A Segment-Based Approach to Protein Secondary Structure Prediction†

Scott R. Presnell,‡ Bruce I. Cohen,‡,§ and Fred E. Cohen*,‡,∥

Departments of Pharmaceutical Chemistry and Medicine and Section on Medical Information Science, University of California, San Francisco, San Francisco, California 94143

Received April 22, 1991; Revised Manuscript Received August 5, 1991

ABSTRACT: Amino acid sequence patterns have been used to identify the location of turns in globular proteins [Cohen et al. (1986) *Biochemistry 25*, 266–275]. We have developed sequence patterns that facilitate the prediction of helices in all helical proteins. Regular expression patterns recognize the component parts of a helix: the amino terminus (N-cap), the core of the helix (core), and the carboxy terminus (C-cap). These patterns recognize the core features of helices with a 95% success rate and the N- and C-capping features with success rates of 56% and 48%, respectively. A metapattern language, ALPPS, coordinates the recognition of turns and helical components in a scheme that predicts the location and extent of α-helices. On the basis of raw residue scoring, a 71% success rate is observed. By focusing on the recognition of core helical features, we achieve a 78% success rate. Amended scoring procedures are presented and discussed, and comparisons are made to other predictive schemes.

Under suitable conditions, many proteins adopt a compact, globular fold which is dictated by the amino acid sequence of their polypeptide chains. However, the precise relationship of sequence to structure remains unresolved. Substantial experimental and theoretical efforts have been directed at understanding the protein folding problem. Experimentalists have uncovered evidence of nativelike intermediates along the folding pathway (Hughson et al., 1990; Goto & Fink, 1990; Ptitsyn et al., 1990). Theoretical methods are also being used to explore possible folding pathways and to gain an understanding of the forces which stabilize folded proteins. To date, three approaches have been employed: energy minimization and/or molecular dynamics (Levitt & Warshel, 1975; Nemethy & Scheraga, 1977; Weiner et al., 1984; McCammon et al., 1977; Karplus & McCammon, 1981; Beveridge & Jorgenson, 1986), lattice models (Skolnick & Kolinski, 1989), and semiempirical substructure condensation (Cohen et al., 1979, 1980).

Energy minimization techniques and molecular dynamics approaches offer the promise of a rigorous treatment of the inter- and intramolecular forces in protein structures. However, several practical details of these methods remain unresolved. The available potential functions do not provide an accurate representation of energy states, bulk solvent is not handled adequately, and optimization algorithms cannot sample the entirety of conformation space. Moreover, computing power enables us to visualize no more than 2 or 3 ns of dynamic variation, while protein folding requires milliseconds to seconds (Udgaonkar & Baldwin, 1988; Roder et al., 1988).

Monte Carlo simulations utilizing simplified lattice frameworks have been performed in an attempt to elucidate general rules for globular protein folding. These simulations have incorporated several different lattice types (cubic, diamond, and "210" or "knights walk") to simulate the folding pathways of four-helix bundles such as apoferritin and somatotropin (Sikorski & Skolnick, 1989) and β-barrel proteins such as plastocyanin (Skolnick & Kolinski, 1990). These methods provide an interesting vignette of the possible folding pathways, though the simulations currently require that position-specific conformational preferences be built into the backbone atom representation of each simulated sequence.

Semiempirical methods also suffer some limitations, but they have found utility in the development of structure models. The present semiempirical condensation methods are based on a hierarchical definition of globular proteins. The classical structure hierarchy presents three levels: primary, secondary, and tertiary structure. Typically, condensation schemes fold primary structure into secondary structure, then secondary structure is assembled into tertiary structure. This strict ordering is not intended to be an accurate reflection of a protein folding pathway: some aspects of secondary structure formation may be influenced by specific tertiary interactions. However, current work on the molten globule state indicates that the intermediate stages in protein folding show a large fraction of the secondary structure apparent in the native state (Hughson et al., 1990; Goto & Fink, 1990; Ptitsyn et al., 1990). The body of work that focuses on the transition of secondary structure to tertiary structure will not be considered here (Cohen et al., 1979, 1980, 1982). This work concentrates on advanced techniques for relating primary structure to secondary structure.

Several groups have reported methods for the prediction of secondary structure in globular proteins [for a review see Schulz (1988)]. These methods group loosely into two classes. The collection of known structures provides a database of information about the propensity of the individual amino acid to reside in specific types of secondary structure. The first class of methods is based on statistical analyses of this data. The progenitor of these methods was developed by Chou and Fasman (1974, 1978). Specialized treatments of context in the amino acid sequence (e.g., Markov dependence) or the information content within the sequence were developed into prediction methods by Garnier et al. (1978, 1987; referred to here as the GOR method). The most advanced of these methods now combines many predictive schemes or combinations of predictions from sequences homologous to the se-

† We acknowledge the support of the Computer Graphics Laboratory at UCSF (NIH RR1081), the Macromolecular Workbench project (DARPA, ONR N00014-86-K-0757), and the National Institutes of Health (NIH GM39900).

‡ Department of Pharmaceutical Chemistry.
§ Section on Medical Information Science.
∥ Department of Medicine.

0006-2960/92/0431-983$03.00/0

quence of interest (Levin & Garnier, 1988; Nishikawa & Ooi, 1986). Recently, computational neural networks have also been used to investigate the mapping of protein sequence to secondary structure (Qian & Sejnowski, 1988; Holley & Karplus, 1989; Kneller et al., 1990).

The second class of methods rely on biophysical principles as a basis for the prediction of interactions among the amino acids. This approach was first outlined by Nagano (1973, 1974), and Lim (1974a,b). For globular proteins, these principles include compactness of form, the presence of a hydrophobic core or cores, and a polar outer shell. The geometries inherent in the two archetypes of secondary structure, α-helix and β-sheet, afford restrictions on the types of amino acid side- and main-chain interactions. Our methods for the generation and analysis of patterns that recognize sequence–structure correlates have followed from this second class of techniques.

Long-range interactions are believed to play a critical role in the formation of complete tertiary structure. Kabsch and Sander (1984) were among the first to note that identical or similar sequences of up to five residues can adopt decidedly different three-dimensional structures. Hence, five residues of context is not enough to define unique three-dimensional structure. Classical secondary structure prediction methods typically achieve results of no greater than 65% accuracy. This limitation has often been attributed to the absence of long-range interactions in the prediction algorithms. At first glance, the specification of long-range interactions seems a daunting problem. However, general knowledge of the nature or types of long-range interactions expected can be included into predictive schemes. In previous work, Cohen et al. (1986) described the specification of regular expression patterns that could incorporate long-range interactions from the estimated turn distribution in proteins. The patterns developed under the PLANS system were able to accurately locate turns in the three classes of globular proteins (α/α, α/β, and β/β) with success rates approaching 90%.

We report here the extension of the PLANS work to predict regular secondary structure in α/α proteins. This advance involved two explicit developments. First, the general problem of regular secondary structure prediction was divided into subtasks. Regular expression patterns were developed to recognize the individual components parts of helices, the amino terminus (referred to hereafter as N-cap region), the core, and the carboxy terminus (C-cap). Those patterns were designed to function in an autonomous fashion. Second, A Language for the Prediction of Protein Substructures (ALPPS) was formulated to coordinate the development and analysis of metapatterns: patterns of patterns. In the case of helical structure prediction, metapatterns coordinate the recognition of the helix components.

The patterns developed in this work recognize three helical features to differing extents. Scoring the success of feature prediction, we are able to detect 95% of the helix core structures, with a 10% overprediction rate (for every 10 helix core features predicted correctly, one overprediction will occur). N- and C-terminal helix caps are much more difficult to recognize. Half of these features are detected, but a 25% overprediction rate is observed. The recognition of individual features at these rates produces prediction accuracies that exceed those of the statistically based prediction algorithms. The residue-based scores from the pattern-based work presented here do not surpass the scores obtained by the latest neural network algorithms. However, pattern-based methods allow complete inspection, and consequent structural interpretation, of the tools used to predict structure. Interpretation of the internal weights of neural network connections is difficult for all but the simplest architectures.

In summary, we can analyze protein structures for specific features of helical secondary structure and develop patterns to recognize those features. We can also orchestrate the recognition of these features in specific orderings via a language that describes metapatterns. The developments described in this work mark the emergence of a prediction scheme that uses the hierarchical organization in protein structure to facilitate the inclusion of sequential, long-range interactions.

