Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 5814–5818, June 1996
Biophysics

Global properties of the mapping between local amino acid sequence and local structure in proteins

KAREN F. HAN*† AND DAVID BAKER‡§

*Graduate Group in Biophysics, University of California, San Francisco, CA 94143; and ‡Department of Biochemistry, University of Washington, Seattle, WA 98195

Communicated by Leroy Hood, University of Washington School of Medicine, Seattle, WA, February 7, 1996 (received for review October 7, 1995)

ABSTRACT **Local protein structure prediction efforts have consistently failed to exceed ~70% accuracy. We characterize the degeneracy of the mapping from local sequence to local structure responsible for this failure by investigating the extent to which similar sequence segments found in different proteins adopt similar three-dimensional structures. Sequence segments 3–15 residues in length from 154 different protein families are partitioned into neighborhoods containing segments with similar sequences using cluster analysis. The consistency of the sequence-to-structure mapping is assessed by comparing the local structures adopted by sequence segments in the same neighborhood in proteins of known structure. In the 154 families, 45% and 28% of the positions occur in neighborhoods in which one and two local structures predominate, respectively. The sequence patterns that characterize the neighborhoods in the first class probably include virtually all of the short sequence motifs in proteins that consistently occur in a particular local structure. These patterns, many of which occur in transitions between secondary structural elements, are an interesting combination of previously studied and novel motifs. The identification of sequence patterns that consistently occur in one or a small number of local structures in proteins should contribute to the prediction of protein structure from sequence.**

Most studies of local sequence–structure relationships have involved the tabulation of statistics on sequences that occur in structural motifs of interest (1–4) (Fig. 1*A*). Our approach (Fig. 1*B*) is essentially the inverse. Instead of investigating the sequence patterns found in predefined local structural environments, we first identify recurring sequence patterns and then investigate their structural correlates.

It is well established that the local sequence-to-structure mapping is not one to one over all of sequence space: identical pentapeptide sequences exist in completely different tertiary structures in proteins (5). Furthermore, local structure prediction efforts consistently fail to exceed ~70% accuracy (6), suggesting that the mapping from local sequence to structure is likely to be degenerate for a significant fraction of sequence space. In this paper we characterize the degeneracy of the mapping by determining the number of sequence segments in neighborhoods (regions of sequence space) in which the sequence-to-structure mapping is one to one, one to two, and one to three (Fig. 1*B*).

The definition of local sequence neighborhoods requires a measure of distance between short sequence segments. Most sequence comparison methods rely on a single global substitution matrix compiled by averaging over all positions in a large set of aligned protein sequences (7). However, at different positions in proteins, different amino acid residues are likely to substitute for each other, and thus the use of a global substitution matrix is potentially problematic. These problems can be circumvented if the two segments being compared are both derived from protein families with multiple members: sequence profiles (8) constructed from sets of aligned sequences contain position-specific information on amino acid substitution patterns.

In previous work (9), we utilized a measure of the distance between sequence profiles generated from multiple sequence alignments to identify sequence patterns that transcend protein family boundaries. A similar distance measure is used in this paper, and the term "segment" below refers to a segment of a profile generated from a multiple sequence alignment. The earlier work focused on the identification and characterization of recurring sequence patterns; the focus of the current paper is on the structural correlates of these patterns.

Methods

The clustering procedures have been described in detail in ref. 9. In brief, 29,921 segments of profiles derived from a nonredundant subset [PDB-select 25 (10)] of the HSSP database (11) of multiple sequence alignments were subdivided into 1200 neighborhoods containing sets of related segments using the *K* means algorithm (12) and the city block metric

$$d(i,j) = \sum_{n=1}^{N} \sum_{k=1}^{20} |F_i(k,n) - F_j(k,n)|$$

where $F_i(k,n)$ (a profile segment) is the frequency of the kth amino acid in the nth position of segment i and N is the segment length. Because the PDB-select 25 subset contains very few pairs of alignments from even distantly related families, segments in a given neighborhood are necessarily derived from quite different protein families. To capture patterns of different lengths, the procedure was repeated for segment lengths ranging from 3 to 15 residues. Frequently, segments of length 9 to 15 that belonged to neighborhoods with strong sequence-to-structure correlations contained shorter segments, which also belonged to such neighborhoods. To avoid overcounting, the statistics in the tables for a given segment length exclude positions already included in the statistics for a longer segment length.

