Article No. mb981943 *J. Mol. Biol.* (1998) **281**, 565–577

JMB

Prediction of Local Structure in Proteins Using a Library of Sequence-Structure Motifs

Christopher Bystroff* and David Baker*

Department of Biochemistry
University of Washington
Seattle, WA 98195-7350
USA

**Corresponding authors*

We describe a new method for local protein structure prediction based on a library of short sequence pattern that correlate strongly with protein three-dimensional structural elements. The library was generated using an automated method for finding correlations between protein sequence and local structure, and contains most previously described local sequence-structure correlations as well as new relationships, including a diverging type-II β-turn, a frayed helix, and a proline-terminated helix. The query sequence is scanned for segments 7 to 19 residues in length that strongly match one of the 82 patterns in the library. Matching segments are assigned the three-dimensional structure characteristic of the corresponding sequence pattern, and backbone torsion angles for the entire query sequence are then predicted by piecing together mutually compatible segment predictions. In predictions of local structure in a test set of 55 proteins, about 50% of all residues, and 76% of residues covered by high-confidence predictions, were found in eight-residue segments within 1.4 Å of their true structures. The predictions are complementary to traditional secondary structure predictions because they are considerably more specific in turn regions, and may contribute to *ab initio* tertiary structure prediction and fold recognition.

Keywords: I-sites library; sequence patterns; clustering; initiation sites; protein folding

Introduction

Although almost all local structure prediction methods have focused on three-state (helix, strand and loop) secondary structure prediction, there clearly are relationships between amino acid sequence and more specific local structure elements. Most attempts to identify such relationships have proceeded by identifying a common structural motif, then characterizing the frequencies of occurrence of each amino acid at each position in the motif (Aurora *et al.*, 1994; Chan *et al.*, 1993; Efimov, 1993; Hutchinson & Thornton, 1994; Jiménez *et al.*, 1994; Muñoz & Serrano, 1995; Unger & Sussman, 1993; Zhu & Blundell, 1996). In particular, sequence patterns for tight turns (Hutchinson & Thornton, 1994; Yang *et al.*, 1996) and helix caps (Aurora *et al.*, 1994; Donnelly *et al.*, 1994; elMasry & Fersht, 1994; Jiménez *et al.*, 1994; Muñoz *et al.*, 1995) have been described. More comprehensive approaches have clustered structural segments into classes using measures of structural similarity, and then tabulated sequence preferences for each of the classes (Oliva *et al.*, 1997; Rooman *et al.*, 1990; Unger *et al.*, 1989). However, most of the sequence-structure correlations identified using this approach have not been developed into local structure prediction methods, perhaps because the sequence-structure relationships have not always been strong. The challenge for local structure prediction is to identify the structural features that have strong sequence preferences.

A systematic approach for identifying such structural features has been described (Han & Baker, 1995, 1996). The approach is based on the idea that if there are a finite number of different local structural elements in proteins, and each structural element has a distinct set of preferences for the different amino acids, then there should be a finite number of distinct local sequence patterns in multiple sequence alignments. Recurrent amino acid sequence patterns that transcend protein family boundaries were identified by clustering sequence segments from a large set of proteins of known structure and, as anticipated, many of the

* E-mail address of the corresponding authors:
bystroff@ben.bchem.washington.edu
baker@ben.bchem.washington.edu

0022–2836/98/330565–13 $30.00/0

sequence patterns were found to occur primarily in a single type of local structure. The advantage of unsupervised learning approaches such as this one is that, since the important properties are not specified in advance, new patterns and groupings can readily be identified; however, the groupings are generally not optimal for prediction (Duda & Hart, 1973).

Here, we make use of insights gained during characterization of the structures adopted by the sequence patterns (Han *et al*., 1997) to develop a procedure that utilizes structural information to increase the structural selectivity of the sequence patterns. The procedure may be viewed as a combination of previous sequence-based and structure-based clustering approaches (Han & Baker, 1995, 1996; Oliva *et al*., 1997; Rooman *et al*., 1900; Unger and Sussman, 1993). Starting with sequence-based clusters, the most frequently occurring structure in each cluster is chosen as the structural ''paradigm''. We then iteratively exclude members with structures different from the paradigm from the cluster, recalculate the sequence pattern (profile) from the remaining members, and search for new members in the database. The result of this refinement procedure, the I-sites library, consists of 82 profiles that can be roughly grouped into 13 different sequence-structure motifs. Predictions of local structure using the library are more specific than and complementary to traditional three-state secondary structure predictions.

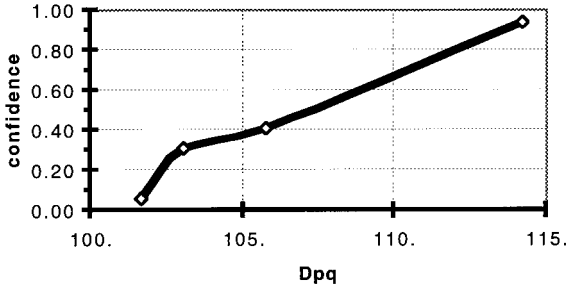


Figure 1. A confidence curve maps similarity score to the probability of correct local structure based on a ten-fold jack-knife test. All nine-residue segments in a test set composed of 10% of the database were scored using the profile for a nine-residue ''serine-containing β-hairpin'' cluster, which had been refined using the remaining 90% of the database, and the top-scoring 40 segments were kept. The structures of the top-scoring segments were compared to the paradigm structure for the cluster, chosen from the 90% training set (e.g. 2bbkH 346-354). The list was sorted by score and the fraction true-positives determined in bins of 30 (◇, highest four bins are shown). The refinement was repeated ten times using a different 10% as the test set and the resulting curves were averaged. A plot such as this was generated for each cluster, and used to translate scores into confidences.

Results

As described in detail in Methods, sequence segments from 471 proteins of known structure were partitioned into clusters, and the clusters were then refined using structural information to produce the I-sites library. Each cluster is represented by a log odds scoring matrix (a sequence profile) and the backbone torsion angles of the paradigm structure. We first describe the results of local structure predictions using the library, and then provide a brief overview of the components of the library, focusing on the more novel sequence-structure relationships.

Local protein structure prediction

To predict protein structure using the I-sites library, profiles for each of the 82 clusters were used to score all sub-fragments of the target sequence(s). Because of the differences in length, the similarity scores of different clusters were not directly comparable; instead we compared the associated ''confidence'' values. The confidence of a fragment prediction is the probability that a sequence segment with a given score has the predicted structure; confidence curves such as that shown in Figure 1 were derived for each cluster using the jack-knife procedure described in Methods. The fragment predictions were sorted by confidence, then the backbone torsion angles of the target sequence were assigned to be those of the paradigm for the fragments at the top of the list (see Methods, Algorithm 2). In the following paragraphs we present prediction results for the training set of 471 sequence families and for an independent test set of 55 sequence families.

The evaluation of local three-dimensional structure predictions requires a choice of the length scale over which the predicted and true structure must agree (Lesk, 1997). We chose to evaluate the I-sites torsion angle predictions using an eight residue window because this is the average length of a cluster. Eight-residue segments were considered to be predicted correctly if none of the predicted torsion angles differed by more than 120° from those of the true structure (*mda* measure) or if the *rmsd* between the predicted and corrected structure was less than 1.4 Å (*rmsd* measure). To avoid counting the same position more than once, the Tables report the number of positions for which at least one of the overlapping eight-residue segments was predicted correctly. The *mda* measure was used in addition to the *rms*, because changes in *mda* were found to better correlate with loss of conserved contacts than changes in *rmsd* (see Methods). The 120° *mda* cutoff for correctness is based on the observed natural boundaries of the clusters (see Natural boundaries and Figure 2). Since the predictions consist of fragments taken from the database, only very rarely does an incorrect prediction have all of its backbone angles within 120° of the true angles. The average backbone *rmsd* between pairs of 8-mers having an *mda* less than 120° is 1.0 Å, *versus* 2.5 Å for 8-mers with *mda* greater than 120.