Experimental Procedures

Definitions. Our previous work on the prediction of turns considered protein sequences from three structural classes of proteins α/α, α/β, and β/β (Levitt & Chothia, 1976). In this work, we have focused on the features that stabilize individual α-helices within protein structures constructed primarily of helices (α/α protein structures). This is not intended to be a representative sampling of all globular proteins. Instead, it provides a limitation on the variety of structural interactions to identify, characterize, and predict. For α/α proteins the polypeptide chain has only two conformational states: helix and nonhelix. All forms of helical structure (3_{10} and α) are treated identically.

One drawback to class-specific structure prediction algorithms is the problem of determining the structure class of the protein under scrutiny. Protein class determination in the absence of a crystal structure remains a difficult problem in biochemistry (Sheridan et al., 1985; Klein & Delisi, 1986; Deleage & Roux, 1987). However, work in sequence analysis (Bowie et al., 1990) and machine learning techniques (S. H. Kim, personal communication, 1991) are beginning to provide new algorithms for this task. Advanced experimental methods (Lee et al., 1990) are also providing new methods beyond the classical circular dichroism techniques for determining structure class from experimental data (Johnson, 1990).

Data Sets. Twenty polypeptide chains from the collection of α/α proteins in the Brookhaven Protein Data Bank (PDB; Bernstein et al., 1977) were pooled and split into two sets of ten chains each. One set was used for the development and analysis of patterns. The sequences of the other set were sequestered from examination for the purposes of unbiased evaluation after pattern development was complete (Table I). Except for the structure of tobacco mosaic virus coat protein, the data sets were selected from crystal structures that have an atomic resolution better than 2.5 Å. Identity scores were evaluated after alignment with the multiple sequence alignment program PIMA (Smith & Smith, 1990). The greatest amount of identity between any two polypeptide chains was 42% (between fetal human hemoglobin γ chain and human adult human hemoglobin β chain). The next highest identity was 27%.

Secondary Structure Assignment. Secondary structure assignment was performed with the aid of the program DEFINE (Richards & Kundrot, 1988). Briefly, the algorithm within DEFINE evaluates the fit of actual secondary structure backbone C_α atoms to an ideal secondary structure through a difference distance matrix technique. Specific structure "masks" are used to make secondary structure assignments to the three-dimensional structure. The cumulative root mean squared difference between the ideal and actual structures is used to evaluate the assignment of the desired structure type. In this study, 0.75 Å was used as the maximum rms difference between the ideal and actual helical structures. This yielded a

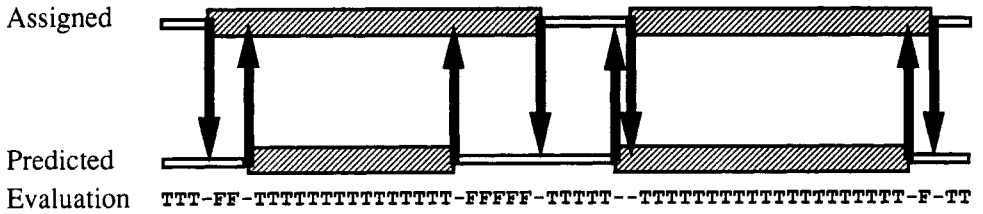

FIGURE 1: Trimming in residue-based scoring. A total of three residues flanking the N- and C-caps in both the assigned and predicted helices are not evaluated in the residue-based scoring of a prediction. The "trimming" of the helical end points concentrates evaluation on the core of the helical structure. T = true positive; F = false negative; - = not evaluated.

Table I: Proteins Used for Pattern Development and Analysis

PDB	protein	reference
	Development Set	
1ccr	cytochrome *c*	Ochi et al., 1983
1fdh	human fetal hemoglobin (γ chain)	Frier & Perutz, 1977
2ccy	cytochrome *c'*	Finzel et al., 1985
2cts	citrate synthase	Remington et al., 1982
2lh1	leghemoglobin	Arutyunyan et al., 1980
2lhb	hemoglobin V	Honzatko et al., 1985
2lzm	T4 lysozyme	Weaver & Matthews, 1987
3c2c	cytochrome *c2*	Bhatia, 1981
3cln	calmodulin	Babu et al., 1988
3cpv	parvalbumin B	Moews & Kretsinger, 1975
	Test Set	
156b	cytochrome *b562*	Lederer et al., 1981
1cc5	cytochrome *c5*	Carter et al., 1985
1ecd	erythrocruorin	Steigemann & Weber, 1979
1hmq	hemerythrin	Stenkamp et al., 1983
1mbd	myoglobin	Phillips & Schoenborn, 1981
2cyp	cytochrome *c* peroxidase	Finzel et al., 1984
2tmv	tobacco mosaic virus coat protein	Namba et al., 1989
3hhb	human hemoglobin (α chain)	Fermi et al., 1984
3icb	vitamin D-dependent calcium-binding protein	Szebenyi & Moffat, 1986
3wrp	Trp aporepressor	Lawson et al., 1988

secondary structure assignment consistent with most authors' assignments. While other programs are available for structure assignment (Kabsch & Sander, 1983; Sklenar et al., 1989), DEFINE was employed because it more closely matched the crystallographers' helical structure assignments.

Evaluation. The accuracy of the algorithms presented here was evaluated using several measures. Feature-based scoring was used to evaluate the predictive capabilities of the PLANS patterns. If a pattern appears within four residues of the targeted feature of structure (e.g., the N-cap of an α-helix), that event is considered a true positive (tp); otherwise the event is registered as a false positive (fp). The absence of a prediction for a targeted feature is registered as a false negative (fn). Since PLANS patterns do not explicitly predict the absence of a structure element, true negatives (tn) cannot be recorded. We represent the success of a feature-based prediction as the quotient of the correctly predicted features and the total number of features. This index, often referred to as Q_2, is defined in eq 1 (Schulz & Schirmer, 1979). This index all but ignores the possibility of overprediction, so we also report the quotient of overpredicted features (fp) and the total number of features (referred to simply as O; eq 2).

$$Q_2 = \frac{\text{tp}}{\text{tp} + \text{fn}} \quad (1)$$

$$O = \frac{\text{fp}}{\text{tp} + \text{fn}} \quad (2)$$

The primary goal of the patterns of algorithms developed in this work was the prediction of helical features. However, structural feature analysis is not a commonly accepted standard for algorithm comparison. Residue-based scoring, in which each residue is given equal weight, was used to evaluate the success of helical predictions for comparison to other algorithms. In a two-state environment, there are four categories of prediction: residues predicted helical and observed helical (TP), residues predicted nonhelical and observed nonhelical (TN), residues predicted nonhelical but observed helical (FN), and residues predicted helical but observed nonhelical (FP) (Matthews, 1975). Two values are often used to assess and compare secondary structure prediction methods. The first is Q_3, the quotient of the number of correctly predicted residues and the total number of predicted residues:

$$Q_3 = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}} \quad (3)$$

This index takes into account both over- and underprediction. To evaluate the prediction accuracy over that expected for a random prediction, a correlation value, C, has also been widely used [Schulz and Schirmer (1977) refer to this index as Q_7]. C is defined as

$$C = \frac{(\text{TP})(\text{TN}) - (\text{FP})(\text{FN})}{[(\text{TN} + \text{FN})(\text{TN} + \text{FP})(\text{TP} + \text{FN})(\text{TP} + \text{FP})]^{1/2}} \quad (4)$$

The exact location of helix termini is often subject to interpretation. Neither automatic assignment methods nor scientists can agree on the precise location of helix termini. The methods of Kabsch and Sander (1983), Richards and Kundrot (1988), and authors' assignments (Bernstein et al., 1977) differ by 10%–14% (Q_3) in their assignments of structure. In addition, only local information is used in a structure prediction to identify helical end points. The final, precise location may be influenced by global forces that are difficult to delineate. Since it is not possible to unambiguously define the termini of helices, it is unreasonable to assign an equal penalty to the incorrect prediction of the central helix structure and the termini. With this in mind, we have developed a simple alteration to the residue-based scoring scheme that discounts the terminal residues in the assigned and predicted helices. Figure 1 demonstrates the "trimming" technique. Residues from the assigned and predicted ends of the secondary structures are removed from consideration in the residue-based scoring procedure. The resulting residue-based comparison focuses more weight on the residues that comprise the core of the secondary structure rather than those residues at the ends of the assigned and predicted structures. Eliminating the terminal residues from evaluation avoids the problem of deciding on a relative weight for the core helical structure and the termini.