Secondary structure and solvent accessibility data for each of the segments in each of the neighborhoods were extracted from the HSSP data base using previously described simplifications (6). The average consistency of secondary structure within a neighborhood was evaluated using the simple formula:

$$\frac{\sum_{i=1}^{N} \max(p_{i,helix}, p_{i,strand}, p_{i,turn})}{N}$$

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

†Present address: Northwestern University Medical School, Box 182, Chicago, IL 60611.
§To whom reprint requests should be addressed.

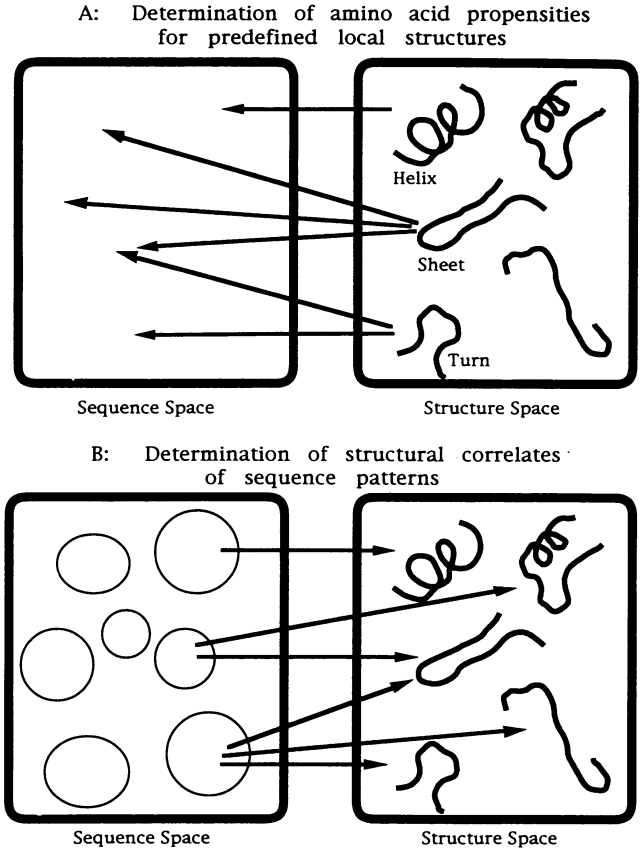

FIG. 1. Approaches to studying local sequence–structure relationships.

where the p_i are the frequencies of occurrence of the indicated secondary structure among the segments in the neighborhood at position i and N is the segment length. For N greater than seven, the lowest scoring position was excluded from the average to allow for ambiguities in secondary structure assignments in transition elements.

To test the statistical significance of the results with the HSSP data set, simulated sequences were generated with the average occurrences and variances of each of the amino acid residues in the HSSP data set, but not the interresidue correlations (9). To preserve the nonrandom sequential correlations in secondary structure elements, the secondary structure assignments were not shuffled in the simulated set.

The secondary structure consistency within the neighborhoods generated from the HSSP data set frequently exceeded 80%, but it almost never reached 80% in the simulated data set. A consistency threshold of 80% is used throughout the paper: for example, the sequence-to-structure mapping was considered to be one to one if the agreement in secondary structure among the segments within a neighborhood averaged 80% or greater over the length of the segments.

Tables 3 and 4 include only a subset of the patterns; unabridged versions are available from the authors by electronic mail.

RESULTS

Sequence segments ranging from 3 to 15 residues in length from a nonredundant subset of the HSSP database of multiple sequence alignments were partitioned into neighborhoods using the K-means algorithm. Because the HSSP database includes at least one sequence of known three-dimensional structure per multiple sequence alignment, the structure adopted by each of the segments in each neighborhood is known with reasonable certainty (12).

