Struck, D. K., Hoekstra, D., & Pagano, R. E. (1981) *Biochemistry 20*, 4093-4099.

Tycko, B., & Maxfield, F. R. (1982) *Cell (Cambridge, Mass.) 28*, 643-651.

Van Renswoude, J., Bridges, K. R., Harford, J. B., & Klausner, R. D. (1982) *Proc. Natl. Acad. Sci. U.S.A. 79*, 6186-6190.

White, J., & Helenius, A. (1980) *Proc. Natl. Acad. Sci. U.S.A. 77*, 3273-3277.

White, J., Helenius, A., & Kartenbeck, J. (1982a) *EMBO J. 1*, 217-222.

White, J., Helenius, A., & Gething, M. J. (1982b) *Nature (London) 300*, 658-659.

White, J., Kielian, M., & Helenius, A. (1983) *Q. Rev. Biophys. 16*, 151-195.

Wilschut, J., Düzgünes, N., Fraley, R., & Papahadjopoulos, D. (1980) *Biochemistry 19*, 6011-6021.

Wilschut, J., Nir, S., Scholma, J., & Hoekstra, D. (1985) *Biochemistry 24*, 4630-4636.

Wilson, I. A., Skehel, J. J., & Wiley, D. C. (1981) *Nature (London) 289*, 366-373.

Yewdell, J. W., Gerhard, W., & Bach, T. (1983) *J. Virol. 48*, 239-248.

Yoshimura, A., Kuroda, K., Kawasaki, K., Yamashina, S., Maeda, T., & Ohnishi, S. I. (1982) *J. Virol. 43*, 284-293.

Turn Prediction in Proteins Using a Pattern-Matching Approach[†]

F. E. Cohen,[‡] R. M. Abarbanel,[§,||] I. D. Kuntz,[*,‡] and R. J. Fletterick[⊥]

Departments of Pharmaceutical Chemistry, Medical Information Science, and Biochemistry, University of California at San Francisco, San Francisco, California 94143

Received May 2, 1985; Revised Manuscript Received August 20, 1985

ABSTRACT: We extend the use of amino acid sequence patterns [Cohen, F. E., Abarbanel, R. M., Kuntz, I. D., & Fletterick, R. J. (1983) *Biochemistry 22*, 4894–4904] to the identification of turns in globular proteins. The approach uses a conservative strategy, combined with a hierarchical search (strongest patterns first) and length-dependent masking, to achieve high accuracy (95%) on a test set of proteins of known structure. Applying the same procedure to homologous families gives a 90% success rate. Straightforward changes are suggested to improve the predictive power. The computer program, written in Lisp, provides a general pattern-recognition language well suited for a number of investigations of protein and nucleic acid sequences.

The recognition by Anfinsen et al. (1961) that amino acid sequence determines protein structure initiated a search for algorithms to predict protein tertiary structure from primary sequence. Two basic theoretical approaches to this problem have developed: energy minimization and semiempirical hierarchical condensation models. Energy calculations offer the advantage of a chemically plausible approach to structure prediction. This method is limited by difficulties in producing adequate energy functions and by problems with convergence [e.g., Momany et al. (1975), Levitt (1976), and Robson & Osguthorpe (1979)]. The semiempirical hierarchical condensation model assumes that the folding problem can be divided into a series of smaller problems. Traditional divisions have been the prediction of secondary structure from sequence [e.g., Chou & Fasman (1974), Lim (1974), Garnier et al. (1978), Taylor & Thornton (1983), and Cohen et al. (1983)], the prediction of approximate tertiary structure from secondary structure (Cohen et al., 1979, 1980, 1982; Ptitsyn & Rashin, 1975), and the refinement of approximate tertiary structure.

In a previous paper, we showed that the successful location of turns facilitated secondary structure assignment for the class of proteins that are formed from a β sheet surrounded by α helices (Cohen et al., 1983). This paper describes a more general approach to the location of turns in proteins of all three major classes of globular proteins: α/α, α/β, and β/β (Levitt & Chothia, 1976). The new algorithms make use of many of the tools of artificial intelligence "expert systems" (Barr & Feigenbaum, 1981) described in our earlier paper. A major advance is the creation of a pattern-matching language, Pattern Language for Amino and Nucleic Acid Sequences (PLANS), written in Lisp. Other computer languages could be employed, but Lisp appears best suited for further development. These algorithms accurately (>90%) locate turns in a large set of proteins. The complete set of patterns and an interpretation guide are presented in this report. Knowledge of turn location is of interest to biochemists and immunologists as a guide to structural or biochemical properties of proteins. Further, we expect that the new programs will ultimately lead to a reasonable, complete, and accurate method of secondary structure prediction.

Theory and Methods

We will consider a subset of globular proteins that consist of one or more sequentially contiguous domains where each domain is a member of one of three classes of protein structure: α/α, α/β, or β/β. The proteins of interest include at least half of the known structures. We treat the protein backbone as a series of regular helical and/or strand conformations connected by turns. Turns for our purposes are defined as regions of the chain between regular secondary structure elements; they may be from two to more than fifteen amino acids in length. They are frequently made from amino acids that are relatively accessible to solvent and hydrophilic in character. These assumptions constitute a simple model for domains in globular proteins.

A *hierarchical* approach is used in the algorithms we use to find turns. We wish to avoid the difficulties associated with "cutoff" or "threshold" parameters. Intuition and experience

[†] This research was supported by Grants 5R23LM03893 (R.M.A.), GM34197 (I.D.K.), and AM26081 (R.J.F.).

[‡] Department of Pharmaceutical Chemistry.

[§] Department of Medical Information Science.

[||] Present address: Intellicorp, Mountain View, CA 94040.

[⊥] Department of Biochemistry.

0006-2960/86/0425-0266$01.50/0

suggest that some combination of over- and underprediction is unavoidable [e.g., Chou & Fasman (1974) and Garnier et al. (1978)]. Hierarchical schemes offer a solution to this dilemma. Turns may be selected by a variety of criteria, each of which is optimized to avoid overprediction. The most reliable characteristics are identified first, and these turns are assigned. Weaker characteristics are identified in subsequent analysis of the sequence.

Definition of Terms. The strongest criterion is the hydrophilicity of amino acid chains in turns leading to a set of "patterns" of amino acid sequences common to all classes of proteins studied here. An additional set of patterns is developed that recognizes the specific physical requirements of particular secondary structural features. Certain side chains may be in turns or not depending on their neighbors. These are termed "*swing residues*". They can play different roles in different environments and will vary with protein class (see below).

When a pattern is found in a particular protein sequence, it is labeled and *masked* from further consideration. This masking is dependent on the characteristic *link length* of each protein class. Each of these concepts is amplified below.

The physical principles that are used to develop the *individual patterns* fall into four groups. These patterns are based on the following: local maxima in hydrophilicity; secondary structure identification and avoidance; regions containing proline; weakly hydrophobic segments sequentially distant from well-defined turns. The above order defines the hierarchical search. We consider each of these ideas in more detail to motivate the actual patterns we have used.

Subsequences rich in hydrophilic residues have been identified as turns by many investigators (Kuntz, 1972; Rose, 1978; Kyte & Doolittle, 1982). Cohen et al. (1983) employed a pattern-based version of this notion. We note, however, that there are certain sequences that are rich in hydrophilic side chains that are not turns; for example, glycine residues flanked by pairs of hydrophilic residues can play a special role in α/α proteins; see Table III. Such "glycine sites" will be discussed further.