Pattern Development. In our previous work, PLANS was developed to facilitate expression of sequence–structure correlates as regular expression patterns (Cohen et al., 1986). We continue to use that regular expression syntax and algorithm

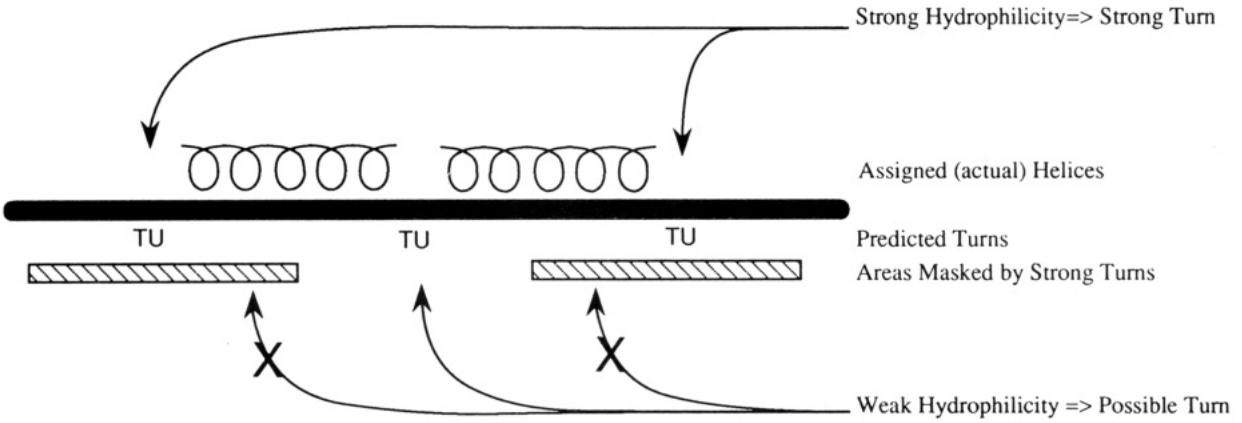

FIGURE 2: Turn prediction reviewed. A high concentration of hydrophilic residues often signals the presence of a turn in globular protein sequences. A block of residues around these strong turn indicators are "masked" from further consideration. Only those regions of the sequence free from the mask are considered when examining the sequence for less likely turn indications (lower concentrations of hydrophilic residues).

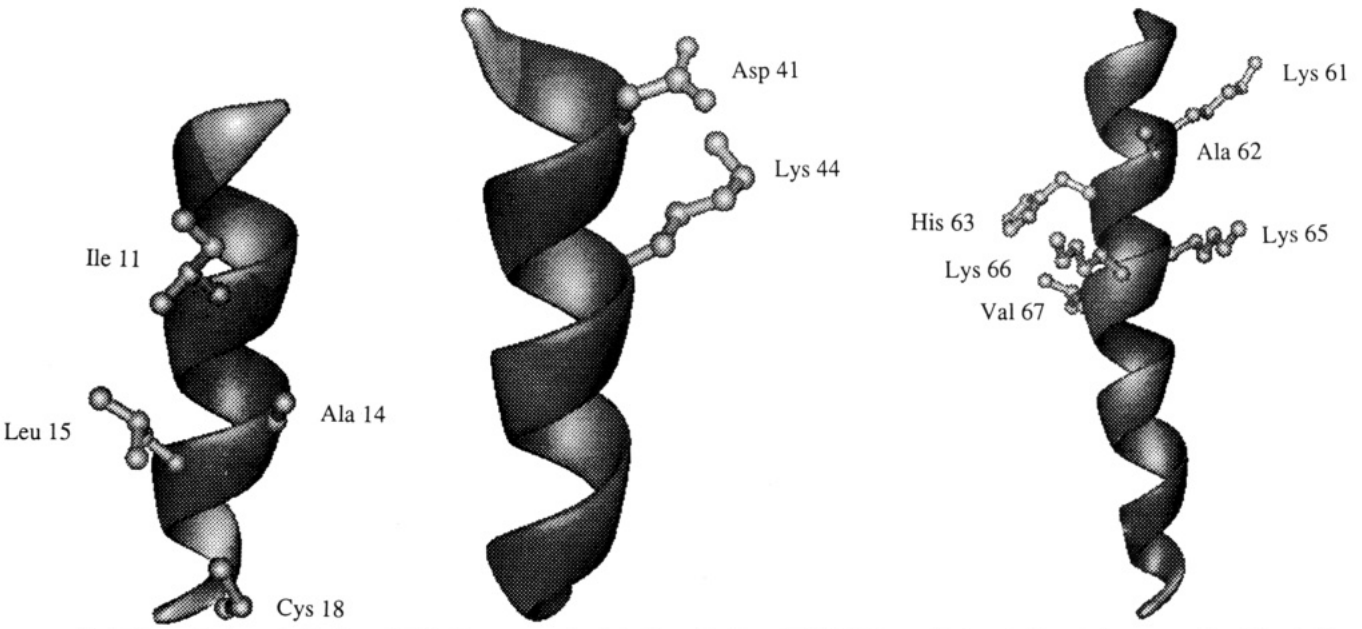

FIGURE 3: Patterns that are indicative of helical sequences include those that would (left) form a hydrophobic patch on one side of the helix (3cpv residues 7–19), (center) form a charged pair of residues (3cpv residues 39–51), and (right) form a putative helix–helix interaction site (1fdh residues 57–77). A low density of hydrophilic residues might also indicate a helical region.

as the basis for turn and helix component pattern development in this work.[1]

Turn Patterns. In the ALPPS-assisted prediction of regular secondary structure, the first step is to predict the likely locations of turns. This is done using the composite turn prediction pattern, TU, developed previously (Cohen et al., 1986). TU is the synthesis of several patterns that predict turn location. A cluster of hydrophilic residues provides the strongest signal. A pattern that explicitly recognizes the special role of proline in initiating and terminating helices is also included. Weaker turns are recognized as an interrupted collection of hydrophilic residues spaced appropriately from the strongest turn indications. The interrelationship of the constituent aspects of TU is outlined in Figure 2.

Helix Component Patterns. The characteristics of helical residues vary as a function of their position within the helical structure. There are often differences in the spatial distribution of residues (e.g., the amphiphilic character of a helix) as well as differences in residue types between the center and the termini. Several authors have noted specific sequential characteristics of helices from globular proteins (Richardson & Richardson, 1988) and others from membrane-associated proteins (Rees et al., 1989; Sternberg & Gullick, 1990). In order to efficiently describe patterns that recognize diverse α-helical characteristics in α/α proteins, we have chosen to subdivide helices into three specific components for study: the central section or core region of the helix, the amino terminal area of the helix (referred to hereafter as the N-cap), and the carboxy terminal area (C-cap). While the principle of structure subdivision is general, the PLANS patterns subsequently described are specific to soluble, globular proteins. The patterns would have to be reformulated for integral membrane proteins.

Helix Core. The criteria for the prediction of the helix core regions incorporate several different biophysical properties (Figure 3). Sequential placement of hydrophobic residues in a pattern suggestive of a *hydrophobic patch* on one face of a helix would facilitate the creation of a hydrophobic interaction with another part of the protein (Schiffer & Edmundson, 1968):[2]

PATCH: Φ..ΦΦ..Φ

[1] The PLANS patterns described here, which show primarily sequential relationships, are presented in a simplified syntax. The supplementary material contain complete pattern descriptions.