Table 1. Overall distribution of sequence patterns for which a single local structure predominates

Length	No. of positions	H	S	T	HT	TH	TS
15	300	0	0	0	0	300	0
13	2,847	691	0	0	1,393	763	0
11	3,399	1,973	0	0	840	586	0
9	2,609	1,376	0	0	433	411	359
7	1,711	819	0	559	0	71	262
5	1,327	208	231	888	0	0	0
3	958	0	103	855	0	0	0
Total	13,151	5,067	334	2,302	2,666	2,131	621

The total number of positions in neighborhoods in which the consistency of the sequence-to-structure mapping was greater than 80% (column 2) and their distribution among different local structures (H, S, T: helix, sheet, or turn throughout the segment; HT, TH, TS: helix–turn, turn–helix, and turn–sheet transitions) is given for different segment lengths. The choice of local structure groupings is primarily for convenience of presentation; other choices would include the 3D building blocks of Unger and Sussman (13).

Approximately 44% of the positions in the input set of multiple sequence alignments fell into a neighborhood in which a single local structure predominated (Table 1). For segment lengths 13 and 15, these predominant local structures are primarily helix caps; for segment lengths 7 to 11, helices; and for segment lengths 3 and 5, turns and loops. Although considerably less frequent than the patterns found in helices and turns, a number of patterns were found in turn-to-sheet transitions for segment lengths 7 and 9, and in β-strands for segment lengths 3 and 5.

To determine the number of distinct structural elements in the neighborhoods in which the sequence-to-structure mapping was not one to one, the K-means algorithm was used to

Table 2. Distribution of sequence segments among neighborhoods in which the sequence structure mapping is one to one, one to two, and one to three

	Positions (%)	H	S	T	HT	TH	ST	TS	HTS	TST
HSSP 1-1	43.9	17.0	1.1	7.7	8.9	7.1	0.0	2.1	0.0	0.0
HSSP 1-2	27.7	11.0	1.1	10.0	1.9	1.3	0.7	0.9	0.8	0.0
HSSP 1-3	8.3	5.0	0.0	2.0	0.2	0.7	0.0	0.0	0.0	0.4
Total	79.9	33.0	2.2	19.7	11.0	9.1	0.7	3.0	0.8	0.4
SIM 1-1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
SIM 1-2	2.0	0.4	0.0	1.6	0.0	0.0	0.0	0.0	0.0	0.0
SIM 1-3	0.5	0.5	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Total	2.5	0.9	0.0	1.6	0.0	0.0	0.0	0.0	0.0	0.0

To identify neighborhoods that contained two or three different local structures, the segments within a neighborhood were subdivided into two or three groups using the K-means algorithm (12) and the distance measure

$$d_s(i,j) = \sum_{n=1,N} \sum_{k=helix,strand,turn} |S(n,i,k) - S(n,j,k)|$$

where $S(n, i, k)$ is the frequency of occurrence of secondary structure type k at position n in segment i, and N is the window length. Column 2 lists the percentage of positions in neighborhoods in which the overall secondary structure consistency within 1 (row 1), 2 (row 2), or 3 (row 3) subgroups was greater than 80%. Comparison with the simulated data set showed that for segment lengths of less than nine, the one to three mapping had little statistical significance, and thus only positions in segments of at least nine residues are included in the 1–3 mapping statistics. Statistics on positions in neighborhoods with one to two and one to three sequence to structure mappings exclude positions falling into neighborhoods with one to one and one to two mappings, respectively. Rows 5–8 give the results of applying the same procedures to a simulated data set.