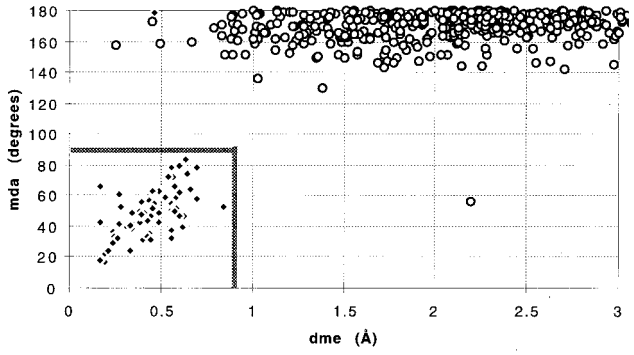


Figure 2. Deviations from the paradigm structure in *dme* and *mda* for the top 600 scores in the database for one cluster (a nine-residue serine β-hairpin). A clear separation (natural boundary) appears between the segments that conserve four specific side-chain:side-chain contacts and two specific backbone H-bonds (◆, true-positives) and those that do not (○, false-positives). *dme* is the distance matrix error, and *mda* is the maximum deviation in backbone angles, as defined in equations (2) and (3), measured against the paradigm (see Methods). These two metrics, especially *mda*, adequately substitute for the specific contacts filter (which was not automated). True/false limits (thick lines) may be chosen automatically, taking advantage of the natural boundaries, whose presence was a condition for keeping a cluster.

Training set

Predictions were made for each of the proteins in the database, totaling 122,510 residues. About 95% of the 471 sequence families had at least one match to a sequence pattern with a confidence of 0.80 or better, and all but one sequence family had a match with a confidence of at least 0.40. Approximately 40% of the residues in the database were included in at least one fragment prediction with confidence greater than 0.60, and these predictions were 71% correct using the *mda* measure (Table 1, first two rows). This measure is considerably more strict than the commonly used ''Q3'' score, which measures the number of positions correctly assigned to one of three states (the probability that a pair of 8-mers, chosen at random, have *mda* less than 120° is about 1 in 20, and secondary structure predictions having an average Q3 score of 70% have an *mda* score of about 45%). Overall, 50% of the positions were found in correctly predicted eight-residue fragments, and prediction accuracy correlated well with the confidence (Table 1).

Comprehensive cross-validation would involve re-discovery of the sequence-structure motifs starting from the initial clustering procedure (see Methods). This strategy was precluded by the very large amount of cpu time required. Instead, cross-validation was carried out by removing the contribution of 10% of the sequence families before refining the sequence profiles, and then predicting the structure of this 10%. The largest decrease in the percentage correct upon removing the contribution of a subset to the sequence profiles was 2%. Because of the large number of sequence segments in the training set for each of the clusters, we do not believe that the results would change significantly if the library was completely rebuilt for each of the cross-validation tests. The results with the test set (see below) and *bona fide* blind predictions for the CASP2 structure prediction experiment (Bystroff & Baker, 1997) using an earlier version of the library further confirm the absence of significant database bias.

Test set

The accuracy of the backbone torsion angle predictions for a test set of 55 protein families, all unrelated to sequences in the training set, was only slightly worse than that of the training set. Of the eight-residue segments covered by predictions with a confidence greater than 0.8, 75% were within 1.4 Å of the true structure. As in the training set, the confidence of the predictions correlated well with their accuracy (Table 2).

Table 1. I-sites structure prediction for the training set and the test set

	Training set		Test set		
				%correct	
Confidence	Residues	%correct (*mda*)	Residues	*mda*	*rmsd*
0.8–1.0	17,394	89	887	75	76
0.6–0.8	33,136	61	2643	65	67
0.4–0.6	46,767	40	8346	48	49
0.2–0.4	18,748	28	2973	35	35
0.0–0.2	6465	15	264	25	30
Totals	122,510	50	15,919	48	50

Predictions of local structure using the I-sites library. The fraction of residues predicted correctly is reported as a function of prediction confidence, for the entire database (training set) of 471 protein families and for an independent test set of 55 proteins (see Methods). The percentage correct was assessed using either the *mda* or the *rmsd* over eight-residue segments; the cutoffs were 120° and 1.4 Å, respectively. For example, using the *mda* measure, %correct is the percentage of positions that fall into at least one eight-residue segment with no backbone angle deviation greater than 120°. The average percentage correct correlates with the confidence. Little bias is observed toward the training set.

Table 2. Comparison of I-sites and PHD for the test set

Confidence	No. of residues	Percent correct Method I-sites	PHD	Combined
A. *All-α (eight proteins)*				
0.8–1.0	95	51	34	60
0.6–0.8	376	55	62	67
0.4–0.6	1312	41	56	55
0.2–0.4	356	28	53	47
0–0.2	128	23	43	40
Total	2267	40	55	55
B. *All-β (six proteins)*				
0.8–1.0	42	79	33	79
0.6–0.8	145	59	42	56
0.4–0.6	483	54	32	52
0.2–0.4	181	40	25	44
0–0.2	80	29	24	38
Total	931	51	32	51
C. *αβ, α+β, and multidomain proteins (41 proteins)*				
0.8–1.0	750	78	48	77
0.6–0.8	2121	67	55	71
0.4–0.6	6551	49	42	54
0.2–0.4	2436	35	31	42
0–0.2	863	24	23	33
Total	12,721	50	41	54
D. *All proteins (55)*				
0.8–1.0	887	75	46	75
0.6–0.8	2642	65	55	69
0.4–0.6	8346	48	44	54
0.2–0.4	2973	35	33	43
0–0.2	1071	25	26	34
Total	15,919	48	43	54

The results of predictions for 55 sequence families in an independent test set are compared to secondary structure predictions. The percentage correct was measured using *mda* for predictions made by the I-sites method, the PHD server (Rost *et al*., 1994) and an optimized combination. For the ''combined'' predictions, the following formula was used to choose which method to use at each residue:

$$\text{if}\begin{Bmatrix}\text{H\&}(0.2r-0.30) > cf\\ \text{E\&}(0.3r+0.05) > cf\end{Bmatrix}\text{use PHD}$$

where r is PHD's reliability (0 to 9), cf is I-sites' weighted confidence (0.0 to 1.8). Thus, most PHD predictions of helix (H) were used if the reliability was over 6 and most sheet (E) predictions were used if the reliability was over 3. PHD loop predictions were not used in the combined approach. The test set is broken down into (A) eight all-α-helix proteins, (B) six all-β-sheet proteins, and (C) 41 others. PHD performed best on α-helix proteins, while I-sites did better on β-sheet proteins. The two methods were the most complementary when both types of secondary structure were present.

Combination of I-sites and conventional secondary structure predictions

To determine whether I-sites local structure predictions are complementary to three-state secondary structure predictions (Rost & Sander, 1993a,b, 1994), the secondary structure of the proteins in the test set was predicted using the PHD server (Rost *et al*., 1994). Comparison of the results of the two methods requires either translating the I-sites torsion angle predictions into three-state secondary structure predictions, or converting the PHD predictions into torsion angle predictions. Not surprisingly, each method performs best on the problem it was optimized to solve: three-state secondary structure prediction *versus* backbone angle prediction. The average Q3 score using three-state assignments from the I-sites paradigms was only 64%; well below the standard set by PHD. The ''failure'' could be traced to underpredicting loop states and overpredicting strand, whose DSSP definitions (Kabsch & Sander, 1983) are based, in part, on non-local interactions. On the other hand, a crude translation of PHD loop positions to backbone angles reproduces none of the detail of the turns predicted by I-sites, and the predictions at the level of torsion angles were considerably worse using PHD (compare columns 4 and 5 in Table 2).