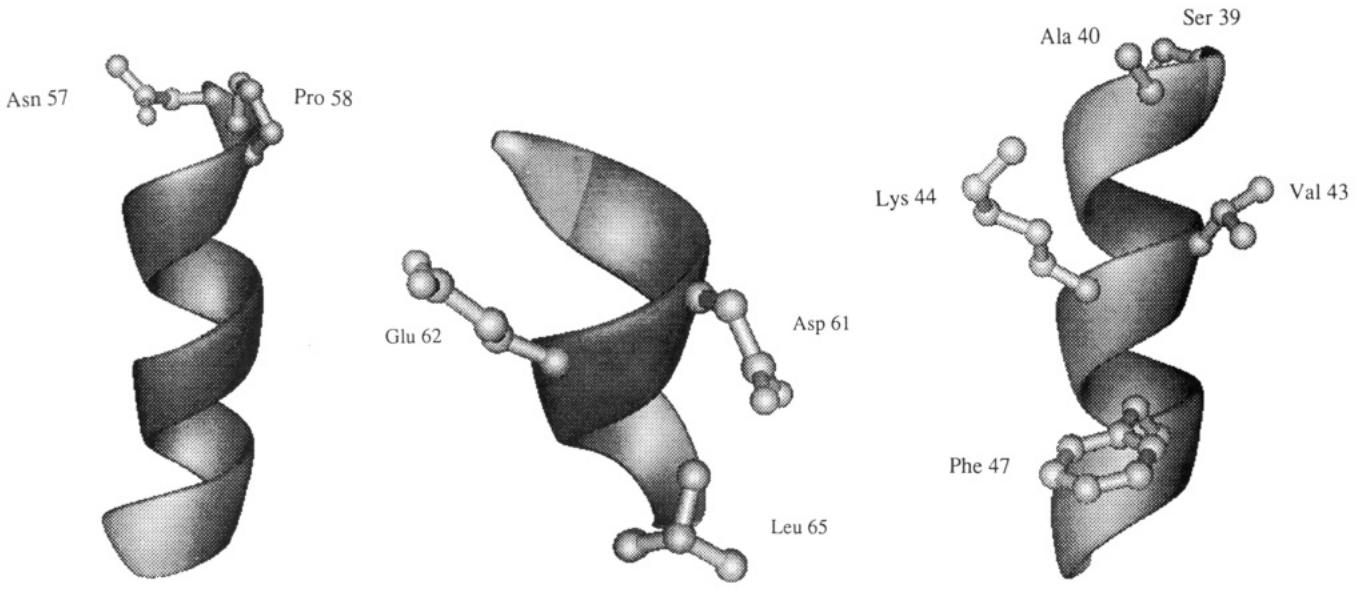

FIGURE 4: Sequence patterns indicative of N-cap sites include (left) simple juxtapositions of specific residues (1fdh residues 57–77), (center) acidic and large hydrophobic residues placed just after a residue with a strong N-cap preference (3cpv residues 59–65), and (right) a strong N-cap residue "terminating" a putative hydrophobic patch for a helix (3cpv residues 39–49).

Acidic and basic residues can be placed in such a way as to generate a *charged pair* on one surface of the helix:

PAIR: −...+

Empirical rules can be used to identify a putative helix–helix interaction site (Richmond & Richards, 1978):

INTER: $\Phi\Psi\Psi[^{A}_{G}]\Psi\Psi\Phi$

Long stretches of hydrophilic residues indicate a turn regions on the surface of the molecule. By process of elimination, the remaining areas are likely to contain helices:

NO-PHIL: $\bar{\Psi}\bar{\Psi}\bar{\Psi}\bar{\Psi}\bar{\Psi}\bar{\Psi}\bar{\Psi}\bar{\Psi}$

We have found that these criteria are effective when given equal weights (i.e., not hierarchical). These individual PLANS patterns, along with others, are collected into an aggregate helix core pattern called HCORE.

Helix N-Cap. Three subtypes of patterns have been developed to describe the amino-terminal capping sites of the helices (Figure 4). Those patterns that give the most reliable indication of a helix N-cap stem from the combination of a residue commonly found at the exact helix N-cap site (in this case N, S, T, or D) and a proline residue (Richardson & Richardson, 1988):

NSIMPLE: $\begin{bmatrix} N \\ S \\ D \\ T \end{bmatrix} P$

Patterns containing a residue from the set N, D, or S one or more acidic residues one to three residues from the N-cap site and one or more large hydrophobic residue five to six residues from the N-cap site are also highly reliable:

NACID: $\begin{bmatrix} N \\ D \\ S \end{bmatrix} E \begin{bmatrix} D \\ E \end{bmatrix} \begin{bmatrix} D \\ E \end{bmatrix} \begin{bmatrix} L \\ F \\ W \\ M \end{bmatrix} M$

The placement of the acidic residue correlates well with the hypothesis that interaction with the helix dipole is important to the stabilization of helical structure. These two patterns are considered at the same level in the N-cap pattern hierarchy. The next most reliable patterns require a residue that strongly suggests a N-cap site "in phase" with a cluster of hydrophobic residues on one face of the putative helix. In the case of N-caps, it seems that there is a requirement for a hydrophilic residue, just after the N-cap position, to terminate the hydrophobic patch:

NPHASE: $\begin{bmatrix} G \\ N \\ D \\ T \\ S \end{bmatrix} \psi . .\Phi\Phi. .\Phi$

This effect will be referred to as the "hydrophobic phasing" of the cap site. These patterns, along with others, are collected together into a composite N-cap pattern referred to as NCAP.

Helix C-Cap. The C-cap prediction scheme follows an analogous hierarchical construction of patterns, but the critical residues differ. The pattern that provides the most reliable, independent indication of a carboxy-terminal site for a helix is the juxtaposition of a G at the C-cap site and a P one or two residues after the helix:

CSIMPLE: GP

Patterns containing either G, H, or K (residues indicative of the C-cap site) and one or more basic residues one to three positions from the C-cap site or a large hydrophobic residue three to four positions upstream of the C-cap site have also proven predictive:

CBASE: $\begin{bmatrix} F \\ M \\ W \end{bmatrix} \begin{bmatrix} L \\ K \end{bmatrix} \begin{bmatrix} R \\ K \end{bmatrix} \begin{bmatrix} Q \\ H \\ K \end{bmatrix} \begin{bmatrix} G \\ H \\ K \end{bmatrix} P$

[2] The general PLANS pattern syntax is pattern name: pattern. Φ represents one of a set of hydrophobic residues, A, V, I, L, M, C, K, F, W, or Y. Ψ represents one a set of hydrophilic residues, A, D, E, H, K, N, Q, R, S, or T. − represents one of the acidic amino acids, D or E. + represents one of the basic amino acids, K or R. A dot (.) represents any amino acid. $^{X}_{Y}$ represents residues X or Y. $\bar{\Psi}$ represents the complement of the set Ψ.

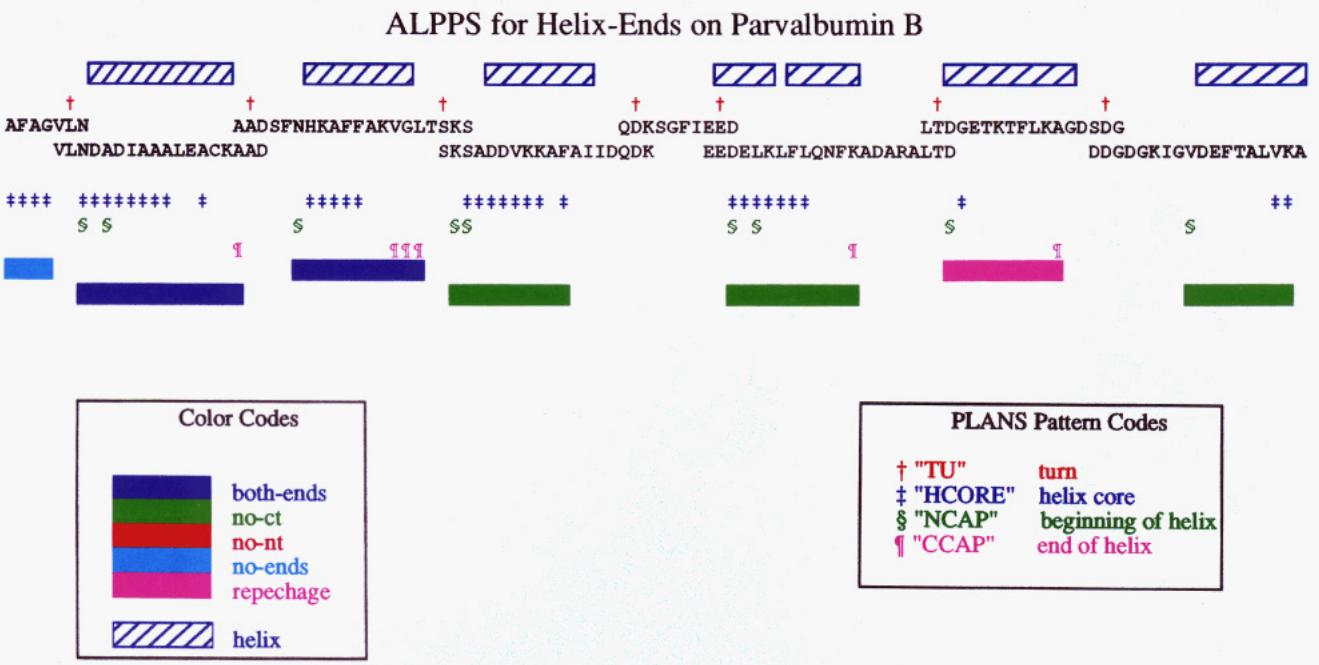

FIGURE 5: A working example of an ALPPS prediction. The first row is a description of the actual secondary structure for carp parvalbumin *b* (3cpv). The next three rows show where the sequence is broken into blocks using the turn prediction algorithm. The next three rows indicate the low-level PLANS patterns that recognize the different helical foundations (N-cap, helix core, and C-cap). ALPPS then uses the location of these PLANS patterns to define secondary structure. In this case, ALPPS first looks for the correct juxtaposition of N-cap core and C-cap and then calls the region between the N-cap and C-cap a helix. If one of the helix caps cannot be found, the ALPPS constructs a helix covering the distance between the remaining cap pattern and the helix core patterns, or possibly to the end of the block. If there are no caps at all, the helical core pattern might be used to define the helical region.