Table 3. Selected sequence patterns that occur predominantly in a single type of local structure

Patterns	Hφ	SA	H	S	T
Amphipathic helices					
#1 25					
[GPa]	.41	.24	23	0	2
[Alr]	.38	.76	24	0	1
[Ark]	.38	.72	23	0	2
[aVI]	.83	.12	24	0	1
[Av]	.81	.28	25	0	0
π	.19	.84	25	0	0
.	.35	.56	24	0	1
[vLY]	.88	.24	25	0	0
.	.44	.72	21	0	4
#2 23					
.	.39	.65	19	0	4
[Anq]	.19	.96	22	0	1
π	.26	.70	22	0	1
[gAv]	.75	.09	21	0	2
[RKd]	.24	.74	22	0	1
π	.24	.91	22	0	1
[ViL]	.86	.26	22	0	1
[Ay]	.75	.13	21	0	2
[gAq]	.19	.91	20	0	3
Amphipathic strands					
#3 58					
π	.22	.82	6	9	43
[G]	.05	.75	4	5	49
π	.12	.79	3	10	45
π	.16	.70	4	33	21
[Vil]	.66	.15	4	46	8
π	.35	.50	3	50	5
[VIL]	.74	.10	5	50	3
#4 39					
π	.19	.64	0	28	11
[VI]	.65	.15	0	34	5
[TsKE]	.11	.66	0	38	1
[Vi]	.71	.10	1	36	2
π	.24	.69	2	23	14
Less amphipathic helix					
#5 28					
.	.52	.50	23	1	4
[gAs]	.74	.18	23	1	4
[vLF]	.89	.18	24	1	3
.	.54	.46	24	1	3
[AvlT]	.59	.29	25	1	2
[AS]	.76	.21	24	1	3
[aVIL]	.70	.36	25	1	2
[Alsk]	.53	.57	22	0	6
[gAvL]	.61	.43	16	0	12
Turns/coils with conserved glycines and prolines					
#6 24					
[P]	.07	.61	4	6	21
[AVTR]	.33	.48	4	4	23
π	.17	.77	4	3	24
π	.24	.80	2	2	27
[G]	.02	.64	2	0	29
#7 33					
[G]	.04	.48	2	4	27
.	.38	.73	2	3	28
π	.22	.58	1	2	30
.	.55	.42	3	2	28
[P]	.06	.49	7	2	24
#8 27					
[P]	.05	.70	0	2	25
.	.42	.81	0	0	27
π	.35	.67	0	2	25
[P]	.10	.78	3	2	22
π	.31	.63	6	3	18
#9 24					
.	.37	.54	0	2	22
[Pf]	.23	.50	0	2	22
[GAVf]	.46	.63	0	1	23
[P]	.10	.58	2	2	20
π	.32	.92	3	2	19
.	.45	.75	4	2	18
[PaLT]	.28	.83	2	1	21
[VTS]	.48	.63	2	1	21
[PLSD]	.29	.75	2	1	21
π	.17	.67	2	1	21
[GpvI]	.47	.67	3	3	18
[PYFE]	.58	.54	3	4	17
Other turn/coil					
#10 32					
[PAtS]	.14	.75	7	1	24
[GASd]	.11	.78	6	0	26
[AS]	.17	.84	4	0	28
[S]	.08	.81	2	0	30
[tSnq]	.11	.71	2	3	27
Helix N-cap					
#11 61					
[GkdE]	.16	.75	9	7	45
π	.16	.80	6	8	47
.	.41	.41	4	9	48
[TSND]	.06	.83	4	7	50
π	.27	.77	44	1	16
[AkDE]	.16	.95	51	1	9
[qDE]	.08	.78	54	3	4
[aVIL]	.64	.13	56	2	3
π	.28	.63	53	2	6
π	.19	.91	51	2	8
[AVLM]	.52	.26	50	1	10
[ILmf]	.49	.42	47	0	14
π	.16	.88	43	0	18
Schellman Helix C-cap					
#12 68					
π	.25	.70	53	2	13
[aILf]	.59	.16	58	1	9
[AVL]	.47	.35	58	1	9
[rKDE]	.12	.79	60	1	7
π	.28	.58	59	1	8
[L]	.69	.16	57	0	11
[ARKE]	.22	.82	52	1	15
π	.16	.89	48	2	18
[ALRk]	.32	.69	27	2	39
[G]	.06	.70	11	3	54
[aviL]	.49	.44	13	8	47
Schellman Helix-Turn-Sheet					
#13 67					
[ArKd]	.33	.75	53	7	7
.	.52	.40	56	7	4
[AL]	.82	.16	58	4	5
[AiLn]	.41	.75	56	4	7
[AqKd]	.27	.88	49	3	15
.	.47	.72	17	2	48
[G]	.07	.79	1	1	65
[AVI]	.68	.33	2	4	61
[arKD]	.17	.74	4	24	39
[VIL]	.62	.37	7	43	17
[VIL]	.72	.31	7	48	12
[GVIL]	.60	.36	6	49	12
φ	.70	.30	6	45	16

