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

Represented by Jianwei

Abstract

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.

Introduction

The challenge for local structure prediction is to identify the structural features that have strong sequence preferences.

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.

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.

Results

- Local protein structure prediction using I-site library
  - Prediction
  - Evaluation
  - Training set
  - Test set
  - Combination of I-sites and conventional sse prediction

- Overview of the I-site library
  - 13 different motifs and 7 listed in the paprer
  - Natural boundaries

Prediction

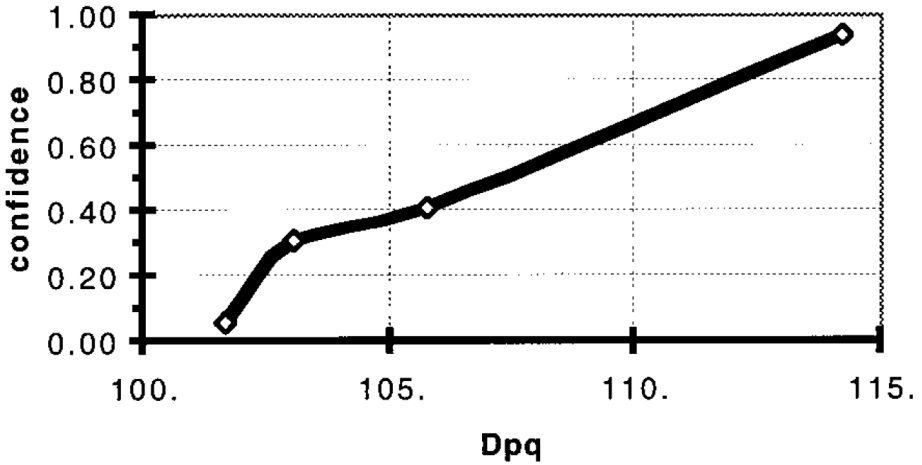


Figure 1. A confidence curve maps similarity score to the probability of correct local structure based on a ten-fold jack-knife test. All nine-residue segments in a test set composed of 10% of the database were scored using the profile for a nine-residue ''serine-containing β-hair-pin'' cluster, which 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.

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:

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

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).

Evaluation

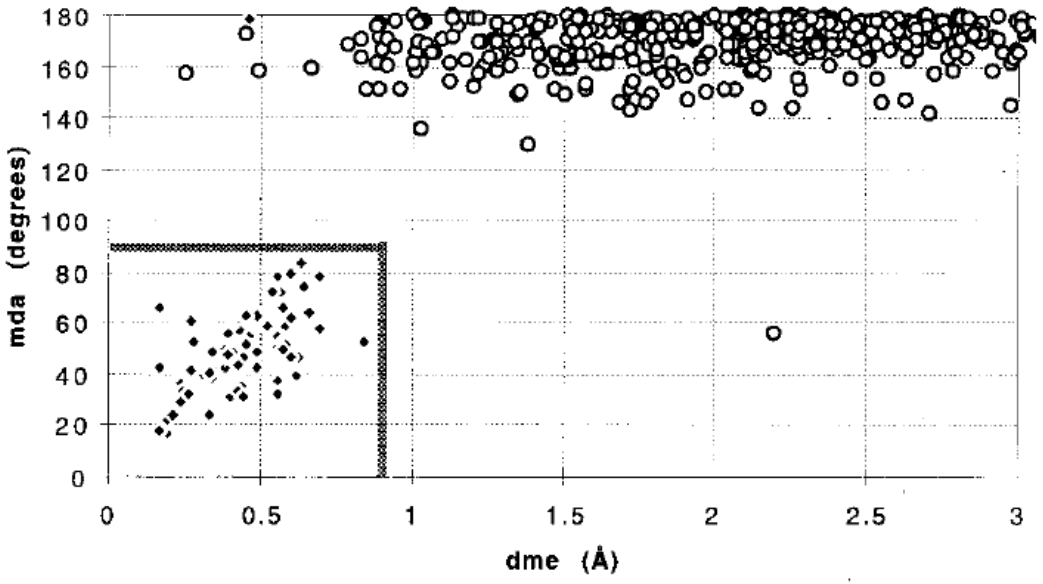


Figure 2. Deviations from the paradigm structure in *dme* and *mda* for the top 600 scores in the database for one cluster (a nine-residue serine β-hairpin). A clear separation (natural boundary) appears between the segments that conserve four specific side-chain:side-chain contacts and two specific 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.

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*):

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

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:

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

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).

Training set

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.

Test set & Combination

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.

13 motifs & Boundaries

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*).