By analogy with the N-cap patterns, a residue that strongly suggests a C-cap specific site "in phase" with a hydrophobic patch on one face of a putative helix constitutes the final class of patterns used in C-cap prediction:

$$\text{CPHASE:} \quad \Phi..\Phi\Phi..\begin{bmatrix} G \\ N \\ H \\ K \end{bmatrix}$$

These patterns, along with others, are collected together into a composite C-cap pattern referred to as CCAP.

ALPPS Pattern Language. While PLANS provides an excellent vehicle for expressing patterns for local sequence interactions it does not provide a convenient means of expressing patterns of patterns, or metapatterns. To evaluate the orderings of helical component patterns in metapatterns, we have developed ALPPS. Metapatterns in ALPPS exploit the relationship of individual PLANS patterns to each other, facilitating the description of entire structure segments in the amino acid sequence. By utilizing the information in the required sequential ordering of the secondary structure component patterns, we taken the next step in introducing long-range interactions into secondary structure prediction.

ALPPS has been implemented in the Allegro Common Lisp (Franz Inc.) on a Sun SparcStation 1 running the UNIX(tm) operating system. The source for the entire pattern-matching system, including the original PLANS pattern matching code, now maintained in the C computer language, and the Lisp code for ALPPS is available from the authors.

The evaluation of an ALPPS pattern begins by segmenting an amino acid sequence into blocks based on the location of a specified PLANS pattern. These blocks may be evaluated against previously assigned target regions in the residue sequence without further processing. Under normal circumstances, additional patterns are used to specify a region: a subsequence within a block. These manipulations are performed by ALPPS functions that take the names of PLANS patterns as arguments. Table II gives a list of functions available to construct the meta-level patterns. A more detailed description of the ALPPS language can be found in Cohen et al. (1991).

Table II: Functions Available in ALPPS

function	operation
def-alpps	split the sequence into several blocks based in the location of a given pattern, a minimum allowed block size, and a tolerance or overlap value
hide-blocks	hide blocks that contain a given pattern
expose-blocks	expose blocks that contain a given pattern
split-blocks	split a block into two blocks on the basis of the occurrence of a given pattern
cat-blocks	concatentate two adjacent blocks into one larger block on the basis of the occurrence of given patterns
make-regions	further specify a subsequence of a block as a region containing a middle pattern, initiated by a starting pattern, and terminated by an ending pattern; regions are associated with targets for the purposes of scoring; regions may also be named

For the prediction of α-helices, the location of the turn predictions generated by the TU pattern is used to segment the sequence into blocks. After initial segmentation, all the sequence blocks are hidden from consideration. Then, those blocks that contain strong evidence for helical structure are exposed for evaluation. This is done using the PLANS pattern for helix core recognition, HCORE. Each visible block is then examined for orderings of PLANS patterns that match the region definitions supplied in the form of an ALPPS pattern. The subsequent PLANS patterns are then evaluated within the context of a block. Helical regions are specified under four possible conditions. Under the best of conditions, the amino terminus, the core, and the carboxy terminus of a helix are recognized by the PLANS patterns NCAP, HCORE, and CCAP, respectively. The sequence beginning with the location of the NCAP pattern and ending with the CCAP pattern is marked as

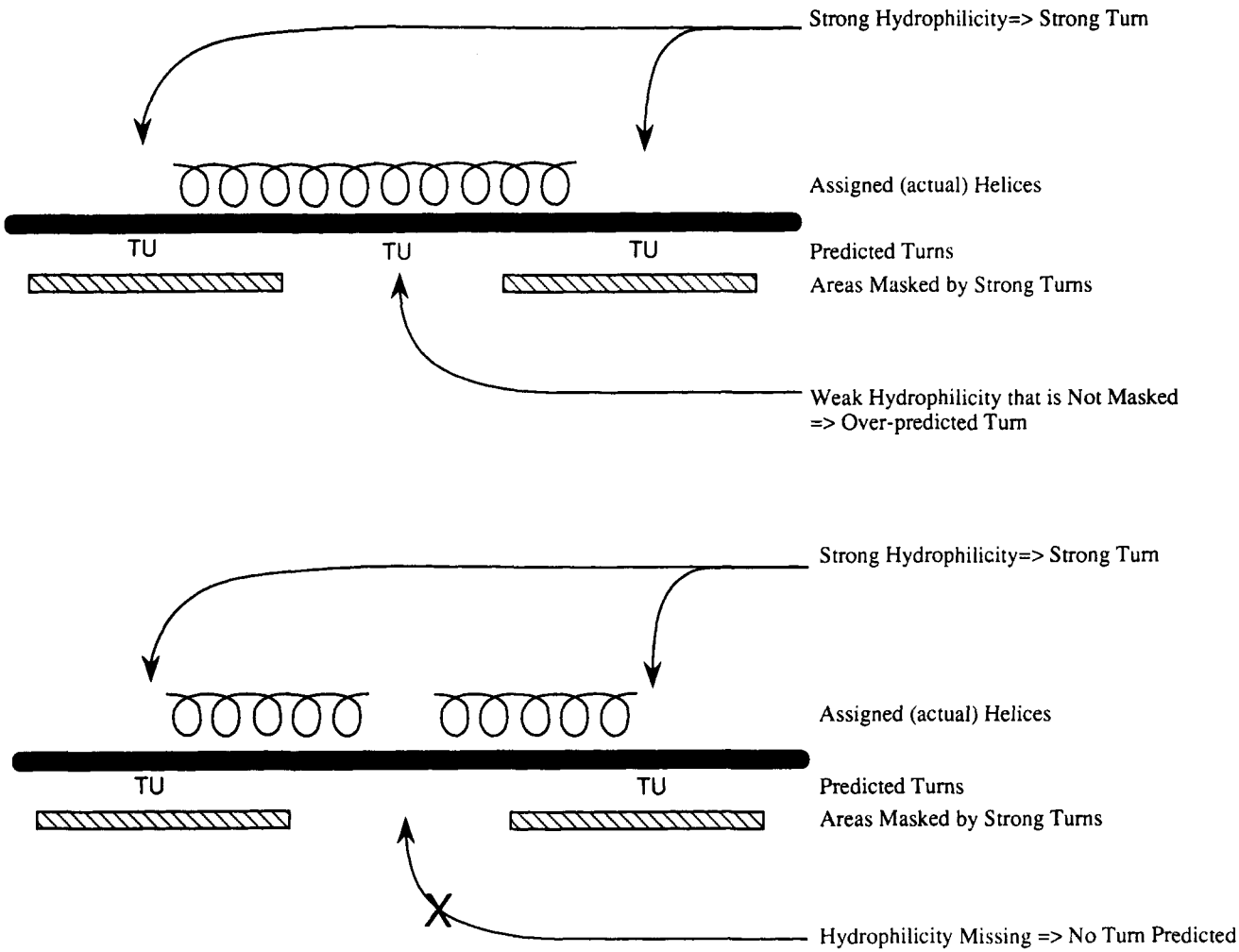

FIGURE 6: (Top) Overpredicted turns. A segment of weakly hydrophilic residues is found sufficiently far away from adjacent strong turns that it is considered valid. However, the prediction of this turn produces two blocks where there should be only one. (Bottom) Underpredicted turns. A segment of sequence, while sufficiently far away from the adjacent strong turns, does not contain enough sequential evidence to predict the location of a turn. This sequence leads to the production of only one block where there should be two.

a helical region. If one or more of the capping patterns is not recognized within the block, helical regions are constructed from the available information. Finally, weak helical indications can be used in a second analysis of the sequence ("repechage"). The ordering of region definitions is significant, and there is at most one helical region per block. Figure 5 presents a symbolic representation of the different stages of prediction for the protein carp parvalbumin B (PDB designation 3cpv).

Results and Discussion

Table III shows the predictive capabilities of individual PLANS patterns on the specific structure subtypes: turns, helical cores, N-caps, and C-caps. Because of a lack of false negatives (fn), strict Q_3 or C values as described earlier cannot be calculated: the PLANS patterns are scored on a basis of feature recognition. Multiple identifications of the same feature had no effect on the accuracy scores.

