J. Mol. Biol. (1995) **251**, 176–187

JMB

Recurring Local Sequence Motifs in Proteins

Karen F. Han[1] and David Baker[2]*

[1]*Graduate Group in Biophysics, University of California, San Francisco CA 94143, USA*

[2]*Dept of Biochemistry University of Washington Seattle, WA 98195, USA*

**Corresponding author*

We describe a completely automated approach to identifying local sequence motifs that transcend protein family boundaries. Cluster analysis is used to identify recurring patterns of variation at single positions and in short segments of contiguous positions in multiple sequence alignments for a non-redundant set of protein families. Parallel experiments on simulated data sets constructed with the overall residue frequencies of proteins but not the inter-residue correlations show that naturally occurring protein sequences are significantly more clustered than the corresponding random sequences for window lengths ranging from one to 13 contiguous positions. The patterns of variation at single positions are not in general surprising: chemically similar amino acids tend to be grouped together. More interesting patterns emerge as the window length increases. The patterns of variation for longer window lengths are in part recognizable patterns of hydrophobic and hydrophilic residues, and in part less obvious combinations. A particularly interesting class of patterns features highly conserved glycine residues. The patterns provide a means to abstract the information contained in multiple sequence alignments and may be useful for comparison of distantly related sequences or sequence families and for protein structure prediction.

Keywords: multiple sequence alignments; sequence comparison; substitution matrices; protein structure prediction; sequence motifs

Introduction

Are there recurring local patterns in the amino acid sequences that encode proteins? Global similarity is often used to classify sequences into families; are there local patterns that transcend family boundaries?

Given that all viable protein sequences must be such that the proteins they encode can fold and have at least marginal stability, it is reasonable to expect that not all 20^N amino acid sequences of length N are equally probable. There are far too few distinct protein families to tabulate meaningful statistics on the frequencies of occurrence of the different peptides of length N for N greater than two (Gonnet *et al*., 1994). An alternative approach is to use cluster analysis to identify recurring sequence patterns. This requires a suitable measure of similarity between two sequences.

Global sequence comparisons almost always rely on amino acid substitution matrices compiled by averaging over large sets of related sequences. The disadvantages of using a single substitution matrix have been pointed out on numerous occasions (Johnson *et al*., 1993; Risler *et al*., 1988). The major problem is that at different positions in protein structures, different sets of amino acid sequences are likely to substitute for one another. In other words, there is no single and universally applicable set of distances (or similarities) between the 20 amino acids. Rather, similarity can be quite context-dependent.

A more natural measure, which does not require the assumption of a single substitution matrix, is available for comparison of protein families if there are a number of sequences in each family. For each position in a set of multiply aligned sequences, one can calculate the frequency of occurrence of each of the amino acids. The resulting sequence of frequency distributions is often called a profile (Gribskov *et al*., 1990). To evaluate the distance between two aligned profile segments, one can compare the frequency distributions at corresponding positions.

Here we use such a distance measure in conjunction with cluster analysis to identify patterns that occur frequently in multiple sequence alignments for proteins of known structure. Because only one multiple sequence alignment is included of each family, the patterns are necessarily common to many different protein families and are distinct from the family-specific patterns compiled in the Prosite database (Bairoch & Bucher, 1994). Because the patterns are universal but still fairly detailed, they present a possible route to overcoming some of the limitations

0022–2836/95/310176–12 $08.00/0

of the global amino acid substitution matrices used in sequence comparisons and the individual residue secondary structure and solvent accessibility propensities used in local protein structure prediction. The work described in this paper is a first step towards correlating local sequence patterns with local structural motifs.

Results

If there are a finite number of distinct chemical environments in proteins, there should be a finite number of patterns of variation in sets of multiply aligned sequences. Here we use cluster analysis to identify recurring patterns of variation at single positions and in short segments of contiguous positions in multiple sequence alignments. A non-redundant set of global multiple sequence alignments for proteins of known structure was extracted from the HSSP database (Sander & Schneider, 1991) as described in Methods. After excluding positions in which fewer than 20 sequences contributed to the alignment, the data set contained approximately 20,000 individual columns from 154 protein families.

Patterns at single positions

The frequencies of occurrence of the 20 amino acids at each position were calculated, and the K-means algorithm was used to group similar frequency distributions using the simple ''city block'' metric (*d1*, see Methods).

The amino acid groupings obtained (Table 1) are consistent with expectation. The mean of the frequency distributions belonging to a given cluster provides a convenient summary statistic. To save space, the mean values of each of the 20 amino acids in each cluster are not shown, instead only the amino acids whose mean frequency of occurrence in a cluster is greater than 0.1 (upper case) or between 0.07 and 0.1 (lower case) are listed (Table 1, column 3). The degree of conservation of these primary components is reflected in the variability index (column 4), which gives the number of amino acid components whose mean frequency of occurrence is greater than 0.05.