The most straightforward combination of the two methods is to simply substitute all PHD loop positions with the corresponding I-sites backbone angles. This combination consistently outperformed PHD, and we believe is a significant improvement over current methods for predicting local structure at a resolution higher than conventional secondary structure prediction. A slightly better combination of the methods was obtained by replacing the PHD helix and strand positions of low reliability by the I-sites prediction (see Table 2).

The I-sites library

Because of space considerations, the 82 individual clusters that make up the library cannot be presented in detail. Instead, they were structurally aligned and grouped into 13 classes (motifs) for analysis. An example of the grouping of clusters into motifs is presented in Table 3; the five clusters in this motif share a common structure and sequence pattern, differing only in register and length. Since the structural filter used in refining the clusters was based exclusively on local structural information, segments with differing DSSP assignments occasionally appeared in the same motif (compare the first and fifth clusters in Table 3). We describe first the sequence and structural features of the more novel of the motifs and then briefly summarize the properties of the others.

Diverging type-II beta-turn

The sequence pattern for this and the other novel motifs are summarized in the log odds matrices on the left side of Figures 3 to 9, the variability in backbone conformation is shown in the structural superpositions in the center, and the key interactions are highlighted on the right. The diverging β-turn (motif 13 in Table 4) contains a conserved Pro-Gly-Asx sequence (positions 3 to 5 in Figure 3), which forms a type-II β-turn followed by a fairly conserved H-bond between the Asx side-chain and the backbone nitrogen atom three residues before it. The diverging geometry is stabilized by the inwardly turned polar side-chain (Asx) and a hydrophobic contact between two side-chains six residues apart. This motif extends the four-residue

Table 3. The five clusters belonging to the diverging turn motif

Cluster ID	Boundaries *mda* (°)	Boundaries *dme* (Å)	Paradigm		2° struct.	Consensus seq.
9024	80	1.07	1eft_	247	EELLLEEEE	LKPGD·V·F
9055	80	1.02	1cpt_	333	LLLLEEEEL	KPGD·VTI·
8300	103	1.00	1mat_	91	LLEEEEEE	GQPVTIDC
7410	80	0.83	2pmgA	496	LLEEEEL	GKPVII·
6923	84	1.06	1qorA	138	LLLLLE	LPPGD·

Five of the 82 clusters in the I-sites library correspond to the ''diverging turn'' structural motif, a type-II β-turn with non-pairing β-strands on either side. Each cluster has a paradigm and two structural boundaries (*mda* and *dme*).

type-II β-turn pattern described by Hutchinson, with the difference that Asx rather than serine is preferred after the glycine residue (Hutchinson & Thornton, 1994). The sequence pattern differs from that of the structurally related ''β-β arch'' described by Efimov ($\beta\beta\alpha_L\beta$ half-turn; Efimov, 1993). The transposition of just two residues (DG→GD) causes the change from a type-I hairpin (below) to a diverging type-II turn. There is experimental evidence for a structure resembling this motif in a short peptide whose sequence matches this pattern (Sieber & Moe, 1996).

Type-I β-hairpin

This motif (12 in Table 4) contains the conserved sequence Pro-Asx-Gly (residues 6 to 8 in Figure 4) where the glycine residue has a positive phi angle ($\beta\alpha\gamma\alpha_L\beta$-turn). Previous descriptions of this turn (Efimov, 1993; Hutchinson & Thornton, 1994) have not included a detailed description of the extended sequence pattern. A conserved Asx side-chain makes a hydrogen bond to the backbone nitrogen atom two residues after it. Hydrophobic residues two before the Pro and two after the Gly initiate the β-sheet pairing; between them there are five polar positions in a row, precluding the stable formation of either α-helix or β-sheet. A short peptide that matches this sequence pattern was found to adopt this hairpin conformation in solution (de Alba *et al*., 1996).

Frayed α-helix

This motif (6 in Table 4) consists of an α-helix that begins to unwind at the C terminus (Figure 5). The frayed helix sequence pattern is summarized as NPPxNPPxN, where N is non-polar and P is polar, as compared to NPPNxPx for the α-helix heptad repeat. If this sequence pattern were folded into a normal α-helix, the final non-polar side-chain would fall on the polar side of the helix; the helix must ''fray'' in order to align the hydrophobic side-chains. A histidine residue, uncommon in helices in general, sometimes hydrogen bonds to the first unpaired carbonyl oxygen atom.

Table 4. A summary of the sequence-structure motifs in the I-sites library

	Motif	Number of clusters	Sites/100 positions Overall	Sites/100 positions Confid. > 0.60	Average boundaries *mda* (°)	Average boundaries *dme* (Å)	Average *rmsd* (len)	Pattern of conserved non-polar residues
1	Amphipathic α-helix	13	3.1	0.9	56	0.71	0.78 (15)	1-4-8, 1-5-8
2	Non-polar α-helix	6	0.9	0.12	54	0.58	0.40 (11)	1-4-8, 1-5-8
3	Schellman cap type 1	6	0.09	0.07	81	1.01	1.02 (15)	1-6-9-11
4	Schellman cap type 2	10	0.3	0.14	76	0.94	0.94 (15)	1-6-8-9
5	Proline α-helix C cap	10	1.8	0.6	92	1.07	0.89 (13)	1-2-5-8
6	Frayed α-helix	2	1.2	0.13	75	0.96	0.69 (15)	1-5-9-13
7	Helix N capping box	10	1.1	0.6	99	0.95	0.65 (15)	1-6-9-13
8	Amphipathic β-strand	8	6.8	2.1	89	0.87	0.87 (6)	1-3, 1-3-5
9	Hydrophobic β-strand	5	2.3	0.3	101	0.91	0.91 (7)	1-2-3
10	β-Bulge	2	0.5	0.15	100	0.97	0.78 (7)	1-4-6
11	Serine β-hairpin	4	1.3	0.3	94	0.76	0.81 (9)	1-8
12	Type-I hairpin	2	0.07	0.04	80	0.94	1.23 (13)	1-7-8
13	Diverging type-II turn	4	0.3	0.14	87	1.04	1.00 (9)	1-7-9

Each grouping (motif) consists of between 2 and 13 clusters, each having related sequence patterns and structures. There are two sequence groups each of α-helix and β-strand. The frequency of sites per 100 residues was estimated as the number of segments of unbroken true predictions (within the *mda/dme* boundaries) of the motif at a minimum confidence of 0.00 (all occurrences) or 0.60 (high confidence occurrences). The average boundaries are the averages of the natural structural boundaries for the clusters within each motif. To indicate the precision described by the structural boundaries in terms of the more familiar *rmsd* measure, the longest cluster for each motif was chosen, and the *rmsd* (all backbone atoms) to the paradigm was averaged over all true positives. The length of that cluster is in parentheses. The last column shows the pattern of conserved non-polar side-chains found within each motif. No two local structure types have the same pattern, consistent with the idea that hydrophobic patterning partially determines local structure (West & Hecht, 1995).