subcluster the segments in each neighborhood into different structural classes (Table 2). A substantial fraction of the neighborhoods contained two different types of local structures (Table 2, row 2). To assess the statistical significance of the results, parallel experiments were carried out on a simulated data set in which the sequence-structure relationships of the individual segments were randomized. Importantly, sequence segments in the same neighborhood are restricted to one, two, or three local structures far more often in the HSSP data sets than in the simulated data sets (Table 2). Thus, the sequence-structure relationships we observe are distinctly nonrandom.

The sequence patterns strongly associated with particular local structures are an interesting combination of previously studied and new motifs (Table 3). Familiar motifs include amphipathic patterns with hydrophobic residues separated by two or three positions almost exclusively found in α-helices (Table 3, patterns 1 and 2), or with hydrophobic residues separated by one position very frequently occurring in surface β-strands (Table 3, patterns 3 and 4). A less strongly amphipathic pattern (pattern 5) was found in somewhat buried helices. A number of short patterns with conserved glycine and proline residues occur predominantly in turns as expected [Table 3, patterns 6–9 (14)]. Pattern 10 is a serine-rich turn. Pattern 11 is similar to a classic N-terminal helix cap motif (3) and indeed is found predominantly in helix N caps. Pattern 12 is close to the Schellman helix C-cap (2, 15) and is found predominantly at the C termini of α-helices.

Several patterns extend and/or refine previously characterized motifs. Pattern 13 is an extension of the Schellman motif; following the characteristic helix–turn transition is a hydrophobic stretch that is almost always part of a β-sheet. Pattern 14 is very similar to a previously described motif [the α-L motif (2)], but surprisingly it appears primarily in strand C-caps rather than in the helix C-caps where it was originally described.

A number of the patterns that correlate very strongly with local structure have not been explicitly singled out in the literature. A strongly hydrophobic stretch in pattern 15 is almost always found in buried β-strands (note low average solvent accessibility in column SA). Patterns 16 and 17 are found in transitions from amphipathic helices through an

Table 3. (*Continued*).

