Secondary structure assignment for α/β proteins by a combinatorial aproach

ARTICLE *in* BIOCHEMISTRY · NOVEMBER 1983

Impact Factor: 3.02 · DOI: 10.1021/bi00290a005 · Source: PubMed

4 AUTHORS, INCLUDING:

Fred E Cohen
University of California, San Francisco
269 PUBLICATIONS **29,868** CITATIONS
SEE PROFILE

Robert Abarbanel
Jonova Inc., Seattle
19 PUBLICATIONS **637** CITATIONS
SEE PROFILE

from the edge to the center of the nucleosome.

Acknowledgments

I thank Dr. Roger Kornberg, in whose laboratory early experiments of this work were carried out.

Registry No. Exo III, 9037-44-9.

References

Finch, J. T., Lutter, L. C., Rhodes, D., Brown, R. S., Rushton, B., Levitt, M., & Klug, A. (1977) *Nature (London) 269*, 29-36.

Gariglio, P., Llopis, R., Oudet, P., & Chambon, P. (1979) *J. Mol. Biol. 13*, 75-105.

Germond, J. E., Hirt, B., Oudet, P., Gross-Bellard, M., & Chambon, P. (1975) *Proc. Natl. Acad. Sci. U.S.A. 72*, 1843-1847.

Klug, A., & Lutter, L. C. (1981) *Nucleic Acids Res. 9*, 4267-4283.

Kornberg, R. D. (1977) *Annu. Rev. Biochem. 46*, 931-954.

Leffak, I. M., Grainger, R., & Weintraub, H. (1977) *Cell (Cambridge, Mass.) 12*, 837-845.

Lutter, L. C. (1978) *J. Mol. Biol. 124*, 391-420.

Lutter, L. C. (1979) *Nucleic Acids Res. 6*, 41-56.

Lutter, L. C. (1981) *Nucleic Acids Res. 9*, 4251-4265.

Maniatis, T., Jeffrey, A., & Van de Sande, H. (1975) *Biochemistry 14*, 3787-3794.

Noll, M. (1974) *Nucleic Acids Res. 1*, 1573-1578.

Noll, M. (1977) *J. Mol. Biol. 116*, 49-71.

Peck, L. J., & Wang, J. C. (1981) *Nature (London) 292*, 375-378.

Prunell, A. (1982) *EMBO J. 1*, 173-179.

Prunell, A., & Kornberg, R. D. (1977) *Cold Spring Harbor Symp. Quant. Biol. 42*, 103-108.

Prunell, A., & Kornberg, R. D. (1978) *Philos. Trans. R. Soc. London, Ser. B 283*, 269-273.

Prunell, A., Kornberg, R. D., Lutter, L., Klug, A., Levitt, M., & Crick, F. H. C. (1979) *Science (Washington, D.C.) 204*, 855-858.

Rhodes, D., & Klug, A. (1981) *Nature (London) 292*, 378-380.

Richardson, C. C., Lehman, I. R., & Kornberg, A. (1964) *J. Biol. Chem. 239*, 251-258.

Riley, D., & Weintraub, H. (1978) *Cell (Cambridge, Mass.) 13*, 281-293.

Simpson, R. T., & Whitlock, J. P. (1976) *Cell (Cambridge, Mass.) 9*, 347-353.

Simpson, R. T., & Kunzler, P. (1979) *Nucleic Acids Res. 4*, 1387-1415.

Strauss, F., & Prunell, A. (1982) *Nucleic Acids Res. 10*, 2275-2293.

Strauss, F., & Prunell, A. (1983) *EMBO J. 2*, 51-56.

Strauss, F., Gaillard, C., & Prunell, A. (1981) *Eur. J. Biochem. 118*, 215-222.

Thomas, J. O., & Butler, P. J. G. (1977) *J. Mol. Biol. 116*, 769-781.

Wang, J. (1978) *Cold Spring Harbor Symp. Quant. Biol. 43*, 29-33.

Wang, J. (1979) *Proc. Natl. Acad. Sci. U.S.A. 76*, 200-203.

Worcel, A., Strogatz, S., & Riley, D. (1981) *Proc. Natl. Acad. Sci. U.S.A. 78*, 1461-1465.

Secondary Structure Assignment for α/β Proteins by a Combinatorial Approach†

Fred E. Cohen, Robert M. Abarbanel, I. D. Kuntz,* and Robert J. Fletterick

ABSTRACT: We describe an algorithm for assigning the secondary structure of α/β proteins. Turns are identified very accurately (98%) by simultaneously considering hydrophilicity and the ideal spacing of turns throughout the sequence. The segments bounded by these turns are labeled by a pattern-recognition scheme based on the physical properties of α-helices and β-strands, in this class of proteins. Long-range, as well as local, information is incorporated to enhance the quality of the assignments. Although the assignment for any one sequence is not unique, at least one of the assignments bears a close resemblance to the native structure. The algorithm successfully divides protein sequences into two classes: α/β and non-α/β. The accuracy of the secondary-structure assignments in the α/β class is sufficient to provide useful input for tertiary-structure assignments.

Protein tertiary structure is specified by the primary amino acid sequence (Anfinsen et al., 1961). There have been many attempts to understand how. The two main avenues have been (1) the direct use of energy-minimization techniques (Momany et al., 1975; Levitt, 1976; Robson & Osguthorpe, 1979) and (2) a two-step process that converts the sequence into a secondary-structure representation followed by the construction of a three-dimensional structure by the packing together of the secondary elements (Richmond & Richards, 1978; Rose, 1979; Cohen et al., 1979, 1980, 1981a,b, 1982; Sternberg et al., 1982; Cohen & Sternberg, 1980; Sternberg & Cohen, 1982).

Energy minimization is based on sound chemical principles, but it has not been particularly successful for protein structure. The difficulties arise from the vastness of the conformation space, the mathematical limits of the optimization algorithms, and the inadequacies of the existing potential functions to

† From the Department of Pharmaceutical Chemistry (F.E.C. and I.D.K.), School of Pharmacy, and the Section on Medical Information Science (R.M.A.) and the Department of Biochemistry and Biophysics (R.J.F.), School of Medicine, University of California, San Francisco, California 94143. *Received March 11, 1983.* This work was supported by grants from the National Institutes of Health (GM19267 and AM26081). F.E.C. was supported by a grant from the American Cancer Society (SPF-21). R.M.A. was partially supported by a training grant from the National Library of Medicine (5T15LM07000).

0006-2960/83/0422-4894$01.50/0

measure the free energy of the protein–solvent system. These problems have been reviewed previously (Nemethy & Scheraga, 1977).

The process of predicting secondary features and then arranging them into tertiary structure has been analyzed by the hierarchic condensation model of protein folding (Schulz & Schirmer, 1979). While such a model can have kinetic, as well as structural, connotations, it is only the latter that interest us here. Specifically, we are motivated by the series of papers by Cohen, Sternberg, and co-workers (Cohen et al., 1979, 1980, 1981a,b, 1982; Cohen & Sternberg, 1980; Sternberg & Cohen, 1982; Sternberg et al., 1982) that show that sequence and secondary-structure information can be combined with a few geometric packing rules to produce a small set of tertiary structures, at least one of which resembles the native structure. We assume that this approach offers a useful first-order solution to the packing problem and, in this paper, will direct our attention to the prior step: the relation of amino acid sequence to secondary structure.

Secondary structure assignment has been an area of some interest since the first protein structures became available. Current methods rely on some combination of two ideas. The first idea is that the large number of known structures provides a valuable data base for the statistical propensities of individual amino acids, or groups of amino acids, to adopt particular secondary conformations [e.g., Chou & Fasman (1974)]. The second idea uses specific physical models of secondary-structure formation and stability [e.g., Lim (1974)]. While many algorithms of both kinds have been developed, with distressing regularity, their overall accuracy has failed to exceed 75%. Further, attempts so far have invariably misassigned one or more of the secondary-structure elements in all the proteins studied. Failure of this type severely limits moving from secondary to tertiary predictions because a near native tertiary structure cannot be found if even a few of the secondary assignments are incorrect. The most likely source of these difficulties is that neither the statistical- nor the physical-model algorithms take proper account of long-range interactions that are likely to be crucial in determining some aspects of backbone conformation.