Figure 3. Diverging Type-II turn. (1eft_ 247-255)

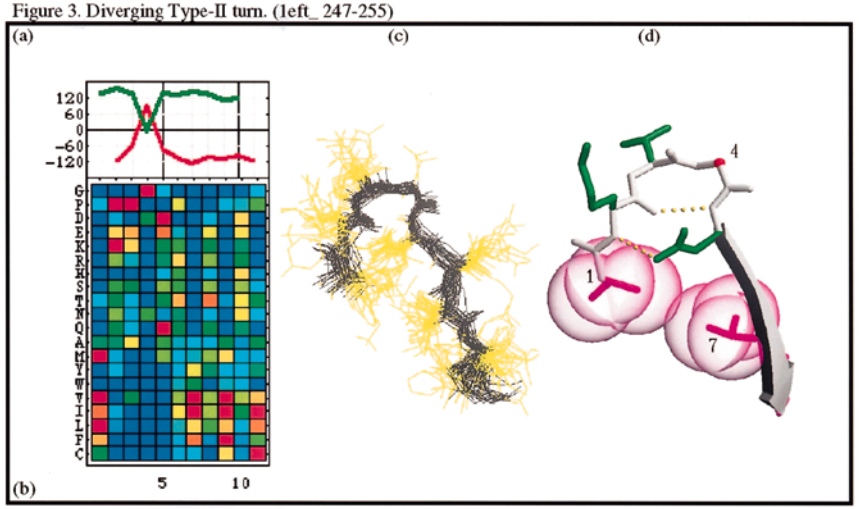


Figure 4. Type-I β-hairpin. (2bbkH 179-191)

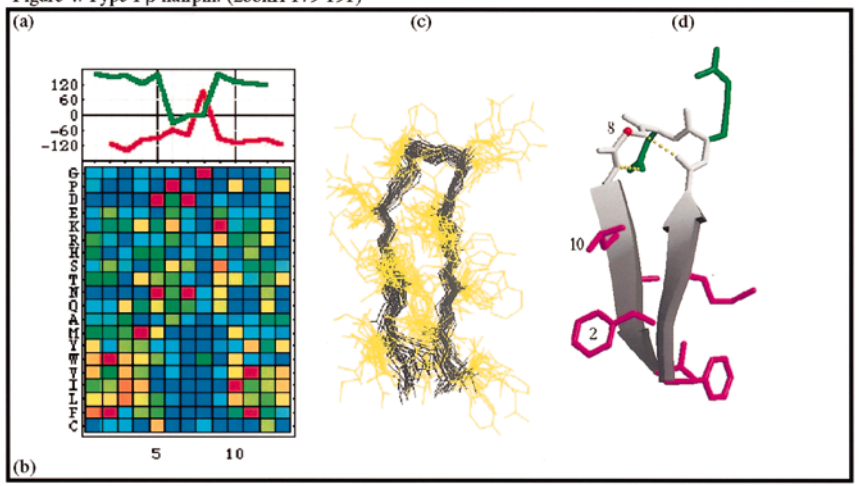


Figure 5. Frayed α-helix. (1minA 26-38)

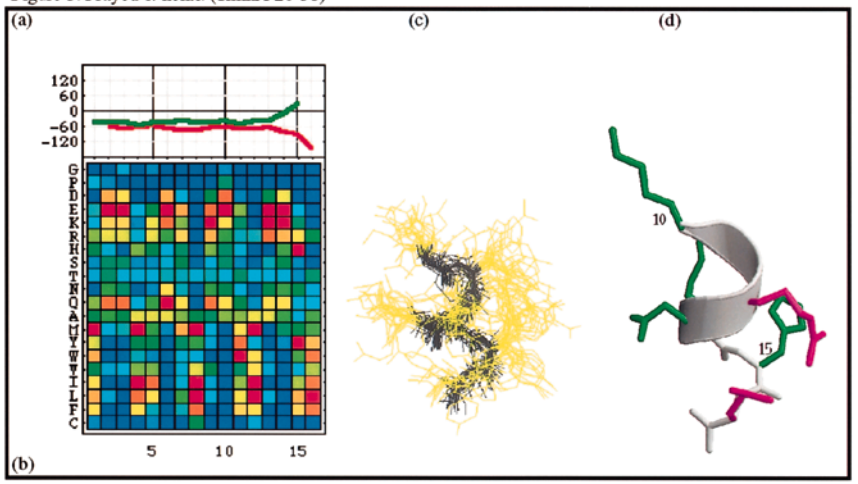
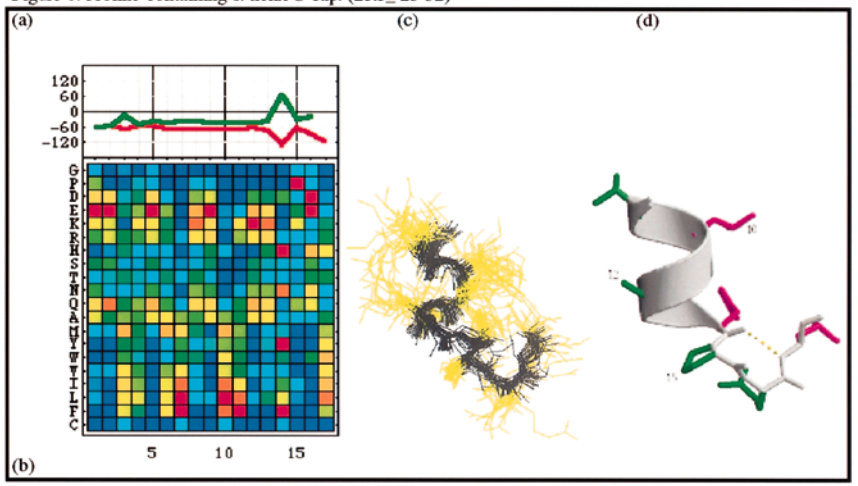
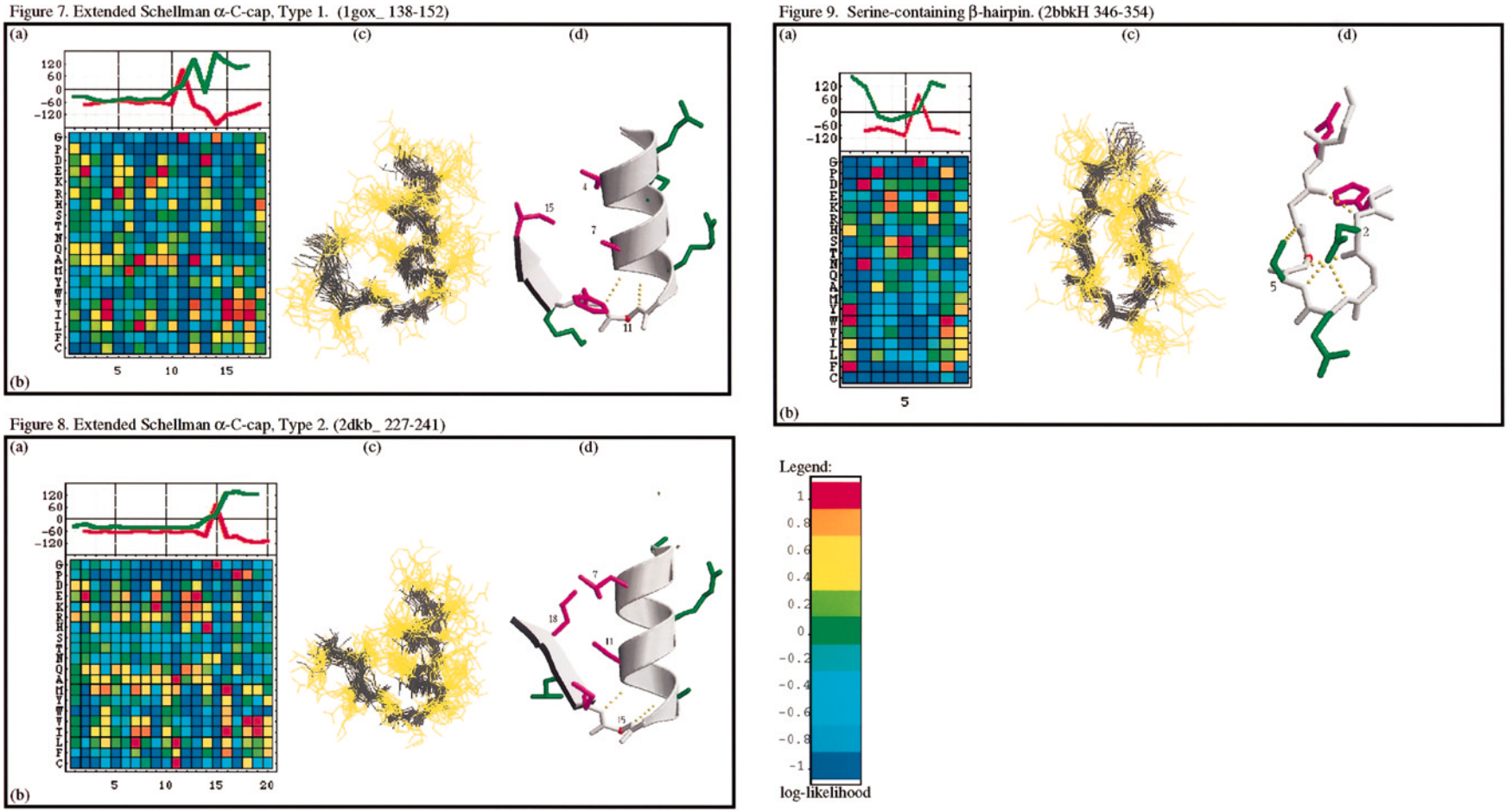