The patterns used to predict the locations of turns show a high level of success. The 2% decrease in prediction accuracy from the development set to the test set suggests the possibility that some patterns recognize specific features of the development set rather than generalized biophysical principles of turn stabilization. Within an error range of four residues, the strong turn patterns rarely generate false indications. However, the weaker turn prediction patterns are not as reliable an indicator of the placement of actual turns. These patterns

Table III: Feature Scores

class	Q_2 (%)	O (%)	tp	fn	fp
		Development Sequence Set			
turns	84	12	63	12	9
cores	95	10	81	4	9
N-cap	61	20	51	33	17
C-cap	58	27	49	35	23
		Test Sequence Set			
turns	82	11	47	10	6
cores	95	10	63	4	7
N-cap	51	30	34	33	20
C-cap	35	32	23	44	21
		Both Sequence Sets			
turns	83	11	110	22	15
cores	95	11	144	8	16
N-cap	56	25	85	66	37
C-cap	48	29	72	79	44

are more dependent on additional signals, such as the expected periodic distance between turns (Cohen et al., 1986). As a consequence, most of the over- or underprediction (false positives or false negatives) is a result of the noise in these secondary signals. In a specific example of an overprediction (false positive), the first helix of cytochrome c' is broken by the presence of a weak turn indicator. Weak turn indications are not utilized if they are within 11 residues of strong turn indications. In this particular instance, the weak turn indications occur 15 residues from the nearest strong turn indi-

cator. Hence, the pattern is accepted as an authentic turn (Figure 6, top). Lysozyme contains an example of an underprediction (false negative). The last block should be split into two blocks, but it is not. There are some weakly hydrophilic residues in the area we would like to call a turn. While the area is sufficiently distant from the previous turn, the signal is below the threshold for accepting a weak turn indicator (Figure 6, bottom).

The patterns used to predict helix core regions also appear successful. The lack of a significant decrease in prediction accuracy is likely to indicate the recognition of general principles in helix core formation. The few false positives for helix core feature stem from a displacement or extension of the predicted helical core, either beginning too soon and/or ending too late in the sequence. It is difficult to associate this character with a specific structure or sequence phenomenon. In addition, short helices (of length 5–10) are underreported (false negatives), as are those helices with a strong hydrophilic character. Prediction of the core region of these types of helices is generally dependent on the recognition of complementary charged residue pairs or a putative tertiary helix–helix interaction site (Richmond & Richards, 1978).

The predictive capability of the N- and C-cap patterns is significantly lower than that of the helix core patterns in both the development and test sets of protein sequences. Similarly, the decrease in predictive accuracy from the development to the test set for the cap patterns is currently greater (approximately 10%–20%) than the decrease for the turn or helix core prediction. This suggests a tendency for the capping patterns to recognize specific features of the proteins in the development set instead of general principles of the amino acid sequences that initiate and terminate helices. There are no specific structural features that identify overprediction. However, the lack of crucial residues in the amino acid sequence near the site of the N- or C-cap typically characterizes underprediction. Often the capping patterns will have as a constituent one of a class of residues commonly found at the terminus of a helix. The N-cap positions in helices are often one of the residues, G, N, S, T, or D. The C-cap positions are usually G, K, H, or N. Further, proline is often a constituent of C-cap areas of sequence, appearing one or two residues after the cap position. If a helix does not begin or end in one of these residues, the likelihood of a correctly predicted helix cap is low.

These helix capping patterns suggest that one of a specific set of residues is required (but not sufficient) to initiate or terminate a helix. A synthetic approach to the mutagenesis of N-caps in barnase by Serrano and Fersht (1989) has provided some additional data on this aspect of helix structure. Threonines found at two different helix N-caps were mutated to several alternative residue types to examine the energetic contributions of different residues to helical structure stabilization. On the whole, the energetic stability provided by the alternative residues commonly found at the N-caps corroborates the statistical data presented by Richardson and Richardson (1988). However, the choice of a "best" residue to terminate a specific helix appears to be dependent on the tertiary interactions at the particular site. In this sense, the statistical data reported by Richardson and Richardson is not sufficient to completely specify the N- and C-caps.

Multiple regions can be defined by the user-specified ALPPS prediction patterns. In the ALPPS pattern for predicting helical structure, four different region types have been specified as described under Experimental Procedures and in Figure 5. These regions reflect the amount of information

Table IV: ALPPS Region Scores

protein	assigned helices	predicted helices	both ends	no NCAP	no CCAP	no ends	repechage
			Development Set				
1ccr	5	4	1	1	0	2	0
1fdh	9	9	3	1	2	3	0
2ccy	5	6	2	1	2	1	0
2cts	21	24	5	7	5	6	1
2lh1	7	8	1	5	1	1	0
2lhb	8	8	0	2	3	3	0
2lzm	11	9	4	1	1	3	0
3c2c	5	7	4	0	3	0	0
3cln	7	8	5	0	3	0	0
3cpv	7	7	4	0	1	1	1
totals	85	90	29	18	21	20	2
			Test Set				
156b	4	5	2	1	1	1	0
1ecd	8	7	2	3	1	1	0
1mbd	8	6	2	2	1	1	0
3icb	6	5	3	0	2	0	0
1cc5	4	3	1	0	1	1	0
2cyp	13	11	3	1	6	0	1
3wrp	6	5	0	1	2	2	0
1hmq	5	5	1	1	0	3	0
2tmv	6	6	1		3	2	0
3hhb	7	8	2	0	3	3	0
totals	67	61	17	9	20	14	1

Table V: ALPPS Residue by Residue Scores

protein	Q_3	C	TP	TN	FP	FN
			Development Set			
1ccr	0.71	0.42	38	41	16	16
1fdh	0.74	0.35	91	17	7	31
2ccy	0.85	0.58	91	18	11	7
2cts	0.70	0.36	220	89	48	80
2lh1	0.69	0.23	87	19	22	25
2lhb	0.66	0.23	79	21	18	32
2lzm	0.71	0.41	87	30	8	39
3c2c	0.61	0.22	48	21	31	12
3cln	0.79	0.46	90	23	14	16
3cpv	0.88	0.75	64	32	8	4
totals	0.73	0.39	895	311	183	262
trimmed	0.80	0.56	543	225	86	101
			Test Set			
156b	0.73	0.49	48	28	5	22
1cc5	0.71	0.41	33	26	13	11
1ecd	0.60	0.01	75	7	11	43
1hmq	0.64	0.12	62	11	15	25
1mbd	0.75	0.26	104	12	11	26
2cyp	0.61	0.23	103	78	62	50
2tmv	0.77	0.54	57	62	15	20
3hhb	0.78	0.45	91	20	8	22
3icb	0.69	0.15	46	6	14	9
3wrp	0.69	0.17	64	6	5	26
totals	0.69	0.33	683	256	159	254
trimmed	0.75	0.47	408	192	72	126

available to specify the extent of the predicted helix. The number of assigned regions (helices) can be evaluated against the number of predicted regions for the protein sequences examined. Table IV presents the number of each type of region identified in each protein. There was no particular relationship between the type of region predicted and the quality of the prediction for that region. Nor was there a more general relationship between the distribution of region types and the overall quality of a sequence prediction. However, some sequences appear to be more difficult to predict than others given any prediction method. For example, both cytochrome *c* peroxidase and erythrocruorin are predicted equally poorly by the methods of Chou and Fasman (1974), GOR (Garnier et al., 1978), neural networks (Kneller et al., 1990), and ALPPS. The rank orderings of the proteins by success rate using the different prediction methods were similar (data not shown).

Table V presents the results of evaluating the predicted helical regions on a residue by residue basis in comparison to helical residues defined by the automatic assignment algorithms. The Chou and Fasman algorithm (Chou & Fasman, 1974) produced a predictive accuracy (Q_3) of 65% when applied to the collection of 20 all-α protein sequences used in this study, the GOR algorithm (Garnier et al., 1978) provided an accuracy of 71%, the neural net scheme of Kneller et al. (1990) was 78% accurate, and the ALPPS algorithm provided an accuracy of 71%. Trimming the N- and C-cap locations in both the assigned sequence and the predicted sequence improved the Chou and Fasman, GOR, and neural net, and ALPPS based predictions 2%, 4%, 6%, and 6%, respectively. Most of the individual prediction scores increase as the end point locations are eliminated from consideration, but some scores stay constant or decrease. This suggests that the overall error rate in prediction stems mainly from the difficulty in assigning the N- and C-caps. It is our experience that those predictions that do not benefit from neglecting the end points were poor predictions from the start. These data give a relative indication of the ability each algorithm possesses to predict the core features of secondary structure.