Patterns	Hϕ	SA	H	S	T
αL,Strand C-cap					
#14 48					
π	.25	.58	2	21	25
.	.40	.54	1	33	14
[VIl]	.88	.15	1	41	6
[pAVi]	.63	.25	1	40	7
[VIl]	.78	.19	2	41	5
[VIF]	.66	.29	2	34	12
[GA]	.30	.19	3	23	22
.	.43	.38	3	15	30
[G]	.05	.52	2	9	37
π	.33	.62	6	6	36
[GPVn]	.34	.58	6	7	35
Buried strand					
#15 37					
[AVIl]	.82	.16	5	29	3
[ViLm]	.75	.13	4	32	1
[IL]	.88	.08	4	32	1
[VILf]	.81	.16	3	30	4
[G]	.06	.21	2	23	12
Other Helix-Turn-Sheet					
#16 34					
.	.54	.44	26	2	6
[aVIL]	.65	.32	25	4	5
.	.40	.73	24	4	6
[ALsK]	.44	.74	25	4	5
[aViL]	.69	.47	22	3	9
.	.42	.76	21	1	12
π	.34	.85	16	1	17
π	.26	.94	12	1	21
π	.30	.73	4	2	28
[PaVi]	.50	.41	5	3	26
[aDE]	.14	.77	5	6	23
[aVIL]	.72	.18	5	22	7
[VIL]	.85	.15	5	25	4
[AVIL]	.71	.17	5	27	2
.	.60	.26	5	26	3
#17 41					
[ASRK]	.22	.87	34	1	6
π	.27	.90	30	0	11
[AiLy]	.61	.46	17	0	24
[G]	.03	.85	3	0	38
[gAV]	.72	.29	3	5	33
π	.20	.75	4	20	17
[VIL]	.77	.36	4	29	8
Other helix C-cap					
#18 56					
[VILf]	.64	.07	49	3	4
π	.25	.44	51	2	3
[AKDE]	.11	.87	52	2	2
.	.43	.23	53	2	1
[ILM]	.59	.10	49	2	5
[atRK]	.13	.78	46	2	8
π	.15	.87	36	1	19
[aLf]	.44	.62	20	1	35
[G]	.04	.83	5	0	51
.	.37	.58	6	2	48
[sNKD]	.09	.76	9	7	40
#19 41					
.	.41	.56	34	0	7
π	.24	.68	35	1	5
[aLd]	.72	.22	38	1	2
[VILy]	.82	.17	37	1	3
[AqrK]	.28	.81	38	2	1
[lqRk]	.26	.68	36	2	3
[IL]	.81	.22	34	1	6
[LRK]	.40	.66	30	1	10
π	.19	.85	24	1	16
[SNKD]	.22	.76	15	0	26
π	.24	.71	9	2	30
.	.37	.81	10	4	27
Sheet N-cap					
#20 44					
π	.17	.86	3	6	35
[G]	.09	.84	1	3	40
π	.16	.72	0	6	38
[VTrK]	.19	.81	2	30	12
[VI]	.71	.20	2	35	7
[AViT]	.37	.47	2	37	5
[VILw]	.74	.11	3	39	2
[ViFT]	.47	.31	3	36	5
[GAiS]	.33	.43	3	28	13
#21 41					
[Gk]	.06	.78	5	3	33
[As]	.72	.29	5	5	31
[avkD]	.25	.68	5	17	19
[aVir]	.57	.29	4	32	5
[VI]	.84	.26	4	34	3
[VILs]	.74	.29	4	35	2
[VIL]	.81	.17	4	33	4
Sheet C-cap					
#22 38					
[ViLy]	.74	.15	3	30	5
[AVi]	.73	.18	3	31	4
[VI]	.88	.15	3	31	4
[VILF]	.80	.13	4	27	7
[G]	.05	.15	3	11	24
[Gasd]	.23	.26	3	6	29
[Gvs]	.30	.50	7	7	24

For each neighborhood, the first row gives the identifier and the number of segments in the neighborhood; the subsequent rows contain summary statistics on each position. Letters within brackets indicate the prominent amino acids at the corresponding position in the neighborhood: capitals indicate frequencies greater than 0.1, lowercase letters, frequencies between 0.07 and 0.1. For example, the third position in the nine-residue pattern characterizing neighborhood 1 is rich in alanine, arginine, and lysine. Positions at which more than seven different amino acids occurred with frequencies greater than 0.05 are represented by π, ϕ, and . for average hydrophobicities of less than 0.35, greater than 0.65, and between 0.35 and 0.65, respectively. Hϕ is the sum of the frequencies of occurrence of alanine, valine, isoleucine, leucine, methionine, proline, phenylalanine, and tryptophan. Solvent accessible surface areas (SA) were taken directly from the HSSP files and then normalized by the exposed area of amino acids in Ala-Xaa-Ala tripeptides. Residues with less than 16% of their surface exposed were considered buried. Columns H, S, and T are the number of segments in the neighborhood that are in helix, strand, or turn–loop configurations. Patterns 13, 14, 16, and 22 have consistency scores slightly below the 80% threshold.

exposed loop to a buried β-strand. Pattern 18 is a helix C-cap with a conserved glycine, but otherwise different than the Schellman motif. Pattern 19 is a helix C-cap with turn-favoring residues (Ser, Asn, and Lys) instead of a conserved glycine. Patterns 20 and 21 are found in transitions from turns to strands, and pattern 22, in transitions from strands to turns. The two latter classes of patterns link well-studied short reverse turns with specific types of β-strands. Analysis of the three-dimensional contexts in which these patterns occur is currently under way and should yield insights into the specific interactions responsible for the prevalence of particular local structures.