Figure 6. Proline-containing α-helix C-cap. (2ctc_ 23-32)



Figures 3 to 6 *(legend opposite)*



Figures 7 to 9.

Figures 3 to 9. Novel or extended sequence-structure motifs included in the I-sites Library are displayed in four parts in Figures 3 to 9. (a) The local structure represented as a plot of backbone dihedral angles ϕ (red) and ψ (green). (b) A color scale representation of the log-odds scoring matrix (see equation (1)); each square represents the preference for an amino acid (*y*-axis) at a position in the motif (*x*-axis). The amino acids are arranged roughly from polar to non-polar from top to bottom, except glycine and proline (at the top) and cysteine (at the bottom). (c) A superposition of 30 cluster members that fall within the cluster's natural boundaries (i.e. true-positives). (d) A cartoon representation of a representative fragment, showing the conserved polar positions in green, non-polar positions in purple and conserved glycine residues as red dots. Conserved hydrogen bonds are indicated by dotted yellow lines. Residue numbering refers to the position in parts (a) and (b). **Figure 3.** Diverging type-II β-turn. The example shown is left 247-255. **Figure 4.** Type-I β-hairpin; 2bbkH 179-191. **Figure 5.** Frayed α-helix; 1minA 26-38. **Figure 6.** Proline-containing α-helix C-cap; 2ctc 25-32. **Figure 7.** Extended Schellman α-C-cap, type 1; 1gox 138-152. **Figure 8.** Extended Schellman α-C-cap, type 2; 2dkb 227-241. **Figure 9.** Serine-containing β-hairpin; 2bbkH 346-354. The colors in Figures 3 to 9 represent log-likelihood values in natural log units according to the legend. Values above 1 and below −1 are truncated. One natural log unit equals 1.44 bits.

Proline-containing helix C-cap plus turn

This motif (5 in Table 4) is a structurally unique helix C-cap and turn in which proline terminates an amphipathic helix (residue 15 in Figure 6). The proline residue occurs three residues after a conserved polar residue and is immediately preceded by His, Asn, Tyr or Phe. A type-I β-turn forms with the proline residue in the second position. Preceding the proline residue by one helical turn are two consecutive non-polar side-chains. These fall between the helix and whatever follows the cap, which can be either helix or stand. Consecutive, conserved non-polar positions do not occur in clusters of purely helix segments. This turn is often a bridge between two helices.

Schellman α-C-cap, extended

Glycine is known to terminate helices when it occurs on the non-polar side of an amphipathic helix (Aurora *et al.*, 1994; Han *et al.*, 1997; Schellman, 1980). Here, we describe two extensions of the Schellman motif.

In the type 1 Schellman cap extension (Figure 7, motif 3 in Table 4), the glycine residue is followed by a β-bulge. An aspartate residue is preferred in the position two residues after the glycine residue. Conserved non-polar side-chains one and four residues after the glycine residue interact with two conserved non-polar side-chains four and seven residues before the glycine residue.

In the type 2 Schellman cap extension (Figure 8, motif 4 in Table 4), the capping glycine residue is often preceded by a histidine residue and may be followed by a non-polar or amphipathic β-strand. Proline instead of aspartate is preferred two positions after the glycine residue. Non-polar side-chains three and five residues after the glycine residue interact with conserved non-polar residues on the helix four and eight positions before the glycine residue, forming a chain reversal that is tighter than that in the type 1 Schellman cap extension.

Serine-containing β-hairpin

This motif (11 in Table 4) may be viewed as a single turn of α-helix capped on either end ($\beta\alpha\alpha\gamma\alpha_L\beta$-turn; Efimov, 1993), almost a merging of the helix N-capping box and the Schellman C-cap. The common unit in this family of four clusters is the central DPxTG sequence preference (residues 2 to 6 in Figure 9), which is in the helical conformation. The aspartate H-bonds with the backbone two residues after it, similar to the serine in the N-capping box. Glycine terminates the three-residue helix in a manner similar to that in the type 2 Schellman cap extension, but by H-bonding to the Ser/Thr side-chain instead of the (missing) next turn of helix. This explains the preference for serine or threonine before the glycine residue. A short peptide matching this sequence pattern has been shown to fold into this hairpin structure in solution (Blanco *et al.*, 1994).

Other motifs

The remaining patterns have been described previously in the literature. There were two classes of helical clusters: those with clear amphipathic periodicity, and those with a predominance of alanine and a lack of polar side-chains. Both classes strongly disfavor Gly, Pro and Cys in all positions. There were also two classes of strand clusters, amphipathic and hydrophobic, with alternating conserved non-polar residues and three consecutive conserved non-polar residues, respectively. Bordering these patterns were various preferences for glycine, proline and/or polar residues, often resembling turn or cap sequence patterns. Ten clusters contained variations on the N-terminal helix capping box motif, or ''hydrophobic staple''. Two clusters contained variations on the well-known β-bulge motif, having the characteristic 1-4-6 pattern of conserved hydrophobic residues, and a polar residue in the position of the kink. Full descriptions of these motifs are presented at the I-sites web site.

Natural boundaries

We observed tight structural boundaries for each of the refined sequence-structure clusters (part (c) of Figures 3 to 9). The boundaries were most evident when measured using the maximum deviation in torsion angles between two segments (*mda*, see equation (3)) combined with the distance matrix error (*dme*, see equation (2)), as suggested by the conclusions of Olivia *et al*. (1997); less sharp boundaries were observed using the RMS deviation in α-carbon atoms (*rmsd*). Most of the clusters are characterized by conserved side-chain:side-chain contacts, and changes in *mda* appear to better correlate with loss of such contacts (see Figure 2) than changes in *rmsd*. The virtual absence of intermediate values of *mda* (120 to 150°) suggests that a significant angular deviation at any point in the segment leads to disruption of the structure. In the refinement of the sequence clusters, the presence of natural boundaries strongly correlated with positive cross-validation results. The structural segments associated with most sequence clusters did not exhibit natural boundaries.