The primary source of error in the ALPPS prediction of helices results from failures in the underlying PLANS patterns. These errors can be subdivided into two levels: the segmentation of the sequence into structural units, or block definition, and the specification of helices within those units, or region definition. The primary source of block definition comes from correct identification of turn location, here with the PLANS pattern TU. Failure at this level results in either the scission of a segment of regular secondary structure or the concatenation of two segments of regular secondary structure. On the basis of our previous work in turn prediction, we had anticipated and planned for these failures. We were able to describe a simple length-based heuristic for splitting exceptionally long blocks. This is analogous to the PLANS work where weak turn predictions were masked from consideration around strong turn indicators on the basis of the expected distance between turns for a given protein class. We were not able to develop a consistently accurate heuristic for recognizing appropriate situations for adjacent block concatenation.

The errors in region specification are of two types: misassignment of the helical core or misassignment of the helical end points. When one or both of the helix termini cannot be determined, helix is defined over the extent of the helical core pattern. This usually leads to underprediction but can lead to overprediction when false positive helical core signals are generated. Predicted helical termini can also be located in a manner which erroneously shortens the helical region.

Conclusions

We have developed PLANS patterns that recognize individual components of secondary structure. In the work presented here, we present concepts used to recognize the distinct structural components of α-helices: N-cap, core region, and C-cap. Currently we can identify almost all of the helical regions via their core structure features. However, we can identify only some of the N-cap and C-cap positions with certainty.

With the success of the pattern-based turn prediction, it was our expectation that the predicted location of turns would be a useful basis for subdividing a protein sequence into regions for independent evaluation. This expectation has been fulfilled. A language to facilitate a segment-based approach to the prediction of regular secondary structure, ALPPS, has been designed and implemented. Initial pattern development was performed on a database of 10 α/α proteins. These patterns were applied to an independent, nonhomologous data set. The results presented here compare favorably with the current secondary structure prediction algorithms (Table VI). The ALPPS method is similar to the GOR method when all residues are considered. However, when the scoring is focused on the core of the helical segments, the ALPPS algorithm fares better. On the basis of individual residue scores, the neural network algorithm and weights developed by Kneller et al. perform 3% better than the ALPPS system.

Table VI: Comparison of Prediction Methods for All Helical Proteins[a]

method	raw Q_3 (C)	trimmed Q_3 (C)	reference
Chou/Fasman	0.65 (0.23)	0.68 (0.30)	Chou & Fasman, 1974
Garnier et al.	0.71 (0.36)	0.75 (0.46)	Garnier et al., 1978
Qian et al.	0.67 (nc)	nc (nc)	Qian & Sejnowski, 1988
Kneller et al.	0.79 (0.55)	0.85 (0.69)	Kneller et al., 1990
this work	0.71 (0.36)	0.78 (0.52)	

[a] nc = not calculated. The values determined for Chou and Fasman, Garnier et al., and the current work were taken by applying the respective algorithms to the 20-protein data set described under Experimental Procedures and comparing that to the assignments as generated by the algorithm of Richards and Kundrot. The weights generated in the neural net simulation by Kneller et al. were also used in a run against the 20-protein data set but were compared against the helical assignment as determined by the algorithm of Kabsch and Sander, as this assignment was used in the training of the neural network. The value for Qian and Sejnowski were taken unmodified from Kneller et al. (1990).

Effective, practical use of secondary structure prediction methods is facilitated by ability to determine the basis for individual predictions. With methods based on statistical analyses, this determination is infeasible because the data are primarily numerical distributions. The problems are similar with computational neural network learning algorithms. However, the database of rules contained in a pattern-matching system such as ALPPS is interpretable by the experimentalist. This allows the experimentalist to perform an evaluation of confidence based on the biochemical knowledge incorporated into the pattern. Secondary structure prediction then over-comes the concept of a "black box" procedure.

Research has suggested that several secondary structure prediction algorithms may in fact be useful when used in concert (Nishikawa & Ooi, 1986). Predictions with the ALPPS system and the neural network software often complement each other. When the neural network software incorrectly predicts a particular region, often the ALPPS evaluation will be closer to the targeted assignment, and vice versa. In future work, we plan to integrate the two methods to obtain the best information from each package.

Work on extensions to the ALPPS method is also underway. PLANS patterns that recognize the components of β-structure and the ALPPS patterns that would coordinate the prediction of β-structure regions are currently under development. The eventual incorporation of combined α and β component structure patterns for the set of α/β proteins is also anticipated. We have only explored the lower half of the structural hierarchy. We are now incorporating concepts into the syntax of ALPPS that will facilitate the description of higher level concepts in protein structure. These concepts will include supersecondary structure and motifs such as four-helix bundles (Presnell & Cohen, 1989) and nucleotide binding folds (Rao & Rossman, 1973).

One of the innovative features incorporated into the ALPPS language was the ability to utilize the sequential ordering of substructure features as recognized by PLANS patterns. We

have exploited that aspect of the ALPPS metapattern language to construct the first example of a predictive algorithm for regular secondary structure that explicitly incorporates nonlocal sequence–structure correlations. The results presented here strongly suggest that processing the sequence into structurally reasonable segments can provide an advantage over nonhierarchical methods of prediction.

The developments in predictive schemes have also brought to our attention the complexities in describing the target. Scoring algorithms that focus on the prediction of individual residues have been used as a standard of comparison for several years. One artifact of these algorithms is that each residue is considered equally important. However, researchers and algorithms cannot agree among themselves on the precise location of the terminal residues of regular secondary structure. Moreover, the semiempirical condensation methods of model construction, which use secondary structure as input, are insensitive the exact end or beginning of a particular stretch of secondary structure (Cohen & Kuntz, 1989). Of much greater importance is the number and location of the individual secondary structure features. The trimming technique presented here affords one method to examine this issue. We are continuing to develop new scoring algorithms that focus on the segmented, feature-oriented nature of regular secondary structure. This should provide an evaluation technique that indicates the utility of secondary structure predictions for subsequent steps in the modeling process.

Supplementary Material Available

Two tables showing the exact patterns used in the ALPPS algorithm (15 pages). Ordering information is given on any current masthead page. A copy of this material will be provided by the authors upon request.

References

Arutyunyan, E. G., Kuranova, I. P., Vainshtein, B. K., & Steigemann, W. (1980) *Kristallografiya 25*, 80.

Babu, Y. S., Bugg, C. E., & Cook, W. J. (1988) *J. Mol. Biol. 204*, 191–204.

Bernstein, F. C., Koetzle, T. F., Williams, G. J. B., Meyer, E. F., Jr., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., & Tasumi, M. (1977) *J. Mol. Biol. 112*, 535–542.

Beveridge, D. L., & Jorgenson, W. L., Eds. (1986) Computer Simulation of Chemical and Biomolecular Systems, *Ann. N.Y. Acad. Sci. 482*.

Bhatia, G. E. (1981) Ph.D. Thesis, University of California, San Diego, CA.

Bowie, J. U., Clarke, N. D., Pabo, C. O., & Sauer, R. T. (1990) *Proteins: Struct., Funct., Genet. 7*, 257–264.

Carter, D. C., Melis, K. A., O'Donnell, S. E., Burgess, B. K., Furey, W. F., Jr., Wang, B. C., & Stout, C. D. (1985) *J. Mol. Biol. 184*, 279–295.

Chou, P. Y., & Fasman, G. D. (1974) *Biochemistry 13*, 211–221.

Chou, P. Y., & Fasman, G. D. (1978) *Annu. Rev. Biochem. 47*, 251–276.

Cohen, B. I., Presnell, S. R., Morris, M., Langridge, R., & Cohen, F. E. (1991) in *Proceedings of the Hawaii International Conference on System Sciences* (Milutinovic, V., & Shriver, B. D., Eds.) Vol. 1, pp 574–584, IEEE Computer Society Press, Los Alamitos, CA.

Cohen, F. E., & Kuntz, I. D. (1989) in *Prediction of Protein Structure and the Principles of Protein Conformation* (Fasman, G. D., Ed.) pp 647–705, Plenum Publishing Corp., New York.

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 285*, 378–382.

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. (1986) *Biochemistry 25*, 266–275.

Deleage, G., & Roux, B. (1987) *Protein Eng. 1*, 289–294.

Fermi, G., Perutz, M. F., Shaanan, B., & Fourme, R. (1984) *J. Mol. Biol. 175*, 159–174.

Ferrin, T., Huang, C., Jarvis, L., & Langridge, R. (1988) *J. Mol. Graphics 6*, 13–37.

Finzel, B. C., Poulos, T. L., & Kraut, J. (1984) *J. Biol. Chem. 259*, 13027–13036.

Finzel, B. C., Weber, P. C., Hardman, K. D., & Salemme, F. R. (1985) *J. Mol. Biol. 186*, 627–643.

Frier, J. A., & Perutz, M. F. (1977) *J. Mol. Biol. 112*, 97–112.