Our purpose in this paper is to suggest a route to more productive secondary-structure assignments. We will use the physical-model approach. The model of a generic protein will lead to a series of statements about the types of amino acid sequences that are expected to be associated with secondary elements. The general strategy, adapted from the "expert systems" formulation of the artificial intelligence field (Barr & Feigenbaum, 1981), consists of stating explicit hypotheses or rules about the system, providing a way of evaluating the validity of the rules, and developing a set of higher order rules (meta rules) to resolve conflicts. In our hands, the rules are simply lists of generalized amino acid sequences or patterns that are associated via the physical model with specific secondary structures. They are evaluated by examination of the known protein structures. The rules can be modified interactively as needed. The meta rules contain statements about the relationship of the secondary-structure elements themselves.

While our model is a natural evolution of the suggestions of Lim (1974), Nagano (1973), and others, the use of artificial-intelligence technology offers some novel departures from earlier efforts: (1) The algorithms are completely defined. (2) Useful results can be obtained even when a unique assignment is not possible. (3) The entire process is refinable so that new information or insights can be added and tested at any stage.

The general method is developed and applied here to α/β proteins (Levitt & Chothia, 1976). The tertiary structure of these proteins is largely specified by the packing of α-helices against a single β-sheet in a three-layer structure (Richardson, 1981). This tertiary structure is particularly well-defined and imposes some powerful constraints on the secondary assignments. We have been able to find a set of rules sufficient to identify all protein domains of this type containing a pure parallel β-sheet—about half of all α/β proteins. Domains that we have tested that do not have this tertiary structure are rejected by the procedure—that is, their sequences produce no acceptable secondary-structure assignments.

Finally, we stress our purpose in pursuing the secondary-structure question is to find an approach that allows continuation to a tertiary level. This goal requires that we focus on those secondary-structure features that form the "core" of the domain. We will ignore loosely packed units that are more peripherally involved. The advantages and disadvantages of this bias are discussed.

Theory and Methods

Model of Protein Structure. We make the following assumptions about globular proteins in aqueous environments: (a) Proteins consist of one or more sequentially contiguous domains. (b) Each domain is made from regular helical and/or β-sheet elements connected by turns. The internal geometry of the turns and the fine details of the helix and sheet structures will not concern us. Irregular chain conformations and regular structures that are not part of the domain core will be treated as "nulls". (c) The core of each domain is that subset of helices and/or β-elements whose primary packing interactions are within the domain. Features that link domains together or that are primarily involved in domain–domain interactions are explicitly excluded from our definition of the core. (d) The major interactions that position secondary elements with respect to each other arise from the packing together of hydrophobic amino acids.

While these assumptions are reasonably accurate for all water-soluble globular proteins, we advance them in the spirit of a starting approximation. More precise descriptions can be added to the physical model as the need arises.

α/β Proteins. For this paper, we specialize this picture to α/β proteins. We will only consider those domains containing a single all-parallel β-sheet or β-barrel with approximately one helix per strand. The helices must cover both sides of the sheet. Helices and isolated extended structures that do not interact strongly with the sheet are not retained as part of the core. The ordering of secondary elements is approximately $\beta/\alpha/\beta/\alpha$, but a number of variants are allowed. A structural motif of this kind allows us to draw several important geometric conclusions:

(A) Turns. Though the path of the polypeptide chain in a globular protein is complex, the important turns occur at fairly regular spatial and sequential intervals. These can be used to delimit or parse the primary sequence into secondary-structure elements. The segments, defined by the turns, are forced into one of three categories—helices, β-strands, or nulls. The number of core secondary-structure elements in α/β proteins is observed to be approximately $N/18$, where N is the number of amino acids in the domain. The number of turns, however, is greater than $N/18 - 1$ because of the noncore segments. It has been known for some years that turns are made from very polar amino acids (Kuntz, 1972; Rose, 1978). Thus, we begin our search for secondary elements by looking for short sequences of hydrophilic amino acids that are separated by approximately 14 residues. This follows from

the characteristic lengths of α-helices and β-strands in α/β proteins. The specific sequence patterns that are used are given below.

(*B*) *Helices.* Helices in the core region of an α/β domain have one face toward the hydrophobic β-structure. The opposite side of the helix is usually at the protein surface. The amino acid sequence should contain a set of hydrophobic residues placed to reflect the periodicity of the helix. Thus, if the hydrophobic patch begins at residue i, other hydrophobic residues are expected at $i - 1$, $i + 3$, $i + 4$, $i + 7$, and $i + 8$. The polar face of the helix would display hydrophilic residues at $i - 2$, $i + 1$, $i + 2$, and $i + 5$ (Richmond & Richards, 1978; Chothia et al., 1977; Schiffer & Edmundson, 1967, 1968).

(*C*) *β-Structures.* There are two types of β-sheets regularly seen in α/β proteins, planar and cylindrical. Planar β-sheets contain two types of strands. Those on the interior of the sheet are well shielded from solvent and should show hydrophobic residues at the interior sequential positions $i + 2$, $i + 3$, $i + 4$, and $i + 5$ in order to provide two hydrophobic faces. The remaining residues (i, $i + 1$, $i + 6$, $i + 7$) are likely to be polar. The strands on the edges of the sheets should show a much less regular pattern and should not have the three to four residue periodicity of helices. β-Barrels do not have edge strands. Each β-segment of a barrel should be roughly comparable to the strands on the interior of the β-sheets.

Although these distributions of hydrophobic and hydrophilic residues are useful, they are not absolute. An example is instructive. The sequence SVIMG[1] is in an α-helix in ADH, and SVIVG is in a β-strand in ADH. This could be due to the helical preference of methionine over valine (Chou & Fasman, 1974). However, VDIIN is in an α-helix in TIM, and MDVIN is in a β-strand in SBT. Since valine and isoleucine are commonly given similar roles, the methionine paradigm must be flawed. This is only one of many such cases. Thus, the importance of tertiary interactions in stabilizing secondary structure must be acknowledged.

(*D*) *Null Structures.* For this paper, either null segments can be irregular structures or they can be isolated helices and extended regions of chain that do not interact with the central β-sheet.

To summarize, our goal is to provide sufficiently accurate descriptions of the core secondary structure to drive the three-dimensional packing programs of Cohen et al. (1982). This goal does not necessarily require a unique secondary assignment. A set of such assignments is acceptable as long as it contains one assignment that is a good approximation to the native structure. A unique but incorrect assignment is of little value.

Algorithm

Overview. The first step is to prepare a list of amino acid sequences or patterns expected for turns, helices, and β-structure. These patterns consist of ordered arrangements of specific amino acid names or classes of amino acids. The detailed physical model of the protein structure often suggests appropriate patterns. Others can be discovered by examination or insight.

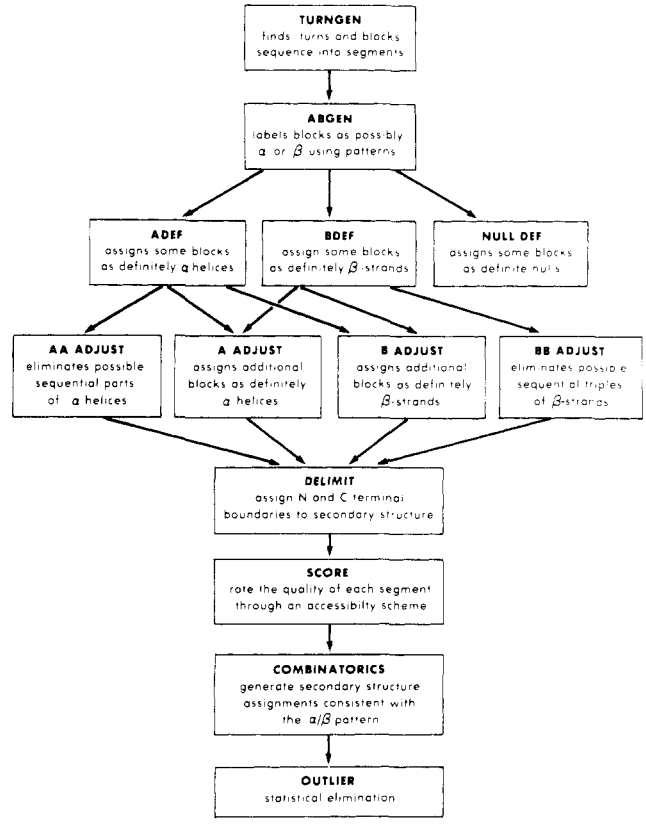

FIGURE 1: Schematic overview of the algorithm. The hierarchic organization of this algorithm is shown, and the nature of the interactions is seen. See Tables III and IV for the results of application of these procedures.