Because nonlocal interactions play an important role in protein three-dimensional structures, local sequence–structure relationships are not absolute. It should be noted that with the 80% consistency threshold used here, up to 20% of the sequence segments in the neighborhoods described in Tables 1 and 2 may adopt local structures different from that of the majority of sequence segments in the neighborhood. Furthermore, the ~20% of positions in the HSSP data set not accounted for in Table 2 belong to neighborhoods in which the consistency of the local sequence-to-structure mapping is not significantly greater than that observed in the simulated data set.

DISCUSSION

Our approach uses the vast amount of available sequence data as a guide to identify natural structural groupings that otherwise may be hidden by the complexity of protein three-dimensional structures. The two major results are the description of the overall features of the local sequence-to-structure mapping (Tables 1 and 2) and the identification of most of the sequence patterns in proteins that consistently occur in a particular type of local structure (Table 3). The identification of sequence patterns that correlate strongly with structure has proceeded in a rather piecemeal fashion in the past (most studies have focused on a particular type of local structure and sought to determine whether the sequences found in the structural element in proteins have any distinguishing features); our automated approach has in one pass probably identified virtually all of such patterns.

Table 4. Selected sequence patterns with two prominent structures

	Pattern	H	S	T	HT	TH	ST	TS
1	[GaqE] ϕ [ILT] [VILF] [AV] [AVI] [aLmt] [AV] π	19	18					
2	. π . [GD] [AII] [ALmE] [AVLE] [iL] [YhRKE]	15		10				
3	[LsR] π [iL] [ATs] π . ϕ π π	15		11				
4	. [N] π . [GVsR] ϕ . π π	11		17				
5	[As] [qhRKE] . [pVIL] [aviLF] π . [pLqe] [aRK]	18			14			
6	π π [LF] . π [AsKDE] [vSnkDE] π π π	11			10			
7	. [pAV] [Lf] [ILf] [GAS] [GAs] [VLh] [G] [G]	10					14	
8	[aVi] . [VIL] . . [VIL] π π [G]	7					10	
9	[gAT] π [GIkdE] π π [IL] [VILk] [aRkE] [aVIk]	16						8
10	[gAs] [GSDE] [IWRk] [VILe] [VIL] [GAs] ϕ [nD] .	14						19
11	π [yF] [TND] [PA] π [AVsR] . [PAVi] π			11		11		
12	π . [AVL] . [IyFw] [NrD] P π [ALSQ]			19			13	
13	π . [ytnDe] π [PsnD] [G] . [VIY] π			20				25
14	[ASK] [G] [SNK] [yFT] π π [AVL] [vIL] [ViL]					11		11
15	. π [G] [nD] [aLQ] . [GAS] [ALs] [aLmF]					10		13

Abbreviations are as in Tables 1 and 3.

An important issue for approaches to protein structure prediction is the extent to which a local "stereochemical code" operates between sequence and structure (2). Our results have both positive and negative implications for the success of such a code. First, we do find a number of patterns that correlate with local structure, and have not been heretofore described (the α-turn-β motif in Table 3, for example). Because the patterns were generated using unsupervised learning methods, they are probably not optimal for the classification problem (12), but refinement of neighborhood boundaries using structural information could yield some improvement in local structure prediction. However, Table 3 shows that currently well-studied motifs dominate the set of patterns that correlate strongly with structure, suggesting that recent success with helix capping motifs (2, 3) may not generalize to a large fraction of other local structure elements. Secondary structure prediction efforts have traditionally had more difficulty with β-strands, presumably because of their greater dependence on nonlocal interactions, and indeed, β-strands are conspicuously underrepresented in the set of patterns that correlate strongly with structure (Table 1). With regard to the question of the contribution of hydrophobicity patterns alone to sequence–structure relationships, we found that considerable resolution was lost, particularly in the case of structural transitions, when sequences were represented using a two-letter hydrophobic-polar code (data not shown).