Turns are found in regions between secondary structural elements. Thus a secondary structure identification algorithm is an obvious approach to turn location. Residues sequentially arrayed to favor secondary structure are, by definition, not in turns. Moreover, since domains in globular proteins have diameters roughly related to the molecular weight, turns must be within a fixed number of residues of the center of secondary structure feature. Thus, some patterns match a high density of hydrophobic residues that corresponds to the center of an internal β strand. Since the typical β strand is no more than nine residues in length, turns are sought five residues before and five after such a site. In a variant of this approach, in the α/α set, the termini of α helices can sometimes be found by a nonhelical alternation of hydrophobic and hydrophilic residues. Turns are located between adjacent sets of these termini.

Proline occupies a special role in turns. It is given a high turn propensity in empirical prediction schemes (Lewis et al., 1971; Kuntz, 1972; Chou & Fasman, 1974, 1977). However, proline is occasionally observed in "kinks" in the middle of α helices or in "bulges" in β strands (Richardson et al., 1978). In helical domains, we find that prolines are not in turns when strong hydrophilic turns are among neighboring residues on the N-terminal side. We speculate that the longer helix with a kink at the proline is favored over two short helices divided by an internal turn. The occurrence of proline in β bulges limits its usefulness in turn identification in the β/β class.

Table I: Link Lengths and Protein Classes

class	approx domain size	r^a	pitch (Å/residue)	link length		used^e
				Å^b	residues^c	
α/α	150	20.6	1.5	33.2	22.1	22
α/β	200	22.7	2.25^d	37.4	16.6	14
β/β	100	18.0	3.0	28.0	9.3	8

[a] Computed from $(3 \times 110n/0.602 \times 4\pi\bar{v})^{1/3}$, where n is the number of residues per domain and $\bar{v}$ is the partial specific volume (0.75 cm^3/g). [b] Computed as $2(r - 4)$ Å, where 4 Å is taken as the width of a turn. [c] Computed as $(2r - 8\text{Å})/\text{pitch}$. [d] Assumed to be the average of the pitch for α and β structures. [e] To obtain the average length between turns, add 4 residues to the value used for the secondary structure length. Only even values appear in this column, as $^1/_2$ of this value appears as a natural parameter in many pattern expressions and nonintegral values are not tolerated. In Cohen et al. (1983), a value of 14 ± 4 was used for α/β turn separation equivalent to the value chosen here.

Weakly hydrophilic segments in the sequence are also candidates for turns. Although not all such "weak spots" are turns, those sequentially distant from other "stronger" turns frequently are. Specific patterns of amino acids with reduced requirements for local hydrophilicity are searched for at the desired turn spacings.

The term *swing residue* was introduced in a previous paper to suggest that certain residues occupy both hydrophobic and hydrophilic roles. Swing residues are not a fixed set of amino acids. Instead, they vary with protein class. We have learned from the present analysis that they are GKPSTY in β/β, AKY in α/α, and KY in α/β. Swing residues generally occur in physical locations where a small change in torsional bond angles of the main chain or side chain can move a side chain from a buried to an exposed position without altering the surrounding protein architecture. In β/β proteins, each strand has at least two such positions, one at either end (Cohen et al., 1981). When β bulges are included, the number of potential positions for swing residues increases by two per bulge. In α/β proteins, the side-chain packing tends to be more regular and β bulges are rarely observed (Cohen et al., 1982). Whatever the precise structural, kinetic, or evolutionary reasons for the existence of swing residues, it is clear, empirically, that these residues can fit both hydrophobic and hydrophilic environments.

In summary, we find that the local maxima in hydrophilicity identify 50% of the turns in the domains studied here. Patterns relating to proline and the avoidance of secondary structure account for another 30%, and weakly hydrophilic regions that are sequentially distant from well-defined turns account for about 10% more. These categories are made mutually exclusive through the use of sequence *masking*.

Masking means that certain regions of the chain are removed from consideration as possible turns. Regions of a sequence are masked because they are considered likely to be α helices or β strands or because a turn has already been located in their immediate vicinity. The length of sequence that is masked follows from the rules presented below and in Table I.

In previous work on α/β proeins (Cohen et al., 1983), we noted that a knowledge of the expected spacing between turns offered an excellent means for exploiting strongly signaled turns to aid in the location of more weakly signaled turns. Though the path of the polypeptide chain is complex, the important turns occur at fairly regular spatial and sequential intervals. To approximate the average length of secondary structure or *link length*, the following calculation is useful. The average domain size for each protein class is known empirically. Precision is not needed since the link length varies as the $^{1}/_{3}$ power of the domain size. If the proteins are assumed to be spherical and the turns taken to occupy the outermost 2 Å of the sphere, then link length as a function of protein class can be quickly calculated (see Table I). For example, in an α/α protein, if residues 20 and 40 were identified as strong turns, the program would not allow additional turns within the 18 residues from 21 to 39 as the link length for α helices is 22 residues. If the same strong turns were found in a β/β protein, an attempt to find a weaker turn pattern in between residues 28 and 32 would be made. This avoidance or masking of the sequence close to strong turns improves the accuracy of the algorithm.

We note that some proteins show large axial ratios and occasionally long α helices. This is the case in influenza hemagglutinin (Wilson et al., 1981). Such structures would be outside the limits of our current models, but they could be included by developing rules for subclasses.

Algorithm. We have developed a complete pattern-matching language to facilitate this work called PLANS (Abarbanel, 1984). We had four basic requirements: (1) pattern specification must be flexible, readable, and general; (2) patterns might be defined in terms of other patterns; (3) the patterns must be easily created and edited by a user; (4) patterns might contain the names of other patterns. The future development of rule-based inductive and deductive reasoning should be able to use these patterns for making inferences about structure. Though we will discuss patterns that have been applied to protein sequences, there is nothing inherent to the PLANS system to prevent it from being applied to DNA or RNA sequences.

The usual languages for scientific programming, Fortran, C, and Pascal, are not well suited for these goals. Lisp, however, has a history of use in implementation of "languages" like PLANS, and software packages to support rule-based systems are available. The entire pattern-matching system described in the 1983 α/β secondary assignment paper (Cohen et al., 1983) is now supported in Franz Lisp (Foderaro et al., 1983) on a VAX 11/750 running UNIX (TM, Bell Telephone Laboratories).

Lisp has the advantage of being a recursive language. If a pattern is composed of other subpatterns, as in a logical expression, those subparts are available for evaluation, inspection, and display. In this way, patterns may be built up from basic units that might correspond to structural entities in the proteins themselves. Another advantage provided by Lisp is the ability to examine the pattern-matching process at arbitrary points in the computation.

A number of extensions have been made to the earlier work. These include specification of densities and partial matches and the attachment of rules to patterns made possible in the Lisp environment.

We have made use of a new type of data structure called a Flavor (Copyright 1981 Massachusetts Institute of Technology). Flavors allow functions or methods to be grouped in powerful ways. Each pattern and each sequence is an instance of a Flavor. The data structure associated with a *pattern* Flavor includes such things as its name and definition, the associated descriptive comments, and the results of matches of this pattern with sequences. Each *pattern* Flavor can have its own methods for evaluating that pattern type. Some patterns use combinations of methods provided by other Flavors. If a given pattern type is defined in terms of another, then the Flavor system provides automatically for combining their evaluation methods. The several forms of "method combination" provided by the Flavors system allow very simple programming of the evaluation of patterns against sequences.