These patterns, in combination, provide presumptive assignments for the sequence under study. These assignments are processed further to yield consistency with such global considerations as (1) are there enough turns? (2) are there enough helices? (3) are there enough β-strands? (4) do helices and strands occur in approximate alternation? and (5) are there identifiable edge strands? The program makes choices that meet as many of these constraints as possible. All internally consistent choices are examined.

The final outcome of this effort can have three results. First, no consistent assignments may be possible. If the assumptions are valid, this forces the conclusion that the particular sequence is not a parallel α/β domain. Second, there may be a single assignment consistent with all the constraints. This is the most desirable result since it provides a unique—and presumably correct—secondary-structure assignment. Third, there may be many assignments that can be made for a specific sequence. If the assignment list includes a nativelike assignment, our present objective is satisfied. We recognize that this objective is not a common one. However, close inspection of the unresolved ambiguities can often suggest new experiments or new discriminations that can add to the basic knowledge base. The details of the pattern choices (Table I) and the program organization (Figure 1) are given below.

For purposes of this discussion, the method is most conveniently divided into three operations: turn generation, α- and β-pattern recognition, and higher order processing. The input information is the amino acid sequence and the number

[1] Abbreviations: A, alanine; C, cysteine; D, aspartic acid; E, glutamic acid; F, phenylalanine; G, glycine; H, histidine; I, isoleucine; K, lysine; L, leucine; M, methionine; N, asparagine; P, proline; Q, glutamine; R, arginine; S, serine; T, threonine; V, valine; W, tryptophan; Y, tyrosine; ADH, alcohol dehydrogenase; ADK, adenylate kinase; FXN, flavodoxin; LDH, lactate dehydrogenase; PGKa, N-terminal domain of phosphoglycerate kinase; PGKb, C-terminal domain of phosphoglycerate kinase; SBT, subtilisin; TIM, triosephosphate isomerase; RHDa, N-terminal domain of rhodanese; RHDb, C-terminal domain of rhodanese; B5C, cytochrome b_5; CHA, chymotrypsin A; CPAa, hypothetical N-terminal domain with the carboxypeptidase A sequence; CPAb, hypothetical C-terminal domain with the carboxypeptidase A sequence; CPV, carp parvalbumin; CYT, cytochrome *c*; MBN, myoglobin; REI, Bence–Jones immunoglobulin light chain; RSA, ribonuclease S; SOD, superoxide dismutase; TLNa, N-terminal domain of thermolysin; TLNb, C-terminal domain of thermolysin; TRX, thioredoxin; CPU, central processor unit.

Table I: Patterns Used in Secondary-Structure Recognition

label	description	symbols	pattern (any of)
		α-Helices	
S	strong hydrophobic diamond pattern without aromatics (Cohen et al., 1982)	s_1 = I, L, or V; $s_2 = s_1$ + C or M; $s_3 = s_2$ + A; *, any amino acid	$s_2{**}s_1s_1{**}s_1$[a] $s_1{**}s_3s_1{**}s_1$ $s_1{**}s_1s_3{**}s_1$ $s_1{**}s_1s_1{**}s_2$
H	generalized pattern of hydrophobics (Richmond & Richards, 1978)	h = A, C, F, I, K, L, M, V, W, or Y	h**hh**h h***hh**h hh***h**h hh**h***h hh**hh
A	hydrophobic diamond (Cohen et al., 1982)	a_1 = A, C, F, I, K, L, M, P, V, W, or Y; $a_2 = a_1$ + G, S, or T	$a_2{**}a_1a_1{**}a_1$ $a_1{**}a_2a_1{**}a_1$ $a_1{**}a_1a_2{**}a_1$ $a_1{**}a_1a_1{**}a_2$
P	hydrophilic stripe on helix	p_1 = D, E, G, H, K, N, P, Q, R, S, or T	$p_1{***}p_1{***}p_1$ $p_1{**}p_1{***}p_1$ $p_1{***}p_1{**}p_1$
C	charge pair with appropriate separation for interaction in an α-helix	+ = K or R; − = D or E	+***− −***+ +**− −**+
N_α	high density of disruptive hydrophilic residues or four strong hydrophobic residues	n_1 = G, P, or S; n_2 = C, F, I, L, M, V, W, or Y	$(3n_1, 3*)$[b] $n_2n_2n_2n_2$
		β-Strands	
B_1	high density of hydrophobics	b_1 = A, C, F, I, L, M, P, V, W, or Y	$b_1b_1b_1$ $b_1{*}b_1b_1$ $b_1b_1{*}b_1$
B_2	hydrophobic residues bounding hydrophilic residues	o = D, E, G, H, K, N, Q, R, S, or T; i = A, C, F, I, K, L, M, P, V, W, or Y	ooii**oi ooi*i*oi oo*ii*oi ooi**ioi oo*i*ioi oo**iioi
B_d	high density of hydrophobics, largely aliphatics	l = I, L, or V; f = A, F, W, or Y; c = C or M	(3l, 2f) (3l, 1f) (4l, 1*) (3l, 1f, 1c)
N_β	high density of disruptive hydrophilic residues	n_1 = G, P, or S	$(2n_1, 2*)$
		Edges of β-Sheets	
E	edge-type alteration of hydrophilic and hydrophobic	e_1 = D, E, G, H, K, N, P, Q, R, S, or T; e_2 = A, C, F, I, K, L, M, P, V, W, or Y	$e_2e_1e_1e_2$
N_ϵ	high density of disruptive hydrophilic residues or a sequence of five strong hydrophobic residues	n_1 = G, P, or S; n_3 = A, C, F, I, L, M, V, W, or Y	$(2n_1, 2*)$ $n_3n_3n_3n_3n_3$
		Turns	
T_1	high density of hydrophilics	t_1 = D, E, G, H, K, N, P, Q, R, S, or T; $t_2 = t_1$ + Y; $t_3 = t_2$ + A	$(3t_1, 1t_2)$ $(4t_1, 1*)$ $(5t_1, 2*)$
T_2	high density of hydrophilics but less than T_1	as above	$(3t_1, 1*)$ $(4t_1, 2t_3, 1*)$ $(3t_1, 4t_3)$
T_3	high density of hydrophilics but less than T_2	as above	$(4t_1, 4*)$ $(5t_1, 4*)$ $(4t_1, 2t_3, 3*)$

[a] Algebraic expressions of the form *abc* are replaced by the appropriate symbol when any residue equivalent to the symbol *a* is found in an arbitrary position *i* followed by residues of types *b* and *c* in relative positions *i* + 1 and *i* + 2. [b] Algebraic expressions of the form (2*a*, 3*b*) mean that for a sequence of five amino acids, two are of type *a* and three are of type *b*.

of secondary structure segments. The latter may be calculated as a simple function of the number of amino acids for the class of parallel α/β proteins that we consider.

Briefly, the turn-generation process converts the complete amino acid sequence into a set of nonoverlapping segments that each contain at most one continuous piece of secondary structure. Labeling deals with assigning to each segment the type or types of secondary structure that the segment could possibly assume. Higher order processing explores the interactions between the segments that had previously been considered in isolation to produce a secondary-structure prediction or predictions consistent with the complete sequence.