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

Gibrat, J. F., Garnier, J. R., & Robson, B. (1987) *J. Mol. Biol. 198*, 425–443.

Goto, Y., & Fink, A. L. (1990) *J. Mol. Biol. 214*, 803–5.

Holley, L. H., & Karplus, M. (1989) *Proc. Natl. Acad. Sci. U.S.A. 86*, 152–156.

Honzatko, R. B., Hendrickson, W. A., & Love, W. E. (1985) *J. Mol. Biol. 184*, 147–164.

Hughson, F. M., Wright, P. E., & Baldwin, R. L. (1990) *Science 249*, 1544–1548.

Johnson, W. C., Jr., (1990) *Proteins: Struct., Funct., Genet. 7*, 205–214.

Kabsch, W., & Sander, C. (1983) *Biopolymers 22*, 2577–2637.

Kabsch, W., & Sander, C. (1984) *Proc. Natl. Acad. Sci. U.S.A. 81*, 1075–1078.

Karplus, M., & McCammon, J. A. (1981) *CRC Crit. Rev. Biochem. 9*, 293–349.

Klein, P., & Delisi, C. (1986) *Biopolymers 25*, 1659–1672.

Kneller, D. G., Cohen, F. E., & Langridge, R. (1990) *J. Mol. Biol. 214*, 171–182.

Lawson, C. L., Zhang, R., Schevitz, R. W., Otwinowski, Z., Joachimiak, A., & Sigler, P. B. (1988) *Proteins: Struct., Funct., Genet. 3*, 18–31.

Lederer, F., Glatigny, A., Bethge, P. H., Bellamy, H. D., & Mathews, F. S. (1981) *J. Mol. Biol. 148*, 427–448.

Lee, D. C., Haris, P. I., Chapman, D., & Mitchell, R. C. (1990) *Biochemistry 29*, 9185–9193.

Levin, J. M., & Garnier, J. (1988) *Biochim. Biophys. Acta 955*, 283–295.

Levitt, M., & Warshel, A. (1975) *Nature 253*, 694–698.

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

Lim, V. I. (1974a) *J. Mol. Biol. 88*, 857–872.

Lim, V. I. (1974b) *J. Mol. Biol. 88*, 873–894.

Matthews, B. W. (1975) *Biochim. Biophys. Acta 405*, 442–451.

McCammon, J. A., Gelin, B. R., & Karplus, M. (1977) *Nature 267*, 585–590.

Moews, P. C., & Kretsinger, R. H. (1975) *J. Mol. Biol. 91*, 201–225.

Nagano, K. (1973) *J. Mol. Biol. 75*, 401–420.

Nagano, K. (1974) *J. Mol. Biol. 84*, 337–372.

Namba, K., Pattanayek, R., & Stubbs, G. (1989) *J. Mol. Biol. 208*, 307–325.

Nemethy, G., & Scheraga, H. A. (1977) *Q. Rev. Biophys. 10*, 239–352.

Nishikawa, K., & Ooi, T. (1986) *Biochim. Biophys. Acta 871*, 45-54.
Ochi, H., Hata, Y., Tanaka, N., Kakudo, M., Sakurai, T., Aihara, S., & Morita, Y. (1983) *J. Mol. Biol. 166*, 407-418.
Phillips, S. E. V., & Schoenborn, B. P. (1981) *Nature 292*, 81-82.
Presnell, S. R., & Cohen, F. E. (1989) *Proc. Natl. Acad. Sci. U.S.A. 86*, 6592-6596.
Ptitsyn, O. B., Pain, R. H., Semisotnov, G. V., Zerovnik, E., & Razgulyaev, O. I. (1990) *FEBS Lett. 262*, 20-24.
Qian, N., & Sejnowski, T. J. (1988) *J. Mol. Biol. 202*, 865-884.
Rao, S. T., & Rossman, M. G. (1973) *J. Mol. Biol. 76*, 241-256.
Rees, D. C., Deantonio, L., & Eisenberg, D. (1989) *Science 245*, 510-513.
Remington, S., Wiegand, G., & Huber, R. (1982) *J. Mol. Biol. 158*, 111-152.
Richards, F. M., & Kundrot, C. E. (1988) *Proteins: Struct., Funct., Genet. 3*, 71-84.
Richardson, J. S., & Richardson, D. C. (1988) *Science 240*, 1648-1652.
Richmond, T. J., & Richards, F. M. (1978) *J. Mol. Biol. 119*, 537-555.
Roder, H., Eloeve, G. A., & Englander, S. W. (1988) *Nature 335*, 700-704.
Schiffer, M., & Edmundson, A. B. (1968) *Biophys. J. 8*, 29-39.
Schulz, G. E. (1988) *Annu. Rev. Biophys. Biophys. Chem. 17*, 1-21.
Schulz, G. E., & Schirmer, R. H. (1979) *Principles of Protein Structure*, Springer-Verlag, New York.
Serrano, L., & Fersht, A. R. (1989) *Nature 342*, 296-299.
Sheridan, R. P., Dixon, J. S., Venkataghavan, R., Kuntz, I. D., & Scott, K. P. (1985) *Biopolymers 24*, 1995-2023.
Sikorski, A., & Skolnick, J. (1989) *Proc. Natl. Acad. Sci. U.S.A. 86*, 2668-2672.
Sklenar, H., Etchebest, C., & Lavery, R. (1989) *Proteins: Struct., Funct., Genet. 6*, 46-60.
Skolnick, J., & Kolinski, A. (1989) *Annu. Rev. Phys. Chem. 40*, 207-235.
Skolnick, J., & Kolinski, A. (1990) *Science 250*, 1121-1126.
Smith, R. F., & Smith, T. F. (1990) *Proc. Natl. Acad. Sci. U.S.A. 87*, 118-122.
Steigemann, W., & Weber, E. (1979) *J. Mol. Biol. 127*, 309-338.
Stenkamp, R. E., Sieker, L. C., & Jensen, L. H. (1983) *Acta Crystallogr., B 39*, 697-703.
Sternberg, M. J., & Gullick, W. J. (1990) *Protein Eng. 3*, 245-248.
Szebenyi, D. M. E., & Moffat, K. (1986) *J. Biol. Chem. 261*, 8761-8777.
Udgaonkar, J. B., & Baldwin, R. L. (1988) *Nature 335*, 694-699.
Weaver, L. H., & Matthews, B. W. (1987) *J. Mol. Biol. 193*, 189-199.
Weiner, S. J., Kollman, P. A., Case, D. A., Singh, U. C., Ghio, C., Alagona, G., Profeta, S., & Weiner, P. (1984) *J. Am. Chem. Soc. 106*, 765-784.

Triple-Helix Formation Is Compatible with an Adjacent DNA–Protein Complex

Chiao-Chain Huang,[‡] Dat Nguyen,[‡] Ricardo Martinez,[§] and Cynthia A. Edwards*[,§]

Drug Discovery Systems Research Group and Molecular Virology Group, Genelabs Incorporated, 505 Penobscot Drive, Redwood City, California 94063

Received August 21, 1991; Revised Manuscript Received October 24, 1991

ABSTRACT: The effect of oligonucleotide-directed triple-helix formation on the binding of a protein to an immediately adjacent sequence has been examined. A double-stranded oligonucleotide was designed with a target site for the binding of a pyrimidine oligonucleotide located immediately adjacent to the recognition sequence for the herpes simplex virus type 1 (HSV-1) origin of replication binding protein, which is encoded by the UL9 gene of HSV-1. Since the optimal conditions for the binding of the UL9 protein and the pyrimidine oligonucleotide to the duplex DNA are markedly different, a pyrimidine oligonucleotide was designed to optimize binding affinity and specificity for the target duplex oligonucleotide. Consideration was given to length and sequence composition in an effort to maximize triple-strand formation under conditions amenable to the formation of the UL9–DNA complex. Using gel mobility shift assays, a trimolecular complex composed of duplex DNA bound to both a third oligonucleotide strand and the UL9 protein was detected, indicating that the UL9–DNA complex is compatible with the presence of a triple helix in the immediately adjacent sequences.

Sequence-specific recognition of double-stranded DNA by proteins is essential for the regulation of many cellular functions including replication, recombination, and transcription. Displacement of DNA-bound regulatory proteins from their recognition sites might provide a general strategy for the alteration of sequence-specific functions in eukaryotes. In principle, DNA–protein interactions could be disrupted by the presence of a DNA-binding molecule within or near the protein recognition sequence. Possible mechanisms of interference include steric hindrance or localized distortions in the physical

* To whom correspondence should be addressed.
‡ Molecular Virology Group.
§ Drug Discovery Systems Research Group.

0006-2960/92/0431-993$03.00/0