Table II: Symbols Used in Pattern Specifications

special symbol	meaning
^	beginning of sequence
$	end of sequence
*	ZERO or more repeats of the preceding symbol, equivalent to %0,*
%*n*,*m*%	between *n* and *m* repeats of previous symbol or parenthesized set of symbols; *m* may be "*" to indicate *n* or more repetitions
[ ]	logical OR of symbols in brackets, as in [ABC]
-	"through", used in [...] to indicate a range of values; e.g., [A-CG-K] means [ABCGHIJK]
{*m*,*u*}	spreading of previous symbol that hits at position *i*, to all sites between $i + m$ and $i + u$
( )	used for grouping of characters for repetitions or logical combinations of pattern expressions involving AND, OR, or NOT
group	explained in the text
density	explained in the text

Pattern Language. Not only must a general pattern language allow direct residue-for-residue matches, but it must also allow residues to be grouped in various ways. Table II explains the special characters used in pattern definitions. All patterns also have an associated *offset*, a integer that is added to the actual sequence location of a match. This is useful for bringing related patterns together to obtain new patterns. Several examples of patterns from Table III are discussed below.

The simplest pattern is just a quoted string that must match exactly, for example

Proline -4 "P"

A sequence will be marked with pattern **Proline** at offset −4 residues, that is, four residues before each P residue.

PLANS allows for single-residue variability at a given site. These residues may be specified to be either IN a restricted set, NOT IN a set, or completely free. More than one residue may take the role of "P" in the pattern above by using the notation "[...]". For example

alpha_strong_phobic 0 "[ACFILMVW]"

would match a single residue at any of residues in the group "ACFILMVW" with zero offset.

The "." character may be used to represent any residue and is therefore a place marker for specific separations, or where any residue may match. For example

charge_a 2 "[DE]...[HKR]"

will mark a site two residues after the D or E whenever there is an H, K, or R at residue 4 downstream. The *offset* of 2 here places the "hit" location for this pattern at the center of the charged pair. In an α helix, this would constitute a reasonable stabilizing charge pair (Kim & Baldwin, 1984).

Logical operators are provided to allow one to specify that more than one pattern match at a given site (AND), that at least one pattern match there (OR), or that some pattern not match (NOT). Expressions using these connective words are evaluated by using the matches of the patterns included in the expression. Expressions may also include parentheses.

An example of the use of logical connectives and parentheses is

gly-ala-site 0 ((**gly-ala-site1** OR **gly-ala-site2** OR
gly-ala-site3) AND NOT **many_ala**)

which would match wherever any one or more of the three

Table III: α/α Turn Patterns

Pattern Name	Offset	Pattern
gly-ala-site1	4	"[ACFIKLMTVWY][ABDEGHKNPQRSTZ]%2,2 [AG][ABDEGHKNPQRSTZ]%2,2[ACFIKLMTVWY]"
gly-ala-site2	5	"[ACFIKLMTVWY]%2,2[ABDEGHKNPQRSTZ] [ABDEGHKNPQRSTZ][AG][ABDEGHKNPQRSTZ]%2,2 .[ACFIKLMTVWY]"
gly-ala-site3	5	"[ACFIKLMTVWY].[ABDEGHKNPQRSTZ]%2,2 [AG][ABDEGHKNPQRSTZ]%2,2[ACFIKLMTVWY]%2,2" CLUSTER OF HYDROPHILIC RESIDUES BOUNDED BY HYDROPHOBIC RESIDUES, CENTERED ON A GLYCINE OR ALANINE. TYPICAL OF HELICAL AREAS PARTICIPATING IN TIGHTLY PACKED HELIX-HELIX INTERACTIONS.
many_ala	5	(density(>=,3,9,1,"A")) ALTHOUGH ALANINE MAY OCCUPY MANY POSITIONS IN THIS PATTERN, WHEN 3 OR MORE ARE SEEN; THE GEOMETRIC ASSYMMETRY SUGGESTS THAT HELIX-HELIX PACKING IS UNLIKELY. THIS PATTERN MATCHES WHEREVER THERE ARE GREATER THAN 2 ALANINES OUT OF 9 RESIDUES.
gly-ala-site	0	((gly-ala-site1 or gly-ala-site2 or gly-ala-site3) and not many_ala) A GLY-ALA TYPE HELICAL SITE WITHOUT TOO MANY ALANINES.
Proline	-4	"P" A PROLINE RESIDUE WITH OFFSET -4.
alpha_philic	0	"[BDEGKNQRSZ]" HYDROPHILIC RESIDUES.
alpha_begin	-3	"[BDEGHKNQRSZ][ACFIKLMPTVWY][BDEGHKNQRSZ] [ACFIKLMPTVWY]" PATTERN FREQUENTLY SEEN AT THE BEGINNING OF α-HELICES.
alpha_end	0	"[ACFIKLMPTVWY][BDEGHKNQRSZ][ACFIKLMPTVWY] [BDEGHKNQRSZ]" PATTERN FREQUENTLY SEEN AT THE END OF α-HELICES.
alpha_phobic	0	"[ACFIKLMPTVWY]" HYDROPHOBIC RESIDUES.
alpha_strong_phobic	0	"[ACFILMVW]" STRONG HYDROPHOBICS.
charge_a	2	"[DE]...[HKR]" CHARGE INTERACTION FAVORED BY HELICAL GEOMETRY.
charge_b	1	"[DE]..[HKR]" SAME WITH DIFFERENT OFFSET AND SEPARATION OF CHARGES. (KIM AND BALDWIN 1984)
charge_c	2	"[HKR]...[DE]" SAME AS "CHARGE_A" BUT NOT FAVORED BY HELICAL GEOMETRY.
alpha1charges	4	((charge_c{1,1} and charge_a{2,2}) or (charge_c{0,0} and charge_a{3,3}) or (charge_c{0,0} and charge_a{4,4}) or (charge_c{-1,1} and charge_a{4,4}))
alpha2charges	2	((charge_a{1,1} and charge_c{2,2}) or (charge_a{0,0} and charge_c{3,3}) or (charge_a{0,0} and charge_c{4,4}) or (charge_a{-1,1} and charge_c{4,4})) TWO HELICAL PATTERNS WHICH FAVORABLY ARRAYS CHARGED RESIDUES ALONG A HELICAL FACE EITHER +...-...+ OR -...+...-
AA	0	(gly-ala-site or alpha1charges or alpha2charges) REGIONS LIKELY TO BE HELICAL WITH SPECIAL PATTERNS OF CHARGED OR HYDROPHILIC RESIDUES WHICH ARE SUBSEQUENTLY MASKED TO PREVENT THEIR ASSIGNMENT AS TURN REGIONS.
alpha_phob	1	(not density(>=,2,3,1,alpha_philic)) REGIONS UNLIKELY TO BE GOOD TURNS.
AA_Turn1_no-phobics	2	(density(=,0,5,1,alpha_strong_phobic) and not charge_a{-2,2} and not charge_b{-2,2} and not gly-ala-site) FIVE RESIDUES IN SEQUENCE WITHOUT ANY STRONG HYDROPHOBICS. THIS PATTERN AVOIDS KIM/BALDWIN CHARGE PAIRS AND GLYCINE INTERACTION SITES.
AA_T_philic	1	(density(=,4,4,1,alpha_philic)) STRONG TURN WITH FOUR HYDROPHILIC RESIDUES IN SEQUENCE.
AA_Turn2_4-philics	0	(T_philic and not AA{-5,5} and not AA_Turn1_no-phobics{-11,11}) WEAKER TURN PATTERNS WITH APPROPRIATE MASKING OF WHAT HAS COME BEFORE.
AA_Turn3_proline	0	(Proline and not AA_Turn1_no-phobics{-11,0} and not AA_Turn2_4-philics{-11,0}) PROLINE SITE STRONGLY SUGGESTIVE OF TURN IN α/α PROTEINS.
AA_Turn4_helix-ends	1	(alpha_end and alpha_begin{-2,0} and not (AA_Turn1_no-phobics{-11,11} or AA_Turn2_4-philics{-11,11} or AA_Turn3_proline)) TURN SUGGESTED BY ADJACENT HELICAL BEGINNING AND END IN REGION AWAY FROM REGULAR TURN.
AA_Turn5_weak	0	(not AA{-5,5} and not AA_T_possible{-11,11} and not alpha_phob{-1,1}) NOT A TYPICAL CENTRAL HELIX OR PHOBIC, MASKING OUT MORE LIKELY AA_T_POSSIBLE TURN PATTERN
AA_Turn5_group	0	(group(AA_Turn5_weak,7)) Merged AA_Turn5_weak hits for up to 7 adjacent hits.
AA_T_possible	0	(group(AA_Turn1_no-phobics,7) or group(AA_Turn2_4-philics,7) or group(AA_Turn4_helix-ends,7) or group(AA_Turn3_proline,7)) GROUPED POSSIBLE TURN SITES.
AlphaTurns	0	(AA_T_possible or AA_turn5_group) FINAL CONSENSUS TURN PATTERN.