Turn Generation. There is no commonly accepted notion of a turn in protein structure. Rather, there is a spectrum ranging from the specific β-turns described by Venkatachalam (1968) to the more general changes in direction of the polypeptide chain used by Rose (1978). To avoid any controversy over these distinctions, we will define turns as regions separating segments of secondary structure. Turns have previously been identified by probability arguments (Chou & Fasman,

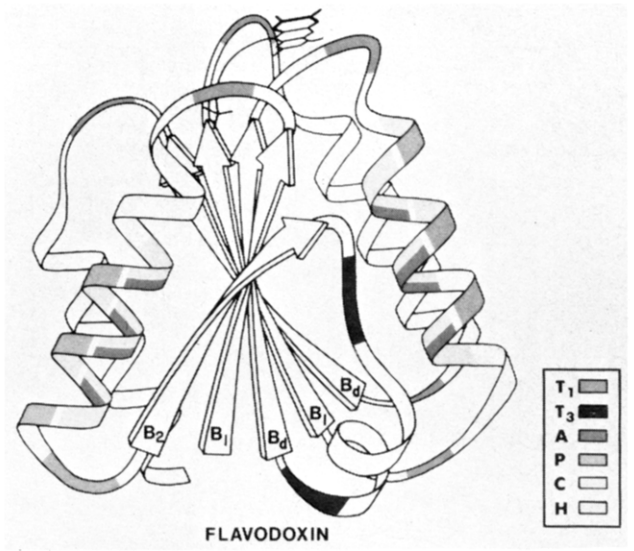

FIGURE 2: Pattern recognition in flavodoxin. An example of the three-dimensional location of the patterns (see Table I) observed along the sequence of flavodoxin. Arrows are β-strands and are labeled with the β-patterns. For clarity, the actual position of the pattern is not marked. Patterns on two of the four α-helices are shown. These indicate that the hydrophobic patterns (A and H) point toward the sheet and so are logically involved in α/β packing. The hydrophilic patterns (P and C) point toward the solvent. Turns are clearly exposed segments separating regions of secondary structure. Dr. J. S. Richardson supplied the pattern ribbon diagram.

1977), by the occurrence of specific amino acids (notably proline or glycine), or as valleys in a sequential plot of chain hydrophobicity (Kuntz, 1972; Rose, 1978). All published methods required the investigator to choose a cutoff that divided the chain into turns and nonturns. Unfortunately, no cutoff can be chosen that simultaneously includes all turns and excludes all nonturns with existing algorithms.

The TURNGEN procedure begins with the hydrophobic valley notion and searches the chain for dense accumulations of hydrophilic residues (see Table I and Figure 2). We label the very dense regions as definite turns (T_1) and use their positions along the chain to find other "weaker" turns (T_2, T_3, ...). The segments defined by T_1 turns are divided into four classes: (1) In short segments, no intervening turn is anticipated. Specifically, if the segment length L is <19, then any T_2 or T_3 turn labels are ignored. (2) In medium-length segments, two subsegments are sought. If the segment begins at residue N and is of length $L \geq 19$ and <30, then a T_2 (or T_3 if no T_2 pattern is found) in the region $N + (L/2) \pm 4$ is called a turn. (3) For longer segments, three subsegments are anticipated. If the segment begins at residue N and is of length $L \geq 30$ and <42, then a T_2 pattern (or T_3 if no T_2 is found) in the region of $N + (L/3) \bullet 4$ or $N + (2L/3) \pm 4$ is called a turn. (4) In the longest segments unsplit by T_1 patterns, no more than four subsegments have been found. If the segment begins at residue N and is of length $L \geq 42$, then a T_2 pattern (or T_3 if no T_2 is found) in the region of $N + (L/4) \pm 4$, $N + (L/2) \pm 4$, or $N + (3L/4) \pm 4$ is called a turn. On occasion, several T_2 (or T_3) patterns are found within a short sequence of residues. In these cases, the average position of the weaker turn indicators is taken as the turn center.

The net effect is to parse the sequence into segments of length 14 ± 4. β-Strands taken together with the turns that follow average 10 residues. In a similar way, α-helices contain approximately 18 residues in α/β proteins. The parsing parameter of 14 ± 4 was chosen to provide reasonable segment lengths, so that each segment contains no more than one secondary-structure unit. Similar values have been obtained independently by Taylor & Thornton (1983). In our experience, no segment longer than 72 residues remains unsplit by a T_1 pattern. Further, no segments with a length >18 residues remain after the second iteration; in principle, this could occur and would be met with an iterative procedure with an additional pattern T_4, which was less hydrophilic than T_3.

α/β Pattern Recognition and Labeling. The turn finding algorithm produces a list of segments with crude boundaries. No segment will contain more than one type of core secondary structure. It remains to assign secondary structure labels to these segments. This is done in several steps.

The first step is to identify patterns associated with structures α (helix), β (sheet), ϵ (edge strand), or φ (no core α, β, or ϵ). In a second step, we recognize some patterns as being so strong so as to allow definite assignments in those segments. Similarly, lack of identifying patterns or mutually contradictory patterns allow some assignments of definite null segments. As with turns, patterns are chosen that represent the physical properties of secondary structure. Special attention is paid to those patterns that facilitate secondary-structure packing.

Having screened all segments in this fashion, the segments for which a definite assignment has not been made can be rescrutinized in light of the assignments of neighboring segments. Further assignment of this group into definites and nulls can be accomplished with higher level rules (e.g., a core helical segment may not follow a core helix). Repeated use of this strategy can significantly reduce the number of segments with alternative assignments. One advantage of this construction is obvious: all unresolved segments remain available for further study, and the effect of application of new knowledge may be measured.

(*A*) *Possible Labels.* α- and β-recognition depend upon the segments having been defined by the turns. Each segment is scanned by the ABGEN procedure for specific amino acid sequence patterns consistent with α- or β-structure (Table I), and the center of each region of the sequence matched is assigned a label. (1) A segment is labeled as a possible α-helix if it contains a sequence of three residues that, taken together, have been labeled with three of the four helical pattern markers: C, H, A, and P (Table I). This guarantees that one portion of the segment is well suited for helix nucleation and subsequent helix–helix or helix–sheet packing. (2) A segment is labeled as a possible β-strand if it contains the B_1 or B_2 pattern without the charge-pair pattern C. The B_1 pattern has a high density of hydrophobic residues and is typical of the internal strands in β-sheets. The B_2 pattern is frequently encountered in edge strands. (3) A segment is labeled as a possible edge strand if it contains the edge pattern E (Table I) within five residues of a β-strand pattern. (4) All segments are considered as possible noncore (null) segments.

(*B*) *Definite Labels.* At the next level (A DEF, B DEF, NULL DEF), some segments are assigned as definite α or β or null (that is, irregular coils or other sequences not in the secondary structure of the α/β core). (1) A segment is labeled as definitely α if it has been labeled as possibly α and contains the pattern S or if it contains no β-signal and less than one-third of the residues are charged (D, E, H, K, and R). If oppositely charged residues are spaced by four along the chain, a pattern that would facilitate ionic interaction without disrupting the helix, they are not counted as charged. (2) A segment is labeled as definitely β if it has been labeled as possibly β and contains the pattern B_d or if it has been labeled as possibly β, contains no α-signal, lacks the N_β pattern, has β-accessibility >250 Å^2 (see Scoring and Ranking), and less than one-fourth

of the residues are charged. If oppositely charged residues are spaced by two along the chain, a pattern that would facilitate ionic interaction without disrupting the strand, they are not counted as charged. A definite β-assignment permits a strand to be either internal or edge, depending on other factors (see below). (3) A segment is labeled as definitely null if (a) it has no possible α- or β-label, (b) it has only possible α-labels and contains the pattern N_α[2] or greater than one-third of the residues are charged (D, E, H, K, and R), (c) it has only possible β-labels and contains the pattern N_β,[3] or (d) it has both possible α- and β-labels and contains both N_α and N_β.

Higher Order Processing. Higher order processing is devoted to identifying long-range interactions that would preferentially stabilize or disrupt secondary structure.

(*A*) *Adjustments to Segment Labels.* The ADJUST modules apply specific observations about known α/β domains to the segments neighboring those with definite assignments. (1) A segment between or neighboring two segments with definite β-labels is not allowed to be β. (2) A segment next to a segment with a definite α-label is not allowed to be α. (3) A segment bounded by two segments that are both labeled definitely α is labeled definitely β if it contains any β-pattern. If not, the protein is not of the α/β class.

(*B*) *End Points of Secondary Structures.* The next stage in the processing begins with the DELIMIT procedure, which assigns specific N and C termini to the segments by looking for logical extensions of the turns that do not obscure the secondary-structure pattern. If fewer than five α-residues or four β-residues remain in the segment, then it is assigned as a null. This procedure has different considerations for α-helices and β-strands. (1) For β-strands, we begin at the residue marked by the β-patterns and proceed in both the N- and C-terminal directions until specific stop signals are encountered. These signals are charged residues[4] or three sequential hydrophilic residues. Sequences of the form iioo[5] on the C terminus and ooii on the N terminus are used whenever possible to delimit the strands with the three inner residues said to be in the strand. The labels divide residues between hydrophilic and hydrophobic. Note that lysine (K) is given an ambivalent role. The long aliphatic side chain can frequently participate in hydrophobic interactions without disrupting the tendency of the polar NH_3^+ group to seek the solvent or an internal carboxylate. In edge strands, lysine (K) and proline (P) are allowed to be both hydrophilic and hydrophobic. This more conservative handling of edge strands results in consistently longer strands that are less hydrophobic on average. (2) Similarly for α-helices, stop signals are sought to produce a region bounded by prolines or three hydrophilics.[6] Sequences of the form ioio on the C terminus and oioi on the N terminus are used whenever possible to delimit the helices with the three inner residues said to be in the helix.