Discussion

As with any new method, it is important to compare the I-sites predictions to those of previous methods. The vast majority of methods have focused on three-state secondary structure prediction; the only method we are aware of for predicting local structure at higher resolution is that of Rooman and Wodak (Rooman *et al.*, 1991, 1992). Short peptide fragments were assigned to three-

dimensional structures provided that their energy (using an empirical energy function) was substantially lower than that of other possible structures. We have not been able to make direct comparisons of the prediction accuracy of the two methods, because the earlier method is not publicly available. Although accuracy is ultimately the most important criterion, the I-sites method has several apparent advantages: first, local structure is predicted throughout a sequence; second, each segment prediction has an associated confidence value that accurately describes the probability that the prediction is correct; and third, the method is extremely fast, since predictions require only sequence-sequence profile comparisons. Because of the very different nature of the predictions, the I-sites method is complementary to conventional secondary structure methods. While the method does not improve Q3 scores, the improvement is evident when measuring the number of eight-residue fragments correctly predicted (Table 2), either by *rmsd* of backbone atoms or by maximum deviation in backbone torsion angles (*mda*). Predictions of backbone torsion angles in a query sequence can be made through the I-sites web server.

The iterative refinement procedure described here was critical for developing the sequence patterns found using the sequence-based clustering method of Han and Baker into a method with predictive power. New sequence-structure relationships discovered or extended during this procedure include a diverging type-II β-turn, two extended versions of the Schellman motif, a specific proline-terminated helix, a ''frayed'' helix, and a hairpin containing a type-I β-turn. Because of the automated nature of the method by which the sequence patterns were identified, we believe that the majority of the strong local sequence-structure relationships that occur more than 30 times in our protein database are included in the I-sites library.

Folding initiation sites

The amino acid sequence patterns described here adopt a constant and context-independent, three-dimensional structure in folded proteins. But this does not itself prove that such sites become structured early in folding, or exist in their native states in unfolded proteins. Experimental evidence for this comes from NMR studies of isolated peptides. For the most part, peptides of 30 residues or less are found not to have a well-defined structure in water (Itzhaki *et al*., 1995; Yang *et al*., 1995), but many of the notable exceptions correspond to I-sites motifs, including the Schellman cap (Viguera & Serrano, 1995), the N-capping box (Muñoz & Serrano, 1995), the serine β-hairpin (Blanco *et al*., 1994), the type-I β-hairpin (de Alba *et al*., 1996; Ilyina *et al*., 1994; Searle *et al*., 1995), and the diverging type-II turn (Sieber & Moe, 1996). In each case, a predominant solution structure was found that closely resembled the paradigm structure of the I-sites cluster that best matched its sequence. In fact, the highest confidence I-sites predictions for a protein sequence may correspond to the segments that adopt structure independent of the rest of the protein. In the case of protein G, a peptide corresponding to the highest confidence prediction, a serine β-hairpin, was the only one of a number of peptides that were studied that was stable in isolation (Blanco & Serrano, 1995). In the SH3 domain, a peptide corresponding to the highest confidence prediction is also the only peptide to adopt structure in isolation (Viguera *et al*., 1996; Yi *et al*., 1998). The correspondence with the experimental data suggests that high confidence I-sites predictions may identify folding initiation sites in protein sequences.

Applications

Because of the increase in structural detail over conventional three-state secondary structure predictions, the I-sites method may contribute to both *ab initio* and fold recognition approaches to structure prediction. *Ab initio* folding approaches could attempt to generate tertiary structures from I-sites local structure predictions by keeping the local structure of the regions predicted at highest confidence constant and varying the local structure in low-confidence regions. The non-local interactions neglected in the I-sites method could be captured using one of many scoring/energy functions developed over the past several years. The I-sites predictions in the CASP2 prediction experiment provide an illustration of the potential power of the approach: one of the longest approximately correct *ab initio* tertiary predictions reported (T0022 215A-259A, 4.9 Å rmsd) was the result of successful prediction of three successive helical cap motifs that, when combined, folded into a roughly correct tertiary structure. We view this result as somewhat fortuitous because non-local interactions were not considered, but it augurs well for the combination of I-sites predictions and non-local scoring functions. With regard to fold recognition, I-sites predictions should contribute to sequence-structure compatibility assessment in much the same way that secondary structure predictions have recently been utilized (Fischer & Eisenberg, 1996): sequence-to-structure alignments that are consistent with the I-sites predictions may be better choices than alignments that are inconsistent with the I-sites predictions. Other applications include gene finding and sequence comparison; promising results have already been obtained in the former area (unpublished results).

Bona fide blind predictions of the CASP2 targets made using a preliminary version of the I-sites library with shorter profiles are described elsewhere (Bystroff & Baker, 1997). Predictions of the four as yet (April 1997) unsolved CASP2 structures using the library described here have been submitted to the Livermore prediction center (http://PredictionCenter.llnl.gov/). We believe that blind

tests are critical for assessing new methods and look forward to making predictions in CASP3. More information about the I-sites library is available *via* the World Wide Web site: http://ganesh.bchem.washington.edu/ ~ bystroff/Isites/. Linked to the site is a server that will predict the backbone angles of multiple-aligned or single sequences.

Methods

The sequence and structure database

The database for this work consisted of 471 protein sequence families from the HSSP database (Sander & Schneider, 1994; Schneider & Sander, 1996), with an average of 47 aligned sequences per family. Each family contains a single known structure (parent) from the Brookhaven protein Data Bank (Bernstein *et al*., 1977). These are a subset of the PDBSelect-25 list (Hobohm *et al*., 1992; November 1996 release), having no more than 25% sequence identity between any two alignments. Families in the PDBSelect-25 list were excluded if the parent structure was not well determined, if the protein was membrane bound, or if it contained a large number of disulfide bonds. Disordered loops were omitted. Gaps and insertions in the sequence were ignored.

Clustering of sequence segments

Each position in the database was described by a weighted (Vingron & Argos, 1989) amino acid frequency profile (Gribskov *et al*., 1990), *P*. A similarity measure in sequence space between a segment (p) and a cluster of segments (q) was defined as:

$$D_{pq} = \sum_{ij} \log\left[\frac{P_{ij}(p) + \alpha F_i}{(1+\alpha)F_i}\right] \log\left[\frac{\sum_{k\in q} P_{ij}(k) + \alpha' F_i}{(N_q + \alpha')F_i}\right] \quad (1)$$

where $P_{ij}(p)$ is the frequency of amino acid i in position j within the segment p. N_q is the number of sequence segments k in the cluster q. F_i is the frequency of amino acid type i in the database overall. The optimal values of α and α' were determined empirically to be 0.5 and 15, respectively. Using this similarity measure, segments of a given length (3 to 15) were clustered *via* the 'Kmeans' algorithm (Everitt, 1993).

Assessing structure within a cluster; choice of paradigm

The structural similarity between any two peptide segments was evaluated using a combination of the RMS distance matrix error (*dme*):

$$dme = \sqrt{\frac{\sum_{i=1}^{L}\sum_{j=i-5}^{i+5}(\alpha_{i\to j}^{s1} - \alpha_{i\to j}^{s2})^2}{N}} \quad (2)$$

where $\alpha_{i\to j}$ is the distance between α-carbon atoms i and j in the segment *s1* of length L, and the maximum deviation in backbone torsion angles (*mda*) over the length of the segment is given by:

$$mda(L) = \max_{i=1,L-1}(\Delta\Phi_{i+1}, \Delta\Psi_i) \quad (3)$$

The paradigm structure for a cluster was chosen from the top-scoring 20 segments in the database as that with the smallest sum of *mda* values to the other 19.

Other structural measures were tried before settling on these two: RMS deviation of α-carbon atoms (*rmsd*), *dme* alone, and a structural filter that looked for specific conserved contacts. The latter worked best in discriminating true and false positives, but could not be easily automated. The *rmsd* and *dme* were found to be poor discriminators of the two types of helix cap. The *mda-dme* combined filter best simulates the conserved contacts filter and is rapidly computed (Figure 2).