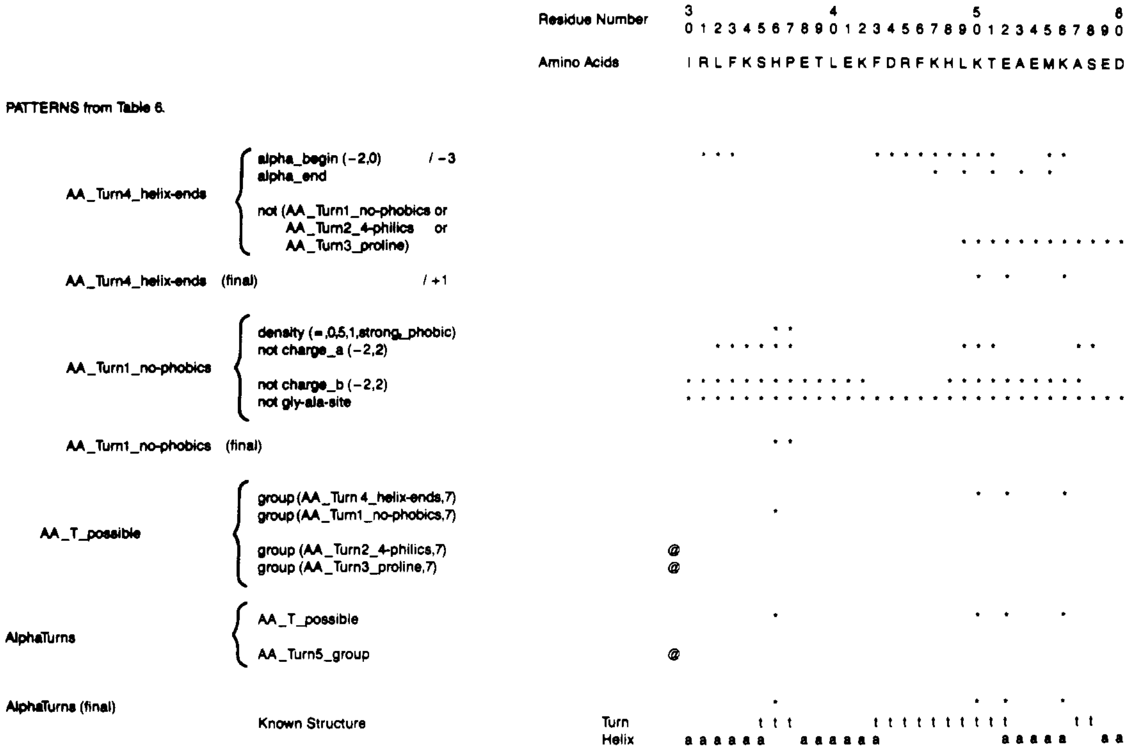

FIGURE 1: Myoglobin residues 30–70: assignment of turns. @ indicates no matches in this region.

OR'd patterns match, as long as the **many_ala** pattern is not also found there. PLANS both notes the match with pattern **gly-ala-site** and remembers which of the four subpatterns was found.

It is convenient to define densities of matches within the pattern syntax. One can specify an exact number (using =) or some relational operator (>, >=, <, <=, !=) applied to a number of patterns. For example

many_ala 5 (density(>=,3,9,1,"A"))

would match wherever the sequence contained at least three alanines out of nine contiguous residues. The "A" here could be replaced by the name of any other pattern.

PLANS also provides "merging" and "spreading" functions. The "group" function merges up to a specified number of matches into a single mark. In this way, local signals spread over adjacent residues may be combined in a single hit. For example

AA_Turn5_group 0 (group (**AA_Turn5_weak,7**))

will merge up to seven **AA_Turn5_weak** matches that are adjacent. These groups will be marked at the beginning of such a group in sequences matched.

Spreading allows the location of a match to be marked over a number of adjacent residues. This allows patterns to act as masks so that regions of a sequence may be effectively excluded from matching during the evaluation of another, possibly less specific, pattern. This *masking* idea is used powerfully in the turn assignment patterns where a precedence of turn types is brought about by each stronger or more definite type of turn masking a region that then will not be hit by weaker patterns.

The symbol for spreading is incorporated in the pattern as "{*m*,*n*}"

AA_Turn2_4-philics 0(**AA_T_philic** AND NOT AA {-5,5} AND NOT **AA_Turn1_no-phobics**{-11,11})

In this case, matches will be found where the pattern **AA_T_philic** is found provided that those matches are NOT also within five residues of a match with pattern **AA**, AND also NOT within eleven residues of a match with pattern **AA_Turn1_no-phobics**.

Since patterns are defined in terms of each other, PLANS must keep track of related patterns and note any missing pattern definitions. All of this information is kept in the *pattern* Flavor objects. In addition, these objects keep the results of previous matches so that a given pattern needs only to be run once on a sequence. Any pattern defined in terms of another will cause the other pattern(s) to be evaluated. This process may be monitored so that complex matches may be decomposed and understood.

Further details about the pattern-matching evaluation methods and the use of Flavors will be published elsewhere. A complete example is worked out in Figure 1 for the turns in part of the myoglobin sequence.