(*C*) *Scoring and Ranking.* At this point, all definite assignments have been made. Those residue associated with each segment assignment are known. To assign a relative preference to any remaining nondefinite segments, we use a numerical evaluation based on a sequence-weighted accessibility formula (Richmond & Richards, 1978; Cohen et al., 1982). This is calculated in the SCORE procedure. The accessibility weights (*A*) for each amino acid are as per Richards & Richmond (1978). The secondary-structure accessibility ($\sum$) formulas are as follows:

$$\sum_\alpha = 1.5(A_i + A_{i+7}) + 1.0(A_{i+3} + A_{i+4}) + 0.5(A_{i-1} + A_{i+8})$$

$$\sum_\beta = 1.0(A_{i+2} + A_{i+3} + A_{i+4} + A_{i+5}) + 0.5(A_{i+1} + A_{i+6}) - 0.5(A_i + A_{i+8})$$

$$\sum_\epsilon = 500 - \sum_\beta$$

$$\sum_\varphi = 250 \text{ if } \beta\varphi\beta$$
$$= 150 \text{ if } \alpha\varphi\beta \text{ or } \beta\varphi\alpha$$
$$= 300 \text{ if the N or C terminus is next to a null}$$

The scoring system is not used to resolve the nondefinite assignments at this point. Instead, it allows the rank ordering of assignments on the basis of the hydrophobic packing of the secondary structure.

(*D*) *Structure Generation.* The COMBINATORICS procedure generates all possible lists of secondary-structure assignments, each containing all segments assigned as definite α, β, or null, as well as all possible permutations of the remaining segment assignments. Only secondary-structure assignments with the following properties are retained: (1) There are no $\alpha\alpha$, $\beta\beta\beta$, or $\beta\beta\alpha\beta\beta$ regions. (2) There are zero (for β-barrels) or two edge strands. (3) Edge strands occur in the sequentially first but not second or last position, or edge strands occur in the sequentially last but not in the first or penultimate positions. (4) The number of residues between centers of secondary-structure segments in the $\beta\alpha\beta$ unit fall within a prescribed range (see Table II). (5) The difference in the number of helices on opposite sides of the β-sheet is no more than one.[7] (6) The number of secondary-structure segments is $N/18 \pm 1$.[8]

(*E*) *Outliers.* In reviewing the remaining structures at this juncture, it was clear that certain structures had accumulated a number of minor flaws without ever possessing a major flaw, which would have prompted their elimination. To make this statistical impression explicit, the OUTLIER procedure was

Table II: Ranges for Intersegment Distances in Residues[a]

	minimum	maximum
Adjacent Segments		
$\alpha\alpha$	n.a.	n.a.
$\alpha\beta$	4	50
$\beta\alpha$	4	50
$\beta\beta$	15	50
Next to Adjacent Segments		
$\alpha X\alpha$	15	90
$\alpha X\beta$	30	60
$\beta X\alpha$	30	60
$\beta X\beta$	15	90

[a] Distances between secondary-structure midpoints allowed in final structures. X represents any intervening segment type.

[2] This pattern is disruptive of normal α-helical secondary structure.

[3] This pattern is disruptive of normal β-strand secondary structure.

[4] Charged residues are D, E, H, K, and R. Residues are not counted as charged if they are oppositely charged and separated by one intervening residue.

[5] o stands for A, C, F, I, L, K, M, P, V, W, or Y (hydrophobics); i stands for D, E, G, H, K, N, Q, R, S, or T (hydrophilics). The charged residues are not considered hydrophilic if they are oppositely charged and separated by one intervening residue.

[6] The hydrophilic residues are D, E, G, H, K, N, P, Q, R, S, and T. Charged residues are not called hydrophilic if they can form a salt bridge with an oppositely charged residue displaced along the sequence by one, three, or four residues.

[7] If one assumes all connections between consecutive β-strands are right-handed, the number of helices on one side of the β-sheet is equal to the number of helices between the two edge strands. The count on the other side is simply the remaining helices.

[8] Although segments average 14 ± 4 residues in length, the frequent presence of noncore excursions in the chain dictates that the number of core segments is less than the number of segments bounded by hydrophilic regions. The number 18, which attempts to account for this discrepancy, is chosen empirically.

Table III: Results for 10 $\alpha/\beta_{parallel}$ Domains

protein	residues considered	correct no. of segments	no. of secondary structures: before combinatorial processing[a]	no. of secondary structures: finally proposed	position of native	% of structure common to all predictions (% equating edge and central β's)
ADH	165–361	12	31 104	113	8	41 (57)
ADK	1–194	11	384	1	1	100 (100)
FXN	1–138	9	128	4	4	44 (66)
LDH	1–201	11	36 864	12	2	54 (63)
PGKa	1–192	11	4 608	24	5	46 (63)
PGKb	193–416	11	248 832	67	4	40 (54)
RHDa	1–158	9	186 624	91	43	31 (31)
RHDb	159–293	9	34 992	51	3	27 (27)
SBT	1–180	9	589 824	45	7	25 (44)
TIM	1–247	15	294 912	14	14	71 (71)

[a] Number of secondary structures before combinatorial processing includes all possible labels for all segments including nulls. Structures with too few segments and those violating the regular $\beta\alpha\beta$ pattern will be eliminating during the combinatorial generation of predictions. If one considers only information about the location of turns, the number of possible secondary-structure predictions will be on the order of 10^4–10^8, depending on the length of the domain. Rank of native is the location of the native structure on an accessibility score sorted list as produced by the SCORE procedure.

Table IV: Results for 13 Non-$\alpha/\beta_{parallel}$ Domains

protein	residues considered	no. of secondary structures: before combinatorial processing[a]	no. of secondary structures: finally proposed	comments
B5C	1–93	54	0	not enough secondary structure found
CHA	1–245	12 582 900	0	$\beta\beta\alpha\beta\beta$ forced; not consistent with class
CPAa	1–200	331 776	0	definite $\alpha\alpha$ pair
CPAb	100–307	110 592	0	definite $\alpha\alpha$ pair
CPV	1–109	15 552	0	sheet possible; not consistent
CYT	1–104	1 152	0	not enough secondary structure found
LYZ	1–129	13 824	0	$\beta\beta\beta$ pattern not consistent with class
MBN	1–153	2	0	definite $\alpha\alpha$ pair
REI	1–107	96	0	too few helices to make pattern
RSA	1–124	2 592	0	$\beta\beta\beta$ pattern not consistent with class
SOD	1–152	192	0	too few helices to make pattern
TLNa	1–156	52 488	0	too few helices to make pattern
TLNb	157–316	10 368	0	helices not evenly spaced by strands
TRX	1–108	216	0	anti-parallel (short) strand connection

[a] Number of secondary structures before combinatorial processing includes all possible labels for all segments including nulls. Structures with too few segments and those violating the regular $\beta\alpha\beta$,pattern will be eliminating during the combinatorial generation of predictions. If one considers only information about the location of turns, the number of possible secondary-structure predictions will be on the order of 10^4–10^8, depending on the length of the domain.

developed. It eliminates structures that lie in the lower half of the distribution defined by SCORE. OUTLIER also counts the number of times any segment is used as a helix or strand as a fraction of the total number of times it could possibly be used in that role. Those that occur infrequently (less than 30% of the frequency of the most favored role for that segment) are considered to be insignificant and are eliminated from further investigation.

Implementation. The computer programs for performing these calculations are written in the C language (Kernighan & Ritchie, 1978) and run on a PDP 11/70 or VAX 750 with the UNIX operating system. An important feature of the program is a general pattern matching module that uses the powerful string processing routines of the UNIX system. A complete examination of a protein sequence of 300 residues requires between 40 and 50 s of CPU time.