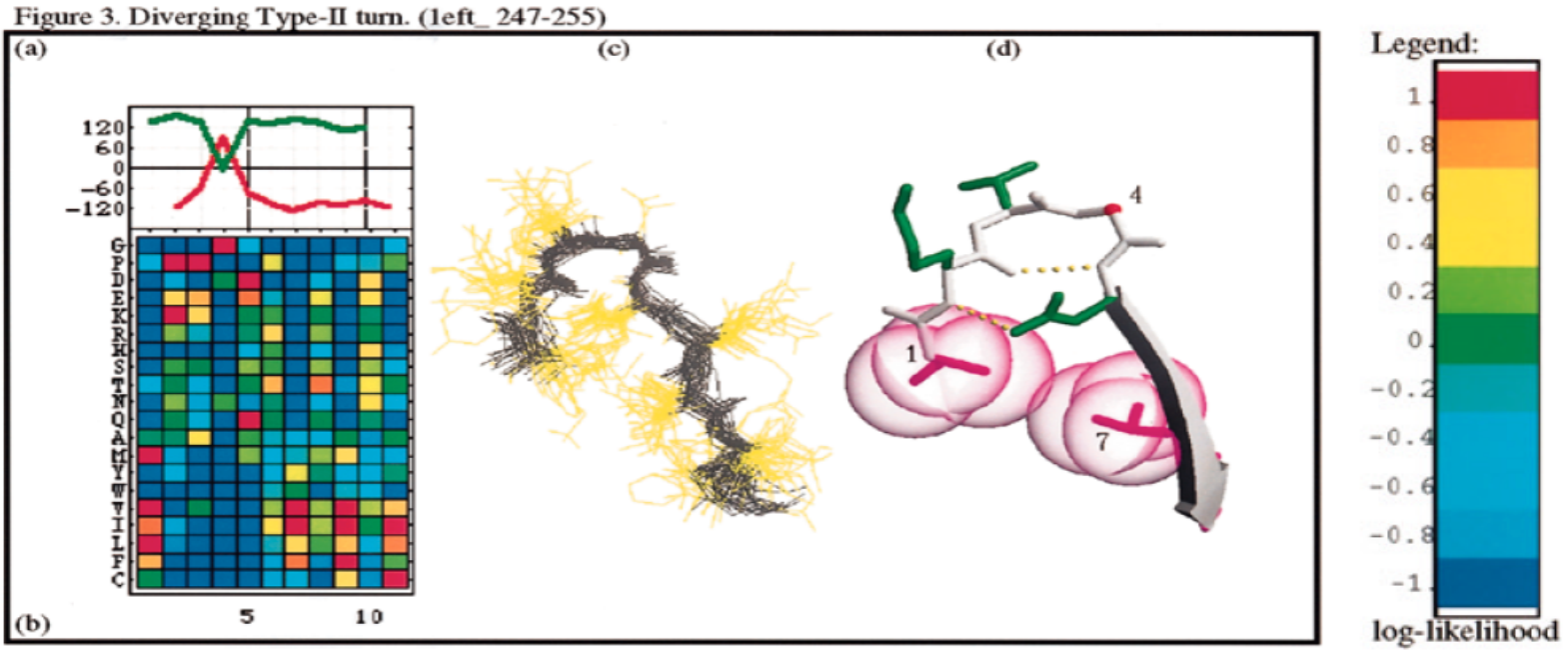
13 motifs & Boundaries

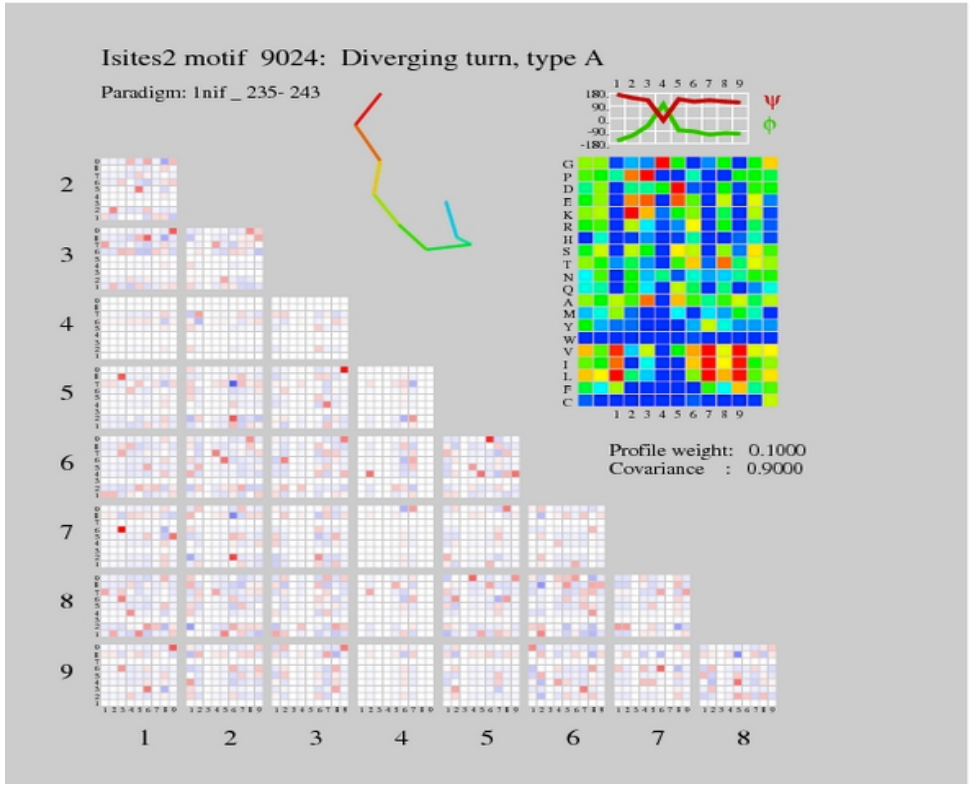
Table 4. A summary of the sequence-structure motifs in the I-sites library

	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).

13 motifs & Boundaries





Novel or extended sequence-structure motifs included in the I-sites Library are displayed in four parts.

(a) The local structure represented as a plot of backbone dihedral angles f (red) and c (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.

The colors 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.

Discussion

Folding initiation sites

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,

Applications

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.

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 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).

Methods

- The sequence and structure database
- Clustering of sequence segments
- Assessing structure within a cluster; choice of paradigm
- True/false boundaries in structure space
- Iterative refinement of clusters
- Cross-validation and confidence
- Iterative peak removal
- Cluster weights
- Prediction protocol
- Independent test set

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 percentage 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:

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

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.