Turns for Each Type of Domain. Table III contains the complete list of α/α turn patterns. The hierarchy or order of precedence of turn expressions is **AA_Turn1_no-phobics**, **AA_Turn2_4-philics**, **AA_Turn3_proline**, **AA_Turn4_helix-ends**, and **AA_Turn5_weak**. **AlphaTurns** is the aggregate final turn expression. Each of the turn expressions is a combination of patterns, which are also listed in the table. Many of these terms are defined by other terms, possibly by using the spread notation and masks. As an example, see the pattern **AA_Turn2_4-philics** explained above. The "**not AA**" part of this expression avoids **gly-Ala-site**, **alpha1charges**, and **alpha2charges** patterns. The "**not AA_Turn1_no-phobics**" logical expression prevents looking for turns in regions where a stronger turn has already been noted. By continued substitution of terms, one finally arrives at a list of specific amino acid sequences. This list is then checked against the sequence under study.

See Table IV for a similar hierarchy used in α/β proteins. The patterns for β/β proteins are specified in Table V.

RESULTS

The patterns described in Tables III–V were developed on a test set consisting of 7 α/α proteins with 45 turns, 11 β/β

Table IV: α/β Turn Patterns

```
Pattern Name     Offset   Pattern

yphil1              0 "[BDEGHKNPQRSTYZ][BDEGHKNPQRSTZ]%3,3"
yphil2              0 "[BDEGHKNPQRSTZ][BDEGHKNPQRSTYZ][BDLGHKNPQRSTZ]%2,2"
yphil3              0 "[BDEGHKNPQRSTZ]%2,2[BDEGHKNPQRSTYZ][BDEGHKNPQRSTZ]"
yphil4              0 "[BDEGHKNPQRSTZ]%3,3[BDEGHKNPQRSTYZ]"

yphilic             1 (yphil1 or yphil2 or yphil3 or yphil4)
         4 HYDROPHILIC RESIDUES WITH AT MOST 1 SWING RESIDUE Y.

5of7                3 (density(>=,5,7,1,"[BDEGHKNPQRSTZ]"))
         5 OF 7 RESIDUES ARE HYDROPHILIC.

charge_a            2 "[DE]...[KR]"   CHARGE PAIR WITH FAVORABLE INTERACTION
charge_b            1 "[DE]..[KR]"    IN THE KIM/BALDWIN (1984) SENSE.

charge_c            2 "[KR]...[DE]"   UNFAVORABLE INTERACTION.
charge_d            2 "[KR]..[DE]"

charges             0 (charge_a or charge_b or charge_c or charge_d)

beta_middle         2 (density(>=,4,5,1,"[CFILMVW]"))
         HIGH DENSITY OF HYDROPHOBIC RESIDUE PATHOGNOMONIC FOR
         AN INTERNAL  β-STRAND.

ab_phob             1 (density(>=,2,3,1,"[ACFILMTVW]"))
         PATTERN NOT SUITABLE FOR TURNS.


AB_Turn1_4-philics 0 (group(yphilic,5))
         MOST DEFINITE TURNS.

AB_Turn2_Proline   0 ("P" and not AB_Turn1_4-philics{-7,7})
         PROLINES MASKING TURNS ABOVE.

AB_Turn3_5of7       0 (5of7
                       and not charges{-2,2}
                       and not AB_Turn1_4-philics{-7,7}
                       and not AB_Turn2_Proline{-7,7})
AB_Turn3_group     0 (group(AB_Turn3_5of7,5))
         GROUPED REGIONS WITH 5 OF 7 RESIDUES HYDROPHILIC,
         MASKING TURNS ABOVE.

AB_Turn4_not-beta  0 ((beta_middle{7,7} or beta_middle{-3,-3})
                       and not AB_Turn1_4-philics{-7,7}
                       and not AB_Turn2_Proline{-7,7}
                       and not AB_Turn3_group{-7,7})
         BORDERS OF SEQUENCE MOSTLY LIKELY A MIDDLE OF  β-STRAND.

AB_Turn5_weak       1 (     not AB_Turn1_4-philics{-7,7}
                       and not AB_Turn2_Proline{-7,7}
                       and not AB_Turn3_group{-7,7}
                       and not AB_Turn4_not-beta{-7,7}
                       and not ab_phob{0,2})
AB_Turn5_group     0 (group(AB_Turn5_weak,5))
         GROUPED WEAK HYDROPHILIC REGIONS, MASKING OTHER TURNS.

ABTurns             0 (AB_Turn1_4-philics or
                       AB_Turn2_Proline   or
                       AB_Turn3_group     or
                       AB_Turn4_not-beta  or
                       AB_Turn5_group)
     FINAL CONSENSUS TURN PATTERN.
```

proteins with 127 turns, and 8 α/β proteins with 145 turns. The errors were α/α 4% ($^{2}/_{45}$), β/β 8% ($^{10}/_{127}$), and α/β 5% ($^{7}/_{145}$) when the appropriate turn algorithm was applied to each class. Turns were judged correct if they occurred between secondary structure elements or if they were not more than three residues from either end of a secondary structure element. An incorrectly predicted turn divides a secondary element. We do not assign actual "turn" residues for these purposes, merely partitions for secondary structures. We also applied the turn algorithm for each class to sequences in other classes. The expected and predicted number of turns usually agreed best when the appropriate class algorithm was applied, but the data were not predictive.

Although the overall error rate is low, both overprediction and errors of underprediction occurred. Overprediction implies that a turn was predicted in the middle of a secondary structure segment. Errors of this variety are difficult to recover from, if one is intent upon predicting tertiary structure from se-