Results

The secondary structure assignment algorithm was developed on eight α/β domains [ADH (Eklund et al., 1976), ADK (Schulz et al., 1974), FXN (Burnett et al., 1974), LDH (Holbrook et al., 1975), PGKa and -b (Banks et al., 1979), SBT (Drenth et al., 1971), TIM (Banner et al., 1975)] and uniformly applied to 10 α/β domains [RHDa and -b (Bergsma et al., 1975) added] and 13 non-α/β domains [B5C (Mathews et al., 1971), CHA (Segal et al., 1972), CPAa and -b (Hartsuck & Lipscomb, 1971), CPV (Kretsinger & Knockolds, 1973), CYT (Takano et al., 1973), LYZ (Imoto et al., 1972), MBN (Takano, 1977), REI (Epp et al., 1975), RSA (Richards & Wyckoff, 1971), SOD (Richardson et al., 1976), TLNa and -b (Colman et al., 1972), TRX (Holmgren et al., 1975)]. The results of these calculations are compiled in Tables III–V.

Central to our procedure is the success of TURNGEN, the turn-finding algorithm. In the 10 α/β domains examined, there are 123 regions between secondary-structure segments. All of these turns are found so every segment contains at most one type of secondary structure. This is a crucial feature. Two additional turns are found that split α-helices. The sequence QETK (145–148) in TIM splits the helix that runs from 138–154. While this shortens the helix, it does not cause a misassignment of the number of segments. The sequence QRNVN (111–115) in LDH splits the helix that includes residues 102–118. This LDH helix is not involved in α/β packing and is not included in our assignment. No other complications result. The overall accuracy of TURNGEN is 98%.

The other elements of this algorithm are less successful, but they are adequate for our purposes. The modules that identify

Table V: A Comparison of Experimental Secondary Structure and Assignment Results

protein	exptl secondary structure	best assignment	protein	exptl secondary structure	best assignment
ADH	β, 193–199, 218–224, 263–269, 287–293	β, 195–201, 216–223, 263–270, 287–294	PGKb	β, 207–212, 231–236, 332–336, 367–371	β, 207–212, 230–236, 332–336, 365–371
	ϵ, 238–243, 312–318	ϵ, 237–243, 313–321		ϵ, 277–282, 388–392	ϵ, 275–283, 386–393
	α, 179–187, 202–212, 229–236, 250–259, 275–283, 304–311, 328–338	α, 180–191, 204–210, 227–236, 246–259, 273–286, 328–338		α, 187–202, 218–229, 239–249, 317–330, 348–365, 375–380	α, 193–203, 216–229, 237–251, 320–330, 351–363, 376–384
ADK	β, 10–14, 90–94, 114–118	β, 9–15, 89–93, 113–119	RHDa	β, 29–33, 94–98, 122–126	β, 29–37, 92–99, 121–127
	ϵ, 35–38, 169–173	ϵ, 33–39, 169–172		ϵ, 7–10, 55–58	ϵ, 2–11, 52–58
	α, 1–8, 23–30, 69–84, 100–107, 144–158, 179–194	α, 2–7, 23–32, 53–63, 102–111, 143–153, 179–194		α, 11–22, 42–50, 76–87, 107–119	α, 12–25, 46–51, 78–91
FXN	β, 1–6, 48–55, 80–87	β, 2–6, 47–55, 81–86	RHDb	β, 176–181, 242–246, 268–271	β, 174–180, 208–214, 267–271
	ϵ, 29–35, 108–119	ϵ, 30–35, 106–116		ϵ, 159–162, 207–210	ϵ, 158–163, 241–249
	α, 10–27, 66–74, 93–106, 124–138	α, 11–26, 65–79, 93–105, 124–138		α, 163–174, 183–189, 224–235, 251–264, 274–282	α, 164–173, 225–238, 252–265, 273–283
LDH	β, 21–26, 46–51, 88–93, 130–134	β, 20–28, 47–52, 90–96, 130–137	SBT	β, 27–32, 89–94, 120–125, 148–152	β, 24–30, 89–94, 120–124, 146–151
	ϵ, 75–79, 155–158	ϵ, 68–76, 155–159		ϵ, 45–49, 174–177	ϵ, 42–51, 173–181
	α, 1–7, 29–43, 55–70, 120–127, 139–151	α, 2–14, 29–44, 57–67, 114–122, 140–154		α, 64–73, 103–117, 132–145	α, 63–75, 105–118, 133–143
PGKa	β, 17–22, 56–61, 114–119, 158–163	β, 17–23, 54–63, 114–118, 158–162	TIM	β, 6–12, 38–42, 60–63, 89–93, 122–129, 159–167, 205–209, 227–231	β, 3–8, 38–43, 59–66, 89–94, 121–128, 159–165, 199–202, 228–232
	ϵ, 91–96, 182–187	ϵ, 92–97, 184–190			
	α, 36–42, 77–89, 101–109, 144–155, 173–178	α, 39–44, 76–88, 102–110, 146–156, 171–179		α, 17–31, 44–55, 79–87, 105–120, 138–154, 177–196, 213–223, 237–246	α, 18–28, 44–54, 78–85, 108–120, 147–155, 182–195, 238–245

definite segment assignments (A DEF, B DEF, NULL DEF) label 20% (20 segments) of the secondary structures solely from a consideration of sequence. The modules that assign segments to α-helix or β-strands by considering both sequence and properties of the neighboring segment sequences (A ADJUST, AA ADJUST, B ADJUST, BB ADJUST) label an additional 30% (29 segments) of the secondary structure. Both the DEF and ADJUST modules make no errors of commission; that is, no segment is assigned incorrectly. Tertiary constraints (DELIMIT, COMBINATORICS, OUTLIER) label an additional 17% (17 segments). Thus 67% (66 segments) of the secondary structure in the 10 α/β proteins can be assigned unambiguously. There are no mistaken assignments in this subset. For the remaining 33% (32 segments) of the segments, multiple assignments remain unresolved. Most of these are α vs. null or β vs. null choices.

Given the unresolved assignments, we are frequently left with lists of possible secondary structures. In general, the length of the list is small and the rank-ordered position of the native structure is high on the list (Tables III and IV). To serve as input to a combinatorial secondary-structure packing algorithm [e.g., Cohen et al. (1982)], the quality of the input secondary-structure assignments is most important. We can check this quality by examining, for all domains, the structures that most closely approximate the native (Table V). All of the 58 β-strands are correctly found, and 48 of the 53 helices are located. Missing are helices 304–311 in ADH, 69–84 in ADK, 107–119 in RHDa, 183–189 in RHDb, and 213–223 in TIM.

The problematic helices in ADH and TIM are marked by the A and P patterns (Table I), but since neither contains the H or C pattern, the algorithm does not consider them as possible helices. Both helices pack against aromatic residues in the native structure, the ADH helix against a tryptophan from the β-sheet and the TIM helix against a tyrosine and phenylalanine from the β-sheet. Aromatic residues are rarely observed in the β-sheets of α/β parallel proteins (Chothia & Janin, 1980).

The ADK helix is labeled as a possible helix. Since its neighboring segment is labeled as a definite helix and connected helices are not allowed, the 69–84 helix is excluded. In the crystal structure, 53–63 is a peripheral (noncore) helix.

The RHDa helix 107–119 is marked with the H and A patterns but fails to have a hydrophilic stripe (P) or favorable charged pair (C). This helix resides in the domain–domain interface and is appropriately hydrophobic on both faces. This sequence is atypical of helices in α/β proteins. The RHDb helix 183–189 is marked only with a hydrophilic stripe (P) and shows none of the hydrophobic or charge characteristics frequently seen in helices that pack against β-sheets. In fact, the observed packing of this helix against the sheet is very marginal. When the distinction between edge and internal β-strands is made, all but one (RHDb 207–210) of the 18 edge strands are correctly identified. The quality of these assignments, calculated as the percent true positive (α or β predicted and observed) plus percent true negative (α or β not predicted and not observed), is 83% α, 93% β, 97% ϵ, 74% coil, and 87% overall. This compares favorably with previous prediction schemes. More importantly for our goal, none of these omissions is so drastic as to prevent the secondary structure packing algorithms from producing an approximate tertiary structure.

