol. 82, 393-420.*
Levitt, M. **(1976).** *J. Mol. Biol.* **104, 59-107.**
Levitt, M. & Chothia, *C.* **(1976).** *Nature (London),* **261, 552-558.**
Levitt, M. & Warshel, A. (1975). *Nature (London), 253, 694-698.*
Lewis, P. N., Momany, **F. A. & Scheraga, H. A. (197 1).** *Proc. Nat. Acad. Sci., U.S.A.* **68, 2293-2297.**
Liljas, A., Kannan, K. **K., Bergsten, P. C., Waara, I., Fridborg, K., Strandberg, B., Carlbom, V.,** Järup, **L.,** Lövgren, **S. & Petef, M.** *(1972). Nature New Biol. 235,* **131-137.**
Lim, V. I. (1974). *J. Mol. Biol. 88, 873-894.*
Ludwig, M. L., Hartsuck, J. A., Steitz, **T. A., Muirhead, H., Cuppola, J. C.,** Reeke, **G. N. & Lipscomb, W. N. (1967).** *Proc. Nat. Acad. Sci., U.S.A. 57,* **511-514.**
Marsh, R. E., Corey, R. B. **& Pauling, L. (1955).** *Biochim. Biophys. Acta,* 16, l-34.
Mathews, F. S., Argos, P. & Levine, M. (1972). *Cold Spring Harbor Symp. Quant. Biol.* **36, 387-395.**
Matthews, B. W. (1975). *Biochim. Biophys. Acta, 405,* **442-451.**
Maxfield, F. R. & Sheraga, H. A. (1976). *Biochemistry,* **15, 5138-5153.**
Mayhew, **S. G. & Ludwig, M. L. (1975). In** *The Enzymes* **(Boyer, P.,** ed.), **vol. 12, pp. 57-119, Academic Press, New York.**
Moras, D., Olsen, K. W., Sabesan, M. N., Buehner, M., Ford, G. C. & Rossmann, M. G. (1975). *J. Biol. Chem. 250,* **9137-9162.**
Muirhead, H. & Perutz, M. F. (1963). *Nature (London),* 199, 633-639.
Nishikawa, K., Ooi, T., Isogai, Y. & Saito, *N.* **(1972).** *J. Phys. Soc. Japan,* **32, 1331-1337.**
Nishikawa, K., Momany, **F. A. & Scheraga, H. A. (1974).** *Macromolecules, 7,* **797-810.**
Padlam, **E. A. & Love, W. E. (1974).** *J. Biol. Chem. 249, 4067-4078.*
Pauling, L. & Corey, R. B. (1951). *Proc. Nat. Acad. Sci., U.S.A.* **37, 729-738.**
Pauling, L., Corey, R. B. & Branson, **H. R. (1951).** *Proc. Nat. Acad. Sci., U.S.A. 37,* **205-211.**
Perutz, M. F. (1951). *Nature (London),* **167, 1053-1054.**
Perutz, M. F., Rossmann, M. G., Cullis, **A. F., Muirhead,** H., **Will, G. & North, A. C. T. (1960).** *Nature (London),* **185, 416-422.**
Phillips, D. C. **(1967).** *Proc. Nat. Acad. Sci., U.S.A.* **57, 484-495.**
Phillips, D. C. (1970). **In** *British Biochemistry, Past and Present* **(Goodwin, T. W., ed.), pp. 11-28, Academic Press, London.**
Poljak, R. J., Amzel, L. M., Cher, B. L., Phizackerley, **R. P. & Saul, F. (1974).** *Proc, Nat, Acad. Sci., U.S.A.* **71, 3440-3444.**

Poulos, T. L., Alden, R. A., Freer, S. T., Birktoft, J. J. & Kraut, J. (1976). *J. Biol. Chem.* **251, 1097-1103.**

Reeke, **G. N. Jr, Becker, J. W. & Edelman, G. M. (1975).** *J. Biol. Chem.* **250, 1525–1547.**