Table V: β/β Turn Patterns

Pattern Name	Offset	Pattern
ypphil1	0	"[BDEGHKNPQRSTYZ][BDEGHKNQRSZ]%3,3"
ypphil2	0	"[BDEGHKNQRSZ][BDEGHKNPQRSTYZ][BDEGHKNQRSZ]%2,2""
ypphil3	0	"[BDEGHKNQRSZ]%2,2[BDEGHKNPQRSTYZ][BDEGHKNQRSZ]"
ypphil4	0	"[BDEGHKNQRSZ]%3,3[BDEGHKNPQRSTYZ]"
		4 HYDROPHILIC RESIDUES WITH AT MOST 1 Y OR P SWING RESIDUE.
loc_min1+	2	"[ACFIKLMTVWY][BDEGHKNPQRSTYZ]%3,4[ACFIKLMTVWY]"
loc_min3+	3	"[ACFIKLMTVWY][BDEGHKNPQRSTYZ]%5,5[ACFIKLMTVWY]"
		LOCAL "MINIMUM" IN HYDROPHOBICITY.
loc_min1-	1	(density(>=,2,3,1,beta_swing1))
loc_min2-	1	(density(>=,3,4,1,beta_swing1))
loc_min3-	2	(density(>=,3,5,1,beta_swing1))
		USED IN PATTERN BB_LOC-MIN TO AVOID HYDROPHILIC LOCAL MINIMA WHERE THE MAJORITY OF THE HYDROPHILIC RESIDUES ARE ACTUALLY "SWING RESIDUES."
bulge_res	0	"[GSY]" ANIMO ACIDS FREQUENTLY SEEN IN \(*B BULGES.
beta_swing1	0	"[GKSTY]" SWING RESIDUES.
beta_swing2	0	"[AY]" ANOTHER SET OF SWING RESIDUES.
many_bulge	1	(density(>=,3,4,1,bulge_res))
		LOCAL CONCENTRATION OF TYPICAL "BULGE" RESIDUES.
betaphilic	0	"[BDEGHKNPQRSTZ]" HYDROPHILIC RESIDUES.
beta_strong_phobic	0	"[ACFILMVW]" STRONG HYDROPHOBICS.
beta_weak_phobic	0	"[ACFGIKLMSTVWY]" WEAK HYDROPHOBICS.
BB_4-philic	1	(ypphil4 or ypphil3 or ypphil2 or ypphil1)
		STRONGEST TURN PATTERN.
BB_loc-min	0	((loc_min1+ and not (loc_min1- or loc_min2-)) or (loc_min3+ and not loc_min3-))
		USING LOCAL HYDROPHOBICITY MINIMA.
BB_no-phobic	2	(density(=,0,5,1,beta_strong_phobic))
		ABSENCE OF STRONG HYDROPHOBIC RESIDUES.
BB_3-philic	1	(density(>=,3,4,1,betaphilic) and not density(>=,4,5,1,beta_weak_phobic))
		HIGH DENSITY OF HYDROPHILIC RESIDUES WITH LIMITED NUMBER OF SWING RESIDUES.
BB_weak-philic	3	((density(>=,4,7,1,betaphilic) and density(>=,2,7,1,beta_swing2)) or (density(>=,3,7,1,betaphilic) and density(>=,4,7,1,beta_swing2)))
		MINIMUM DENSITY OF HYDROPHILIC RESIDUES SEEN IN TURNS.
BB_4-philic_group	0	(group(BB_4-philic,7))
BB_loc-min_group	0	(group(BB_loc-min,7))
BB_no-phobic_group	0	(group(BB_no-phobic,7))
BB_3-philic_group	0	(group(BB_3-philic,7))
BB_weak-philic_group	0	(group(BB_weak-philic,7))
		GROUPING OF BB_ TURN PATTERNS ABOVE.
BB_Turn1_4-philic	0	(BB_4-philic_group and not many_bulge{0,1})
		MOST DEFINITE TURN PATTERN, MASKING TYPICAL BULGE RESIDUES.
BB_Turn2_loc-min	0	(BB_loc-min_group and not BB_Turn1_4-philic{-4,4})
		LOCAL MINIMUM TYPE TURNS MASKING ONES ABOVE.
BB_Turn3_no-phobics	0	(BB_no-phobic_group and not BB_Turn1_4-philic{-8,8} and not BB_Turn2_loc-min{-4,4})
		NO PHOBIC TYPE TURNS MASKING ONES ABOVE.

BB_Turn4_3-philic	0	(BB_3-philic_group
		and not BB_Turn1_4-philic{-4,4}
		and not BB_Turn2_loc-min{-4,4}
		and not BB_Turn3_no-phobics{-4,4})

		THREE PHILIC TYPE TURNS MASKING ONES ABOVE.

BB_Turn5_weak-philic	0	(BB_weak-philic_group
		and not BB_Turn1_4-philic{-8,8}
		and not BB_Turn2_loc-min{-8,8}
		and not BB_Turn3_no-phobics{-8,8}
		and not BB_Turn4_3-philic{-8,8})

		WEAK PHILIC TYPE TURNS MASKING ONES ABOVE.
BetaTurns	0	(BB_Turn1_4-philic or BB_Turn2_loc-min or BB_Turn3_no-phobics or BB_Turn4_3-philic or BB_Turn5_weak-philic)
		FINAL CONSENSUS TURN PATTERN.

Table VI: False Positive Assignments of Turns

		overpredicted turn		
protein class	protein name	sequence no.	amino acids	secondary structure split
α/α	none			
α/β	LDH	36	ADA	*a*
	SBT	111	NGIE	α 103–117[b]
	TIM	145	QETK	α 137–153[b]
β/β	STNV	187	DSSYE	β 183–192[c]

[a] Helix longer than allowed by model. Weak turn is sought. [b] Strong turn pattern. However, crystallographically a good helix. Potential interactions of K 147 with D 151 to form salt bridge. [c] Unusual collection of residues in the center of an internal β strand in a β sandwich.

Table VII: True Turns Not Assigned

		missed		
protein class	protein name	sequence no.	amino acids	comment
α/α	LZM	91–92	LD	*a*
	MBN	18–20	EAD	*b*
α/β	PGK	330–332	AKQ	*b*
	RHD	10–11	VS	*b*
		247–250	CRKG	*b*
	TIM	86–88	GAA	*c*
		203–204	SR	*d*
β/β	ADH	40–41	MV	*e*
	EST	46–47	GG	*e*
		88–89	GV	*e*
	GCR	127–130	LEGS	perhaps *a*
	IMMH	100–102	IAG	*a*
		176–178	SSG	*a*
	SNS	11–12	PA	*e*
		90–91	AY	*e*
		95–97	DGK	*a*

[a] Masked from consideration since turn separation close to desired helix limit (22 residues). Error of model assumption. [b] Spacing between neighboring turns is long enough to suggest additional turns. Length-based parse of sequence would locate turn. [c] Sequence not hydrophilic enough to be recognized as a potential turn. Perhaps footnote *a* applies. [d] Masked from consideration since 208 is too close (<7 residues) from 203–204. Fortunately, the α helix is noncore. [e] Single β strand hydrogen bonding with both sheets with 90° kink in strand direction between turn residues. Note relative hydrophobicity.

Table VIII: Turn Assignment in Homologous Sets

	no. of proteins	no. of turns per protein	split 2° structure	missed turns	% error
HBN β	37	8	2	27[a]	9
MBN	20	8	3	3	4
HBN α	40	8	18[b]	12	9
LDH	5	17	6	0	7
SOD	4	9	1	1	6
TIM	5	19	10	0	10
IMKAP	7	9	4[c]	2	14
total	118	238	44	45	9

[a] AB corner accounts for 20 of these errors. [b] In 14 proteins, the H helix is split at "SLDK" since the separation between the neighboring turns is ~25 residues. [c] Five-residue β strand with a bulge "GQGTK", called a turn in all sequences.

quence. Fortunately, only three overpredictions occur in the test set. These are cataloged in Table VI.[1] The two sequences have in common hydrophilicity and adequate separation from neighboring turns. Both features lead to failures in the present algorithms. We were unable to develop any reasonable rule to recognize these regions and subsequently avoid the errors. The error rate for overprediction is 0.9%.

Errors of omission—the failures to identify a turn—are more common (Table VII) but are more easily addressed. For instance, a more determined search for increasingly weaker turns in segments too long to remain as a single link would yield a suitable turn in most cases. A few of the turns in β/β proteins deserve further mention. Some β strands have N-terminal residues hydrogen bonded to residues in one β sheet while the C-terminal residues are hydrogen bonded to residues in the second sheet. Since there is a 90° kink between the N-terminal and C-terminal residues, this strand is sometimes labeled as two separate strands [for a complete discussion see Chothia & Janin (1982)]. For example, trypsin can be thought of as composed of domains that are either six- or eight-stranded barrels. The separations between some of the neighboring turns are short (zero to three residues), and these turns are hydrophobic. In the tertiary structure, these turns are frequently less accessible to solvent. We miss these turns with the patterns described here but could retrieve them fairly easily with a pattern designed to recognize the situation.

Homologous Sequences. To assess the utility of these algorithms with sequences outside the test set, we applied them to families of homologous sequences available in the sequence data bank. The results for seven protein families are presented in Table VIII. While tertiary structures are not available for all sequences, we assumed that structure is conserved within each family. Then, the overall accuracy of turn assignment was comparable to that seen in the test set. The majority of errors are in three regions: the omission of the turn between the A and B helices in $^{20}/_{37}$ of the hemoglobin β chains, the splitting of the H helix in 14 of 40 hemoglobin α chains, and the prediction of an extra turn in the last β strand (at a β bulge) in $^{7}/_{7}$ of the immunoglobulin κ light chains.