One important result of this work is that the secondary structure assignment algorithm is able to distinguish between sequences compatible with the α/β parallel motif and those that are not. We indicate the basis for discrimination for each of the 13 non-α/β domains in Table IV. Two of these cases, TRX and CPV, deserve further exposition. TRX is an α/β

mixed protein, containing a five-stranded β-sheet with both parallel and antiparallel strands covered by α-helices on both faces. The algorithm finds β-strands at 2–8, 26–30, 52–60, 75–82, and 86–95 and α-helices at 37–49, 68–74, and 97–108. The crystal structure (Holmgren et al., 1975) has β-strands at 2–8, 22–29, 53–60, 77–83, and 88–91 and α-helices at 34–49, 62–66, and 95–107. Because of the short connecting segment between the end of the fourth β-strand [82 (predicted), 83 (native)] and the beginning of the last β-strand [86 (predicted), 88 (native)], a parallel connection is not possible and so no parallel α/β secondary-structure assignments are predicted. We have not as yet extended this algorithm to the more general class of mixed α/β proteins.

CPV, the all-α calcium-binding protein parvalbumin, presents another issue. Five secondary structure assignments, two with four β-strands and three with five β-strands, are produced. Initially, this would seem consistent with an α/β parallel protein. However, if the construction of a β-sheet consistent with the requirements outlined by Cohen et al. (1982) is attempted, no structure with the required anticomplementary hydrophobic patches on opposite faces of the β-sheet is possible. Such arrangements are possible for the all true α/β parallel proteins.

The fact that a collection of lists of possible secondary-structure assignments must be processed by a secondary-structure packing algorithm instead of one list is a source of concern. Fortunately, the different secondary-structure assignment lists can be arranged into families that differ from a parent list by the addition or deletion of one α-helix or β-strand. Had we not chosen the goal of distinguishing between edge and internal β-strands, the lengths of these collections of secondary structure assignments would be shorter (e.g., the β-pattern $\beta\beta\beta\beta\beta$ could be $\beta\epsilon\beta\beta\epsilon$ or $\epsilon\beta\epsilon\beta\beta$). However, since the distinction between internal and edge strands must be addressed before tertiary structure is predicted, it is artificial to ignore this problem at this level to improve the results in Table III. Additional methods for eliminating incorrect secondary-structure assignments from the complete set are being sought. It is useful to note that the success of this algorithm rests heavily on a model of tertiary-structure interactions. It may be unreasonable to expect a unique secondary-structure assignment in the absence of some tertiary-structure interactions. In the case of CPV, structures that looked consistent with the small set of properties of α/β parallel proteins contained in the algorithm were no longer tenable when additional tertiary constraints were added. The generality of this observation is under investigation.

We have ascertained that a tertiary structure that closely approximates the native structure can be generated from our correct secondary-structure assignments, for each domain examined. However, with current packing algorithms, incorrect secondary-structure assignments may also yield plausible structures.

Examples of Application of the Algorithm. Table VI shows the labeling of flavodoxin in the region of residues 44–86 with the patterns in Table I. This initial labeling produces three segments (45–57, 65–75, and 76–87) to which high order processing is applied. To illustrate the power and function of the various sets of rules that are applied to the segments thus labeled, we have shown the number of possible predicted complete secondary structures and the assignments for these three segments in Table VII.

Table VI: Pattern Labeling of a Region of Flavodoxin

residue	turns	α's	β's	edges	resultant labels
L-44	T1, T3				turn
N-45	T2, T3		B2	EB[a]	
E-46	T2, T3				
D-47					
I-48		H, A	B1		
L-49	T3	A	B1 BD		β
I-50		NA	B2, BD		
L-51			B1		
G-52 β		P			
C-53		A			
S-54			B1		
A-55	T3				
M-56	T3	P			
G-57	T2, T3	P			turn
D-58	T2, T3				
E-59	T3				
V-60	T3	P			
L-61	T3	P			
E-62	T2, T3	P			
E-63	T2, T3	P			turn
S-64	T2, T3				
E-65	T2, T3	P			
F-66	T2, T3	H, A, P			
E-67	T3	P	B1	EB	β
P-68	T3		B1	EB	
F-69	T3	H, P			
I-70 α	T3	A, P			
E-71	T3				α
E-72	T2, T3	P		E	
I-73	T2, T3	H, A, P		E	
S-74	T2, T3	C, P			
T-75	T2, T3			E	
K-76	T2, T3	H, A, P			
I-77	T2, T3	H, A			
S-78	T2, T3	A	B2	EB	β
G-79	T2, T3		B2	EB	turn
K-80	T2, T3	H, A			
K-81	T3	H, A			
V-82	T3		B1	EB	β
A-83 β	T3		B1	EB	
L-84	T3	P			
F-85	T3	P			
G-86					

[a] EB, E pattern near a β-pattern.

Discussion

The procedure described here differs from the previous approaches to secondary-structure prediction in both methodology and emphasis. The difference in method lies primarily in the use of artificial-intelligence formalism and not the physical model or pattern descriptions that are a composite of much earlier work. Artificial-intelligence approaches have some attractive features for problems of this complexity. These include separation of the data or knowledge base (the patterns and combination rules) from the program itself and the clear connection between a particular assignment and the rule from which it arose. At this early stage, our subjective impression is that these techniques are powerful and extensible.

The overall aim of this work is the assignment of protein tertiary structure. Hence, we are concerned with the ability to identify and to assign the core secondary-structure elements rather than individual residues. The focus on core features means that we ignore noncore parts of the sequence. While this part of the procedure is done automatically—that is, the core/noncore decisions are made within the program—it results in the deliberate misassignment of the occasional noncore helix or β-strand as a null segment.

A second difference in emphasis is that we do not insist upon a single result for each sequence studied. Our point of view is that any given set of rules or facts may not contain sufficient information to make a particular decision. We prefer to retain the ambiguity of multiple assignment, rather than force a

Table VII: Three Segments in Flavodoxin during Rule Processing and Combinatorial Structure Generation

after rule set applied	no. of possible secondary structures for entire protein	remaining possible assignments[a]
initial state after labeling	20736	B, E, N, 45–57 A, B, E, N, 65–75 A, B, E, N, 76–87
assign definite non-null α's force neighbors to non-α	6912	B, E, N, 45–57 A, B, E, N, 65–75 B, E, N, 76–87
assign definite non-null β's or edges	3456	B, E, 45–57 A, B, E, N, 65–75 B, E, N, 76–87
choose α or null for some non-β segments	1152	B, E, 45–57 A, B, E, N, 65–75 B, E, N, 76–87
assign definite β to some non-α segments	288	B, E, 45–57 A, B, E, N, 65–75 B, E, N, 76–87
assign non-β to segments bordering pairs of definite β's	288	B, E, 45–57 A, B, E, N, 65–75 B, E, N, 76–87
force segments between definite non-null α's to non-α	288	B, E, 45–57 A, B, E, N, 65–75 B, E, N, 76–87
aaaign some definite nulls	216	B, E, 45–57 A, B, E, N, 65–75 B, E, N, 76–87
delimit previous assignments; delete those that produce structures having too few residues	128	B, E, 47–55 {A, N, 65–75 {E, N, 65–70 B, N, 81–86
combinatorial generation	4	B, E, 47–55 {A, N, 65–75 {E, N, 65–70 B, N, 81–86

[a] A, α; B, β; E, edge (β); N, null.

possibly incorrect assignment that would prevent successful operation of the tertiary-packing programs. Note that the unresolved assignment merely causes a longer list of tertiary possibilities that must be reduced by additional information as available.

These differences make comparison with earlier work quite difficult. If we compare results for only the core regions, our treatment has the following strengths: (1) Sequences associated with parallel α/β domains can be identified without error. (2) The turn algorithm is very accurate. (3) Every definite assignment (some 50% of the core segments) is correct. (4) A near native core assignment is always to be found in the output list. (5) When non-α/β proteins are examined, inconsistencies are found that prevent their fitting the α/β mold demanded by this procedure. (6) The accuracy of the segment assignment and identification of the starting and stopping points of each segment is good enough to generate nativelike core tertiary structures (rms deviations ca. 4 Å) with the Cohen–Sternberg procedures.

The weaknesses of our approach include the following: (1) The internal details of the noncore regions are poorly characterized; conventional methods are currently superior for the study of these parts of the proteins. (2) The patterns and rules developed here are specific to pure parallel α/β proteins. While we plan to study other protein classes in the future, at present the conventional methods of Chou & Fasman (1974) or Garnier et al. (1978) are more general.