True/false boundaries in structure space

The refinement procedure described below required that all segments could be assigned a true or false value based on the structural difference with the paradigm. The observation of natural boundaries (see Results) in structure space, as we have defined it above, facilitated the choice of cutoff values (boundaries). Histograms of *dme* and *mda versus* the paradigm were summed for all segments in the cluster. These histograms are generally bimodal when a true sequence-structure correlation exists. The boundary was set to where the histogram first dropped to half its maximum value. If the histogram did not have the bimodal shape, or the drop was reached after 130° in *mda* or 1.3 Å in *dme*, then the cluster was rejected. The boundary values for each structural motif, averaged over all clusters in that motif, are shown in Table 2. The average boundaries for all 82 clusters were 81° in *mda* and 0.89 Å *dme*.

Iterative refinement of clusters

For each of the clusters that was found to have good structural boundaries, an iterative procedure was used to increase the correlation between segments selected based on sequence and those selected based on structure. The word profile as used below refers to the amino acid frequency profile of all positions in the segment plus two residues on either end; i.e. if the cluster segments were seven residues long, a profile of length 11 was calculated, centered on the seven.

Algorithm 1: (1) all member segments that were not within the natural boundaries of the paradigm structure are removed. (2) The frequency profile of the cluster is calculated from the remaining members. (3) Using the new profile, the database is searched for the 400 highest-scoring (equation (1)) segments, which becomes the new cluster. These steps were repeated to convergence (3 to 5 cycles).

Cross-validation and confidence

To show that the procedure was improving the predictive value of the cluster profile, a jack-knife test was performed: 90% of the database was used in the refinement procedure above, while the remaining 10% was set aside and used for validation. Validation consists of assigning a true or false to each high-scoring segment in the validation set based on the paradigm and boundaries. The jack-knife test was repeated ten times, each time using a different 10% of the database and choosing a new paradigm. If the ten paradigms were not structurally the same (within natural boundaries) or if the ten runs did not converge to the same profile, then the cluster was rejected. If the cluster was not rejected, the per-

centage true was determined as a function of the D_{pq} score (equation (1)) in bins of 20, resulting in the ''confidence curve'' (Figure 1). Scores are translated to confidences using these curves, after smoothing by linear interpolation.

Iterative peak removal

In some cases, similar sequence patterns mapped to different structures. When this happened, the predominant pattern occluded the secondary one. To find structurally distinct clusters with similar sequence patterns, the cluster refinements were repeated using subsets of the data in which the members of previously identified clusters were removed. This was important for identifying the two distinct Schellman α-C-cap extensions, which are very similar in sequence. At the end of this procedure, clusters were rejected from the library if they did not have at least 70% confidence in the highest bin.

Cluster weights

The prediction accuracy was improved by requiring that the number of predictions of each paradigm structure match the number of occurrences of that structure in the database. This was done by defining a weight (w) for the confidence curve of each cluster (set initially to 1), and then minimizing the difference between false-positives (F^+) and false-negatives (F^+) in the database, using a gradient descent approach. The update equation for the cluster weights was:

$$w_C^{\text{new}} = w_C^{\text{old}} + \varepsilon\left(\frac{F_C^- - F_C^+}{F_C^- + F_C^+}\right) \quad (4)$$

where ε is a small positive value. Using optimized cluster weights improved the performance of the library in a jack-knife test; when cluster weights were generated using one-half of the database, the total number of true-positives increased significantly in the other half, from 68 to 74% of the predicted positions.

Prediction protocol

To make a local structure prediction starting from a single sequence, the following was done. Algorithm 2: (1) the sequence was submitted to the PHD Predict Protein server (Rost *et al*., 1994) to obtain a set of multiple-aligned sequences and hence a profile. (2) Each segment of the profile was scored against each of the 82 clusters, and the scores were converted to weighted confidences. (3) All predicted segments were sorted from high to low based on weighted confidence. (4) The first segment was assigned the φ and ψ angles of the cluster's paradigm. (5) For all subsequent segments in the sorted list, the prediction was used if none of its φ-ψ values conflicted with any previously assigned φ-ψ values, within a 60° limit.

Independent test set

A recent release of the PDB-select database (October 1997) contained many new structures not included in the training data set. To form an independent test set, 55 sequence families were identified that did not contain any of the sequences used in training, including parent sequences and all homologs used to construct the profiles. Like the training set, the test set is non-redundant, with less than 25% identity between members of any two sequence families. Six members of the test set were all β-strand proteins: 1lcl, 1mspA, 1rie, 1stmA, 2ayh, 2stv (four-letter PDB code + chain identifier, if present). Eight members were all α-helix proteins: 1bmfG, 1cem, 1cpo, 1ignA, 1kxu, 1lbd, 1vnc, 1xsm. The rest (41) were αβ, α + β or multi-domain proteins having both types of secondary structure: 1alo, 1anv, 1apyA, 1ayl, 1bmfA, 1bmfD, 1broA, 1dekA, 1div, 1fieA, 1frvA, 1frvB, 1 gal, 1gnd, 1gplA, 1gtmA, 1havA, 1htp, 1httA, 1hxpA, 1ihfB, 1lbu, 1lnh, 1otgA, 1oxy, 1qba, 1reqA, 1sfe, 1taq, 1tfe, 1tfr, 1vcc, 1vhiA, 1whi, 1xel, 1xvaA, 1zymA, 2ebn, 2eng, 4kbpA.

Acknowledgements

We thank Ed Thayer, Kevin Karplus, Daniel Fischer, Bob McCammon, Dietlund Gerloff, David Shortle and members of the Baker laboratory for helpful discussions. This work was partially supported by National Science Foundation, Science and Technology Cooperative Center Agreement BIR-9214821, and young investigator awards to D.B. from the National Science Foundation and the Packard Foundation.

References

Aurora, R., Srinivasan, R. & Rose, G. D. (1994). Rules for alpha-helix termination by glycine. *Science*, **264**, 1126–1130.

Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Jr, Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T. & Tasumi, M. (1977). The Protein Data Bank. A computer-based archival file for macromolecular structures. *Eur. J. Biochem.* **80**, 319–324.

Blanco, F. J. & Serrano, L. (1995). Folding of protein G B1 domain studied by the conformational characterization of fragments comprising its secondary structure elements. *Eur. J. Biochem.* **230**, 34–649.

Blanco, F. J., Rivas, G. & Serrano, L. (1994). A short linear peptide that folds into a native stable beta-hairpin in aqueous solution. *Nature Struct. Biol.* **1**, 584–590.

Bystroff, C. & Baker, D. (1997). Blind ab initio local structure predictions using a library of sequence-structure motifs. *Proteins: Struct. Funct. Genet. Suppl.* **1**, 167–171.

Chain, A. W., Hutchinson, E. G., Harris, D. & Thornton, J. M. (1993). Identification, classification, and analysis of beta-bulges in proteins. *Protein Sci.* **2**, 1574–1590.

de Alba, E., Jiménez, M. A., Rico, M. & Nieto, J. L. (1996). Conformational investigation of designed short linear peptides able to fold into β-hairpin structures in aqueous solution. *Folding Des.* **1**, 133–144.

Donnelly, D., Overington, J. P. & Blundell, T. L. (1994). The prediction and orientation of alpha-helices from sequence alignments: the combined use of environment-dependent substitution tables, Fourier transform methods and helix capping rules. *Protein Eng.* **7**, 645–653.

Duda, R. O. & Hart, P. E. (1973). *Pattern Classification and Scene Analysis*, John Wiley & Sons, New York.