Importantly, we learn something about our current patterns when used with homology information. It is clear that a turn prediction in one sequence is not sufficient to predict turns in all. Nor is the opposite extreme true: we cannot insist that each homologous sequence agree before predicting a turn for the set. However, a "majority rule" would have recognized the existence of a turn in the hemoglobin β AB corner and could have rejected the splitting of the H helix in hemoglobin α.

The difficulty with the immunoglobulin κ chains suggests a fundamental defect in our current patterns to identify large β bulges in the midst of long β strands.

Figure 2 shows a set of myoglobin sequences together with the actual helices and predicted turns. Beyond the simple demonstration of the accuracy of the turn assignments, this figure suggests the possibility that turn assignments may be

[1] Abbreviations: (α/α proteins) 562, cytochrome *b*-562 (Mathews et al., 1979); CPV, calcium-binding protein, carp muscle (Kretsinger & Nockolds, 1973); HBN, hemoglobin (Ten Eyck & Arnone, 1976); LZM, lysozyme, T4 phage (Matthews & Remington, 1974); MBN, myoglobin (Takano, 1977); MHR, myohemerythrin (Hendrickson & Ward, 1977); TLN, thermolysin (Holmes & Matthews, 1982); TMV, tobacco mosaic virus protein (Bloomer et al., 1978); (β/β proteins) ADH(1–179), alcohol dehydrogenase, liver (Eklund et al., 1976); CNA, concanavalin A (Reeke et al., 1975); EST(1–111), elastase (Sawyer et al., 1978); GCR, γ-crystallin (Blundell et al., 1981); IMMH, Fab fragment of human immunoglobulin; IMML, heavy and light chains (Saul et al., 1978); PRE, prealbumin (Blake et al., 1978); REI, Bence-Jones protein variable dimer (Epp et al., 1974); SNS, nuclease, staphylococcal (or micrococcal) (Arnone et al., 1971); SOD, superoxide dismutase, Cu,Zn (Tainer et al., 1982); STNV, satellite tobacco necrosis virus (Jones & Liljas, 1984); (α/β proteins) ADH(180–374), alcohol dehydrogenase, liver (Eklund et al., 1976); ADK, adenylate kinase (Schulz et al., 1974); FXN, flavodoxin (Smith et al., 1977); LDH, lactate dehydrogenase (Holbrook et al., 1975); PGK, phosphoglycerate kinase (Banks et al., 1979); RHD, rhodanese (Ploegman et al., 1978); SBT, subtilisin (Wright et al., 1969); TIM, triosephosphate isomerase (Banner et al., 1975); (others) CPA, carboxypeptidase A (Hartsuck & Lipscomb, 1971); SRX, thioredoxin, *Escherichia coli* (Holmgren et al., 1975); UTG, uteroglobin (Mornon et al., 1980).

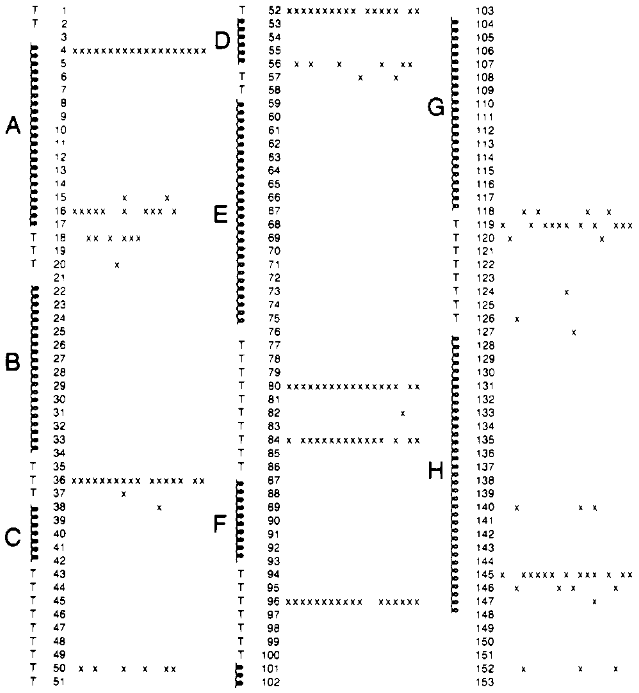

FIGURE 2: **AlphaTurn** pattern matches (X) in 19 myoglobin sequences: helical residues in sperm whale myoglobin (Takano, 1977). Residues in turns are marked with "T". Myoglobin sequences from the NBRF Protein Data Bank: MYBD, badger; MYBO, bovine; MYCA, carp; MYCH, chicken; MYDD, pacific common dolphin, Atlantic bottle-nosed; MYDG, dog, bat-eared fox, and Cape hunting dog; MYEL-I, Indian and African elephants; MYHH, European hedgehog; MYHO, horse and zebra; MYHU, human; MYKG-R, red kangaroo; MYLZ-M, lace monitor lizard; MYOP, opossum; MYOR, platypus; MYPG, pig; MYRK-J, Port Jackson shark; MYSH, sheep and red deer; MYSL-H, harbor seal; MYTT-M, map turtle; MYTU-Y, yellowfin tuna; MYWH-C, whale; MYWH-P, sperm whale and dwarf sperm whale; MYWH-Z, goose-beaked whale and Hubb's beaked whale.

useful for schemes directed at establishing homology and subsequent alignment. This issue merits further investigation.

The turn algorithm was also applied to three proteins not in the set α/β proteins, thioredoxin and carboxypeptidase A, that contain antiparallel strands in the β sheet, and the α/α protein uteroglobin. The α/β turn algorithm located all 31 turns in carboxypeptidase. It also divided two α helices by overprediction of turns. Both helices were split with the **AB_Turn4_not-beta** pattern (Table IV). The class of mixed α/β proteins have β strands with a more diffuse hydrophobic core than pure parallel α/β proteins (Cohen et al., 1982). Thus it is not surprising that a pattern based on the location of hydrophobic sequences frequently seen in internal β strands could cause problems. Turns in thioredoxin and uteroglobin were predicted correctly. The error rate over this set is 6%, which is consistent with the other data sets.

Discussion and Conclusions

We are encouraged by the success of these turn prediction algorithms but recognize the following problems. At this time, no computational method exists for determining which of the three algorithms to apply. Such a scheme is currently under development based on amino acid composition (Sheridan et al., 1985). Experimental information such as circular dichroism or NMR spectra would also be useful. At worst, three possible predictions would have to be carried along until experiment on higher level processing could sort among them. This policy was previously used with α/β proteins (Cohen et al., 1983).

The issue of underpredictions is less problematical. Most occur in regions where a turn is needed to maintain a globular structure. These turns will be sought with increasing diligence. A method for managing overprediction is more difficult. Although a small fraction of the cases, they are potentially disastrous when attempts are made to build tertiary structures. Perhaps what is needed is a procedure to recognize the sequences that code for regions of the protein that do not fit our model. These could result form domain–domain or even subunit–subunit interactions. Much more effort will be required to solve this problem.