Conclusions

This method for secondary-structure assignment offers several advantages over previous schemes. For the class of α/β parallel proteins, we are able to locate turn regions with a very high degree of accuracy and two-thirds of the secondary structure with certainty. Proteins that fit into this class can be distinguished from other proteins by the amino acid sequence. Most importantly, the output for this algorithm appears to be suitable input to the packing algorithms of Cohen et al. (1982), which produces a list of tertiary structures for α/β proteins, one of which resembles the native structure, solely from a consideration of sequence and secondary structure. One of the most pleasant features of this approach is the ability to see quickly the consequences of one's ideas about sequence patterns. The ability to formulate the many structural suggestions of protein chemists into a self-consistent system has proved to be an effective tool for α/β proteins and is likely to be readily extendible to other classes of proteins.

Acknowledgments

We thank Drs. R. L. Baldwin and P. S. Kim for discussions and a copy of their manuscript prior to publication. We are indebted to the Medical Information Science Department and the Computer Graphics Laboratory (RR01801) for the use of their facilities.

Registry No. Alcohol dehydrogenase, 9031-72-5; adenylate kinase, 9013-02-9; lactate dehydrogenase, 9001-60-9; phosphoglycerate kinase, 9001-83-6; rhodanese, 9026-04-4; subtilisin, 9014-01-1; triosephosphate isomerase, 9023-78-3; cytochrome b_5, 9035-39-6; chymotrypsin, 9004-07-3; carboxypeptidase A, 11075-17-5; cytochrome *c*, 9007-43-6; lysozyme, 9001-63-2; ribonuclease, 9001-99-4; superoxide dismutase, 9054-89-1; thermolysin, 9073-78-3.

References

Anfinsen, C. B., Haber, E., Sea, M., & White, F. H. (1961) *Proc. Natl. Acad. Sci. U.S.A. 47*, 1309–1314.

Banks, R. D., Blake, C. C. F., Evans, P. R., Haser, R., Rice, D. W., Hardy, G. W., Merrett, M., & Phillips, A. W. (1979) *Nature (London) 279*, 773–777.

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

Barr, A., & Feigenbaum, E. A. (1981) in *The Handbook of Artificial Intelligence*, Vol. 1, pp 190–199, Heuristech Press, Stanford, CA.

Bergsma, J., Hol, W. G. J., Jansonius, J. N., Kalk, K. H., Ploegman, J. H., & Smit, J. D. G. (1975) *J. Mol. Biol. 98*, 637–643.

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

Chothia, C., & Janin, J. (1980) *J. Mol. Biol. 143*, 95–128.

Chothia, C., Levitt, M., & Richardson, D. (1977) *Proc. Natl. Acad. Sci. U.S.A. 74*, 4130–4134.

Chou, P. Y., & Fasman, G. D. (1974) *Biochemistry 13*, 222–244.

Chou, P. Y., & Fasman, G. D. (1977) *J. Mol. Biol. 15*, 135–175.

Cohen, F. E., & Sternberg, M. J. E. (1980) *J. Mol. Biol. 137*, 9–22.

Cohen, F. E., Richmond, T. J., & Richards, F. M. (1979) *J. Mol. Biol. 132*, 275–288.

Cohen, F. E., Sternberg, M. J. E., & Taylor, W. R. (1980) *Nature (London) 285*, 378–382.

Cohen, F. E., Novotny, J., Sternberg, M. J. E., Campbell, D. G., & Williams, A. F. (1981a) *Biochem. J. 195*, 31–40.

Cohen, F. E., Sternberg, M. J. E., & Taylor, W. R. (1981b) *J. Mol. Biol. 148*, 253–272.

Cohen, F. E., Sternberg, M. J. E., & Taylor, W. R. (1982) *J. Mol. Biol. 156*, 821–862.
Colman, D. M., Jansonius, J. N., & Matthews, B. W. (1972) *J. Mol. Biol. 70*, 701–724.
Drenth, J., Jansonius, J. N., Koefer, R., & Wolthers, B. G. (1971) *Adv. Protein Chem. 25*, 79–115.
Eklund, H., Nordstrom, B., Zeppezauer, E., Solderlund, G., Ohlsson, I., Boiwe, T., Soderberg, B.-O., Tapia, O., Branden, 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.
Garnier, J., Osguthorpe, D. J., & Robson, B. (1978) *J. Mol. Biol. 120*, 97–120.
Hartsuck, J. A., & Lipscomb, W. N. (1971) *Enzymes, 3rd Ed.*, 1–56.
Holbrook, J. J., Liljas, A., Steindel, J., & Rossmann, M. G. (1975) *Enzymes, 3rd Ed.*, 191–292.
Holmgren, A., Soderberg, B.-O., Eklund, H., & Branden, C.-I. (1975) *Proc. Natl. Acad. Sci. U.S.A. 72*, 2305–2309.
Imoto, T., Johnson, L. N., North, A. C. T., Phillips, D. C., & Rupley, J. A. (1972) *Enzymes, 3rd Ed.*, 665–868.
Kernighan, B. W., & Ritchie, D. M. (1978) *The C Programming Language*, Prentice Hall, Englewood Cliffs, NJ.
Kretsinger, R. H., & Knockolds, C. E. (1973) *J. Biol. Chem. 248*, 3313–3326.
Kuntz, I. D. (1972) *J. Am. Chem. Soc. 94*, 4009–4012.
Levitt, M. (1976) *J. Mol. Biol. 104*, 59–116.
Levitt, M., & Chothia, C. (1976) *Nature (London) 261*, 552–557.
Lim, V. I. (1974) *J. Mol. Biol. 88*, 857–872.
Mathews, F. S., Argos, P., & Levine, M. (1971) *Cold Spring Harbor Symp. Quant. Biol. 36*, 387–395.
Momany, F. A., McGuire, R. F., Burgess, A. W., & Scheraga, H. A. (1975) *J. Phys. Chem. 79*, 2361–2381.
Nagano, K. (1973) *J. Mol. Biol. 75*, 401–420.
Nemethy, G., & Scheraga, H. A. (1977) *Q. Rev. Biophys. 10*, 239–352.
Richards, F. M., & Wyckoff, H. W. (1971) *Enzymes, 3rd Ed.*, 647–806.
Richards, F. M., & Richmond, T. J. (1978) in *Molecular Interactions and Activity in Proteins, Ciba Foundation Symposium*, pp 23–45, Exerpta Medica, Amsterdam.
Richardson, J. S. (1981) *Adv. Protein Chem. 34*, 167–339.
Richardson, J. S., Richardson, D. C., Thomas, K. A., Silverton, E., & Davies, D. R. (1976) *J. Mol. Biol. 102*, 221–235.
Richmond, T. J., & Richards, F. M. (1978) *J. Mol. Biol. 119*, 537–555.
Robson, B., & Osguthorpe, D. J. (1979) *J. Mol. Biol. 132*, 19–51.
Rose, G. D. (1978) *Nature (London) 272*, 586–590.
Rose, G. D. (1979) *J. Mol. Biol. 134*, 447–470.
Schiffer, M., & Edmundson, A. B. (1967) *Biophys. J. 7*, 121–135.
Schiffer, M., & Edmundson, A. B. (1968) *Biophys. J. 8*, 29–39.
Schulz, G. E., & Schirmer, R. H. (1979) *Principles of Protein Structure*, Springer-Verlag, New York.
Schulz, G. E., Elzinga, M., Marx, F., & Schirmer, R. H. (1974) *Nature (London) 250*, 120–123.
Segal, D. M., Cohen, G. H., Davies, D. R., Powers, J. C., & Wilcox, P. E. (1972) *Cold Spring Harbor Symp. Quant. Biol. 36*, 85–90.
Sternberg, M. J. E., & Cohen, F. E. (1982) *Int. J. Biol. Macromol. 4*, 137–144.
Sternberg, M. J. E., Cohen, F. E., Taylor, W. R., & Feldman, R. J. (1982) *Philos. Trans. R. Soc. London, Ser. B 239*, 177–189.
Takano, T. (1977) *J. Mol. Biol. 110*, 569–584.
Takano, T., Kallai, O. B., Swanson, R., & Dickerson, R. E. (1973) *J. Biol. Chem. 248*, 5234–5255.
Taylor, W. R., & Thornton, J. M. (1983) *Nature (London) 301*, 540–542.
Venkatachalam, C. M. (1968) *Biopolymers 6*, 1425–1436.