Richardson, J. S., Thomas, K. A., Rubin, **B. H. & Richardson, D. C. (1975).** *Proc. Nat. Acad. Sci., U.S.A.* **72, 1349-1353.**

Robson, **B. & Pain, R. H. (1974).** *Biochem. J.* 141, **186-195.**

Salemme, **F. R., Freer, S. T., Xuong, N. H., Alden, R. A. & Kraut, J. (1973).** *J. Biol. Chem.* **248, 3910-3921.**

Sawyer, L., Shotton, **D. M. & Watson, H. C. (1973).** *Biochem. Biophys. Res. Commun.* **53, 944–951.**

Schmid, **M. F. & Herriott, J. R. (1976).** *J. Mol. Biol.* 103, **175-190.**

Schulz, G. E., Elzinga, M., Marx, F. & Schirmer, R. H. (1974*a*). *Nature (London),* **250, 120-123.**

Schulz, G. E., Barry, C. D., Friedman, J., Chou, P. Y., Fasman, G. D., Finkelstein, A. V., Lim, V. I., Ptitsyn, 0. B., Kabat, E. A., Wu, T. T., Levitt, M., Robson, **B. &** Nagano, **K.** (1974*b*). *Nature (London), 250,* **140-142.**

Srinivasan, R., Balasubramanian, R. & Rajan, S. S. (1975). *J. Mol. Biol. 98, 739-747.*

Steitz, **T. A., Fletterick, R. J., Anderson, W. F. & Anderson, C. M. (1976).** *J. Mol. Biol.* 104, **197-222.**

Takano, T., Kallai, 0. B., Swanson, R. & Dickerson, R. E. (1973). *J. Biol. Chem. 248,* **5234-5255.**

Tanaka, N., Yamane, T., Tsukihara, T., Ashida, **T. & Kakudo, M. (1975).** *J. Biochem. (Tokyo),* **77, 147-162.**

Tanaka, S. & Scheraga, **H. A.** (1976*a*). *Macromolecules*, **9, 142-182.**

Tanaka, S. & Scheraga, **H. A.** (1976*b*). *Macromolecules*, 10, **187-210.**

Timkovich, R. & Dickerson, R. E. (1973). *J. Mol. Biol. 79, 39-56.*

Tulinsky, A., Mari, **N. V., Morimoto, C. N. & Vandlen, R. L. (1973).** *Acta Crystallogr. ser. B, 29,* **1309-1322.**

Venkatachalam, C. M. (1968). *Biopolymers*, **6, 1425-1436.**

Ward, K. B., Hendrikson, W. A. & Klippenstein, *G.* **L. (1975).** *Nature (London), 257,* **818-821.**

Warshel, A. & Levitt, M. (1976). *J. Mol. Biol.* 106, **421-437.**

Watenpaugh, K. D., Sieker, L. C., Herriott, J. R. & Jensen, L. H. (1973). *Acta Crystallogr. ser. B, 29, 943-956.*

Watson, H. *C.* **(1969).** *Prog. Stereochem. 4, 299-233.*

Watson, H. C., Shotton, **D. M., Cox, J. M. & Muirhead, H. (1970).** *Nature (London), 225,* 806–816.

Webb, L. E., Hill, E. J. & Banaszak, L. J. (1973). *Biochemistry,* 12, **5101-5109.**

White, J. L., Hackert, **M. L., Buehner, M., Adams, M. J., Ford, G. C., Lentz, P. J. Jr, Smiley, I. E., Steindel,** S. **J. & Rossmann, M. G. (1976).** *J. Mol. Biol.* 102, 759-779.

Wright, C. S., Alden, R. A. & Kraut, J. (1969). *Nature (London),* 221, **235-242.**

Wyckoff, H. W., Tsernoglou, D., Hanson, A. W., Knox, J. R., Lee, B. & Richards, F. M. (1970). *J. Biol. Chem. 245, 305-328.*