One conclusion of this work is that turns are specified by both local and global features of the polypeptide chain. The hypervariable loops of immunoglobulins are exemplary. These loops are located on one side of the immunoglobulin molecule. When viewed along the sequence, the hypervariable turns alternate with turns that have more conserved sequences. Symbolically, the sequence is

$$[\beta\text{–turn–}\beta\text{–hypervariable turn–}\beta\text{–turn–}\beta]$$

If two turns are strongly determined and sufficiently distant along the chain to require the existence of an intervening turn, the precise sequence within the hypervariable region is not critical for maintaining the core structure of the domain. In this way, hydrophobic residues can be accommodated in positions normally occupied by hydrophilic residues and foster antibody diversity without compromising structural integrity.

Obviously, we need to follow this work with additional investigations into a general scheme for secondary structure prediction. Preliminary work suggests that it will be possible to rapidly identify ~30% of the secondary structure with a >95% degree of accuracy. The location of additional segments will be facilitated by the fact that, in the α/α and β/β cases, it is postulated that only one type of secondary structure occurs. Some success has already been achieved with α/β proteins (Cohen et al., 1983).

This algorithm, its successors that are currently under development, and existing secondary structure packing algorithms have been useful in producing models for tertiary structure from a consideration of sequence. As these methods currently exist, they will only be useful if suitable experimental systems can be developed to test specific features. This could result in a reduction of the set of possible structural alternatives generated by these schemes when joined with other methods.

References

Abarbanel, R. M. (1984) Ph.D. Thesis, University of California, San Francisco.

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

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

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., Milman, J. D., 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.

Blake, C. C. F., Geisow, M. J., Oatley, S. J., Rērat, B., & Rērat, C. (1978) *J. Mol. Biol. 121*, 339–356.

Bloomer, A. C., Champness, J. N., Bricogne, G., Staden, R., & Klug, A. (1978) *Nature* (*London*) *276*, 362–368.

Blundell, T., Lindley, P., Miller, L., Moss, D., Slingsby, C., Tickle, I., Turnell, B., & Wistow, G. (1981) *Nature* (*London*) *289*, 771–777.

Chothia, C., & Janin, J. (1982) *Biochemistry 21*, 3955–3965.

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

Chou, P. Y., & Fasman, G. D. (1977) *J. Mol. Biol. 115*, 135–175.

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

Cohen, F. E., Sternberg, M. J. E., & Taylor, W. R. (1982) *J. Mol. Biol. 156*, 821–862.

Cohen, F. E., Abarbanel, R. M., Kuntz, I. D., & Fletterick, R. J. (1983) *Biochemistry 22*, 4894–4904.

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

Epp, O., Colman, P., Fehlhammer, H., Bode, W., Schiffer, M., Huber, R., & Palm, W. (1974) *Eur. J. Biochem. 45*, 513–524.

Foderaro, J. K., Sklower, K. L., & Layer, K. (1983) *The Franz Lisp Manual*, University of California Press, Berkeley, CA.

Garnier, J., Osguthorpe, D. J., & Robson, B. (1978) *J. Mol. Biol. 120*, 97–120.

Hartsuck, J. A., & Lipscomb, W. N. (1971) *Enzymes* (*3rd Ed.*) *3*, 1–56.

Hendrickson, W. A., & Ward, K. B. (1977) *J. Biol Chem. 252*, 3012–3018.

Holbrook, J. J., Liljas, A., Steindel, S. J., & Rossman, M. G. (1975) *Enzymes*, (*3rd Ed.*) *11*, 191–292.

Holmes, M. A., & Matthews, B. W. (1982) *J. Mol. Biol. 160*, 623–639.

Holmgren, A., Söderberg, B.-O., Eklund, H., & Brändên, C.-I. (1975) *Proc. Natl. Acad. Sci. U.S.A. 72*, 2305–2309.

Jones, T. A., & Liljas, L. (1984) *J. Mol. Biol. 177*, 735–767.

Kim, P. S., & Baldwin, R. L. (1984) *Nature* (*London*) *307*, 329–334.

Kretsinger, R. H., & Nockolds, C. E. (1973) *J. Biol. Chem. 248*, 3313–3326.

Kuntz, I. D. (1972) *J. Am. Chem. Soc. 94*, 4009–4012.

Kyte, J., & Doolittle, R. F. (1982) *J. Mol. Biol. 157*, 105–132.

Levitt, M. (1976) *J. Mol. Biol. 104*, 59–116.

Levitt, M., & Chothia, C. (1976) *Nature* (*London*) *261*, 552–558.

Lewis, P. N. Momany, F. A., & Scheraga, H. A. (1971) *Proc. Natl. Acad. Sci. U.S.A. 68*, 2293–2297.

Liljas, L., Unge, T., Jones, T. A., Fridborg, K., Lövgren, S., Skoglund, U., & Strandberg, B. (1982) *J. Mol. Biol. 159*, 93–108.

Lim, V. I. (1974) *J. Mol. Biol. 88*, 857–894.

Mathews, F. S., Bethge, P. H., & Czerwinski, E. W. (1979) *J. Biol. Chem. 254*, 1699–1706.

Matthews, B. W., & Remington, S. J. (1974) *Proc. Natl. Acad. Sci. U.S.A. 71*, 4178–4182.

Momany, F. A., McGuire, R. F., Burgess, A. W., & Scheraga, H. A. (1975) *J. Phys. Chem. 79*, 2361–2381.

Mornon, J. P., Fridlansky, F., Bally, R., & Milgrom, E. (1980) *J. Mol. Biol. 137*, 415–429.

Ploegman, J. H., Drent, G., Kalk, K. H., & Hol, W. G. J. (1978) *J. Mol. Biol. 123*, 557–594.

Ptitsyn, O. B., & Rashin, A. A. (1975) *Biophys. Chem. 3*, 1–20.

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

Richardson, J. S., Getzoff, E. D., & Richardson, D. C. (1978) *Proc. Natl. Acad. Sci. U.S.A. 75*, 2574–2578.

Robson, B., & Osguthorpe, D. J. (1979) *J. Mol. Biol. 132*, 19–51.

Rose, G. D. (1978) *Nature* (*London*) *272*, 586–590.

Saul, F. A., Amzel, L. M., & Poljak, R. J. (1978) *J. Biol. Chem. 253*, 585–597.

Sawyer, L., Shotton, D. M., Campbell, J. W., Wendell, P. L., Muirhead, H., Watson, H. C., Diamond, R., & Ladner, R. C. (1978) *J. Mol. Biol. 118*, 137–208.

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

Sheridan, R. P., Dixon, J. S., Venkataraghavan, R., Kuntz, I. D., & Scott, K. P. (1985) *Biopolymers 24*, 1995–2003.

Smith, W. W., Burnett, R. M., Darling, G. D., & Ludwig, M. L. (1977) *J. Mol. Biol. 117*, 195–225.

Tainer, J. A., Getzoff, E. D., Beem, K. M., Richardson, J. S., & Richardson, D. C. (1982) *J. Mol. Biol. 160*, 181–217.

Takano, T. (1977) *J. Mol. Biol. 110*, 569–584.

Taylor, W. R., & Thornton, J. M. (1983) *Nature* (*London*) *301*, 540–542.

Ten Eyck, L. F., & Arnone, A. (1976) *J. Mol. Biol. 100*, 3–11.

Wilson, I. A., Skehel, J. J., & Wiley, D. C. (1981) *Nature* (*London*) *289*, 366–373.

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