Efimov, A. V. (1993). Standard structures in proteins. *Prog. Biophys. Mol. Biol.* **60**, 201–239.

elMasry, N. F. & Fersht, A. R. (1994). Mutational analysis of the N-capping box of the alpha-helix of chymotrypsin inhibitor 2. *Protein Eng.* **7**, 777–782.

Everitt, B. (1993). *Cluster Analysis*, Halsted Press, New York.

Fischer, D. & Eisenberger, D. (1996). Protein fold recognition using sequence-derived predictions. *Protein Sci.* **5**, 947–955.

Gribskov, M., Luthy, R. & Eisenberg, D. (1990). Profile analysis. *Methods Enzymol.* **183**, 146–259.

Han, K. F. & Baker, D. (1995). Recurring local sequence motifs in proteins. *J. Mol. Biol.* **251**, 176–187.

Han, K. F. & Baker, D. (1996). Global properties of the mapping between local amino acid sequence and local structure in proteins. *Proc. Natl Acad. Sci. USA*, **93**, 5814–5818.

Han, K. F., Bystroff, C. & Baker, D. (1997). Three dimensional structures and contexts associated with recurrent amino acid sequence patterns. *Protein Sci.* **6**, 1587–1590.

Hobohm, U., Scharf, M., Schneider, R. & Sander, C. (1992). Selection of representative protein data sets. *Protein Sci.* **1**, 409–417.

Hutchinson, E. G. & Thornton, J. M. (1994). A revised set of potentials for beta-turn formation in proteins. *Protein Sci.* **3**, 2207–2216.

Ilyina, E., Milius, R. & Mayo, K. H. (1994). Synthetic peptides probe folding initiation sites in platelet factor-4: stable chain reversal found within the hydrophobic sequence LIATLKNGRKISL. *Biochemistry*, **33**, 13436–13444.

Itzhaki, L. S., Neira, J. L., Ruiz, Sanz J., de Prat, Gay G. & Fersht, A. R. (1995). Search for nucleation sites in smaller fragments of chymotrypsin inhibitor 2. *J. Mol. Biol.* **254**, 289–304.

Jiménez, M. A., Muñoz, V., Rico, M. & Serrano, L. (1994). Helix stop and start signals in peptides and proteins. The capping box does not necessarily prevent helix elongation. *J. Mol. Biol.* **242**, 487–496.

Kabsch, W. & Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, **22**, 2577–2637.

Lesk, A. M. (1997). CASP2: report on *ab initio* predictions. *Proteins: Struct. Funct. Genet.* Suppl. 1, 151–166.

Muñoz, V. & Serrano, L. (1995). Analysis of i,i + 5 and i,i + 8 i-hydrophobic interactions in a helical model peptide bearing the hydrophobic staple motif. *Biochemistry*, **34**, 15301–15306.

Muñoz, V., Blanco, F. J. & Serrano, L. (1995). The hydrophobic-staple motif and a role for loop-residues in alpha-helix stability and protein folding. *Nature Struct. Biol.* **2**, 380–385.

Oliva, B., Bates, P. A., Querol, E., Avilés, F. X. & Sternberg, M. J. E. (1997). An automated classification of the structure of protein loops. *J. Mol. Biol.* **266**, 814–830.

Rooman, M. J., Rodriguez, J. & Wodak, S. J. (1990). Automatic definition of recurrent local structure motifs in proteins. *J. Mol. Biol.* **213**, 327–336.

Rooman, M. J., Kocher, J. P. & Wodak, S. J. (1991). Prediction of protein backbone conformation based on seven structure assignments. Influence of local interactions. *J. Mol. Biol.* **221**, 961–979.

Rooman, M. J., Kocher, J. P. & Wodak, S. J. (1992). Extracting information on folding from the amino acid sequence: accurate predictions for protein regions with preferred conformation in the absence of tertiary interactions. *Biochemistry*, **31**, 10226–10238.

Rost, B. & Sander, C. (1993a). Improved prediction of protein secondary structure by use of sequence profiles and neural networks. *Proc. Natl Acad. Sci. USA*, **90**, 7558–7562.

Rost, B. & Sander, C. (1993b). Prediction of protein secondary structure at better than 70% accuracy. *J. Mol. Biol.* **232**, 584–599.

Rost, B. & Sander, C. (1994). Combining evolutionary information and neural networks to predict protein secondary structure. *Proteins: Struct. Funct. Genet.* **19**, 55–72.

Rost, B., Sander, C. & Schneider, R. (1994). PHD: an automatic mail server for protein secondary structure prediction. *Comput. Appl. Biosci.* **10**, 53–60.

Sander, C. & Schneider, R. (1994). The HSSP database of protein structure-sequence alignments. *Nucl. Acids Res.* **22**, 3597–3599.

Schellman, C. (1980). The aL conformation at the ends of helices. In *Protein Folding: Proceedings of the 28th Conference of the German Biochemical Society, University of Regensburg, Regensburg, West Germany, September 10–12, 1979* (Jaenicke, R., ed.), pp. 53–61, Elsevier/North-Holland Biomedical Press, Amsterdam.

Schneider, R. & Sander, C. (1996). The HSSP database of protein structure sequence alignments. *Nucl. Acids Res.* **24**, 201–205.

Searle, M. S., Williams, D. H. & Packman, L. C. (1995). A short linear peptide derived from the N-terminal sequence of ubiquitin folds into a water-stable non-native beta-hairpin. *Nature Struct. Biol.* **2**, 999–1006.

Sieber, V. & Moe, G. R. (1996). Interactions contributing to the formation of a beta-hairpin-like structure in a small peptide. *Biochemistry*, **35**, 181–188.

Unger, R. & Sussman, J. L. (1993). The importance of short structural motifs in protein structure analysis. *J. Comput. Aided Mol. Des.* **7**, 457–472.

Unger, R., Harel, D., Wherland, S. & Sussman, J. L. (1989). A 3D building blocks approach to analyzing and predicting structure of proteins. *Proteins: Struct. Funct. Genet.* **5**, 355–373.

Viguera, A. R. & Serrano, L. (1995). Experimental analysis of the Schellman motif. *J. Mol. Biol.* **251**, 150–160.

Viguera, A. R., Jiménez a., M., Rico, M. & Serrano, L. (1996). Conformational analysis of peptides corresponding to beta-hairpins and a beta-sheet that represent the entire sequence of the alpha-spectrin SH3 domain. *J. Mol. Biol.* **255**, 507–521.

Vingron, M. & Argos, P. (1989). A fast and sensitive multiple sequence alignment algorithm. *Comput. Appl. Biol. Sci.* **5**, 115–121.

West, M. W. & Hecht, M. H. (1995). Binary patterning of polar and nonpolar amino acids in the sequences and structures of native proteins. *Protein Sci.* **4**, 2032–2039.

Yang, A. S., Hitz, B. & Honig, B. (1996). Free energy determinants of secondary structure formation. III. Beta-turns and their role in protein folding. *J. Mol. Biol.* **259**, 873–882.

Yang, J. J., Buck, M., Pitkeathly, M., Kotik, M., Haynie, D. T., Dobson, C. M. & Radford, S. E. (1995). Conformational properties of four peptides spanning the sequence of hen lysozyme. *J. Mol. Biol.* **252**, 483–491.

Yi, Q., Bystroff, C., Rajagopal, P., Klevit, R. E. & Baker, D. (1998). Prediction and structural charcterization of an independently folding substructure in the src SH3 domain. *J. Mol. Biol.* In the press.

Zhu, Z. Y. & Blundell, T. L. (1996). The use of amino acid patterns of classified helices and strands in secondary structure prediction. *J. Mol. Biol.* **260**, 261–276.

Edited by J. Thornton

(Received 19 January 1998; received in revised form 16 April 1998; accepted 30 April 1998)