The explicit treatment of the ambiguity of the local sequence–structure mapping could have useful application to the prediction of tertiary structure from primary sequence. Examples of sequence patterns that are found in two distinct local structures are shown in Table 4. Most work on local structure prediction has sought to specify uniquely the local structure of a protein segment given the sequence. The demonstration that sequence patterns may correlate with two specific local structures out of a larger set of possible structures has immediate relevance to the global protein structure prediction problem because it suggests a means to greatly reduce the size of conformational space. Such a reduction in the size of the space could readily be incorporated into a search procedure in which only a limited number of local conformations are allowed as a global energy function involving nonlocal interactions is minimized.

The use of sequence patterns to identify structural motifs opens a new paradigm for studies of protein structure. The amount of available sequence data is vast and growing rapidly, and one-dimensional sequences are much more amenable to pattern recognition approaches than are three-dimensional protein structures. The striking correlation we observe between a number of sequence patterns and local protein structure is probably only the first indication of the power of such "inverse" approaches.

We dedicate this paper to the memory of Sharon C. Han (1941–1995). We thank H. Schneider and C. Sander for the HSSP database, D. A. Agard for encouragement and computational resources, and S. Henikoff, J. Henikoff, S. Pietrokovski, K. Zhang, D. Gerloff, N. Hunt, T. Defay, R. Klevit, and members of the Baker laboratory for critical reading of the manuscript. K.F.H. is supported by a Howard Hughes Medical Institute Predoctoral Fellowship. This work was partially supported by National Science Foundation, Science and Technology Center Cooperative Agreement BIR-9214821, the Merck Research Laboratories, and young investigator awards to D.B. from the National Science Foundation and the Packard Foundation.

1. Harper, E. T. & Rose, G. D. (1993) *Biochemistry* **32,** 7605–7609.
2. Aurora, R., Srinivasan, R. & Rose, G. D. (1994) *Science* **264,** 1126–1130.
3. Presnell, S., Cohen, B. & Cohen. F. E. (1992) *Biochemistry* **31,** 983–993.
4. Zhou, H. X., Lyu, P., Wemmer, D. E. & Kallenbach, N. R. (1994) *Proteins* **18,** 1–7.
5. Kabsch, W. & Sander, C. (1984) *Proc. Natl. Acad. Sci. USA* **81,** 1075–1078.
6. Rost, B. & Sander, C. (1993) *J. Mol. Biol.* **232,** 584–599.
7. Dayhoff, M. O., Eck, R. V. & Park, C. M. (1972) in *Atlas of Protein Sequence and Structure*, ed. Dayhoff, M. O. (Natl. Biomed. Res. Found., Washington, DC), pp. 89–99.
8. Henikoff, S. & Henikoff, J. G. (1992) *Proc. Natl. Acad. Sci. USA* **89,** 10915–10919.
9. Han, K. F. & Baker, D. (1995) *J. Mol. Biol.* **251,** 176–187.
10. Hobohm, U. & Sander, C. (1994) *Protein Sci.* **3,** 522–524.
11. Sander, C. & Schneider, R. (1991) *Proteins* **9,** 56–68.
12. Duda, R. O. & Hart, P. E. (1970) *Pattern Classification and Scene Analysis* (California Artificial Intelligence Group, Stanford Research Institute, Menlo Park, CA).
13. Unger, R. & Sussman, J. L. (1993) *J. Comput. Aided Mol. Des.* **7,** 457–472
14. Cohen, F., Abarbanel, R., Kuntz, I. & Fletterick, R. (1986) *Biochemistry* **25,** 266–275.
15. Schellman, C. & Jaenicke, R., eds. (1980) *Protein Folding: Proceedings of the 28th Conference of the German Biochemical Society* (Elsevier/North-Holland, New York), pp. 53–61.