The patterns generally fall into either hydrophobic (clusters 1, 2 and 3) or polar (clusters 4 through 8) classes (Table 1, column 6). However, the different clusters contain different combinations of hydrophobic and hydrophillic groups. For example, cluster 1 contains primarily V, I and L while cluster 2 contains primarily I, L and M. Cluster 3 contains only aromatic residues while cluster 6 contains only negatively charged residues. Amino acid residues with special structural properties are prominent in clusters 9 (P) and 10 (G). Although the RMS deviation of points within a cluster is not dramatically less than that of points in the entire dataset (see Methods), the products of the variances are considerably lower in the former than in the latter (Table 1, column 6). As outlined in Methods, the patterns were independent of the choice of starting cluster centers implicit in the K-means algorithm. Patterns similar to those in Table 1 were obtained in a Dirichlet mixture decomposition of multiple sequence alignments (Brown *et al*., 1993).

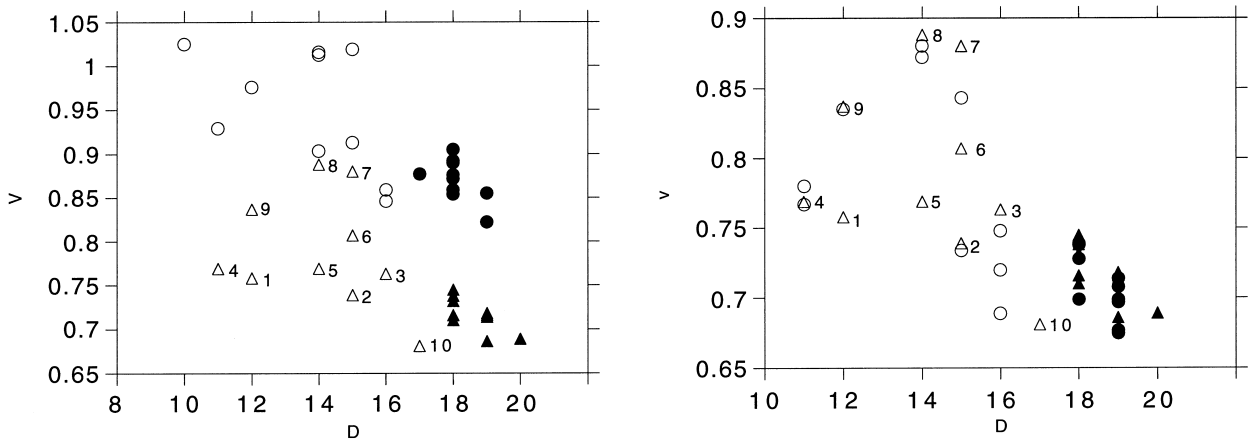
The first ten patterns in Table 1 are the result of a low resolution subdivision of sequence space (ten classes were allowed). More subtle patterns are revealed when the number of classes is increased (see Methods). For example, in cluster 11, primarily L, R and K, the common feature is the long aliphatic side-chain of all three residues. Pattern 13 is dominated by the beta branched residues V, I and T. A cluster with conserved cysteine residues also emerges when more classes are allowed. Thus, although hydrophobicity appears to be the major feature distinguishing the largest clusters, other chemical properties are often important in the smaller clusters.

How clustered are the frequency distributions in sequence space? The K-means algorithm can always

Table 1. Recurrent patterns at individual positions

Cluster no.	No. of members	Dominant substitutions	Variability index	Hydrophobicity	Relative cluster volume
1	2449	V,I,I	3	0.832	2.3e-4
2	1971	L,i,m	5	0.853	5.4e-4
3	1521	Y,F,w	4	0.818	1.6e-3
4	1166	N,H,d	4	0.151	7.4e-4
5	2263	R,K,q	4	0.163	5.8e-3
6	2396	D,E	4	0.148	2.5e-3
7	1401	T,s	3	0.237	2.4e-4
8	1412	S,a,t	3	0.199	3.3e-4
9	2214	P,A	3	0.538	1.2e-3
10	1349	G	2	0.166	4.3e-4
11	84	L,R,K	4	0.450	2.8e-3
12	150	G,N,k	4	0.101	2.4e-6
13	114	V,I,T	4	0.687	1.1e-5

The amino acids which occur with frequencies greater than 0.1 are shown in upper case, those which occur at frequencies between 0.07 and 0.1, in lower case (column 3). The number of amino acids which occur at frequencies greater than 0.05 is given in column 4. The average summed frequency of occurrence of the amino acids A, V, I, L, M, F, W and C is listed in column 5.



(a) (b)

	Pattern conservation	Real data cluster dimensionality Mean	Real data cluster dimensionality Variance	Simulated data cluster dimensionality Mean	Simulated data cluster dimensionality Variance
Weighting scheme					
1	1.0	11.8	0.54	16.3	0.17
2	0.9	11.8	0.54	16.5	0.15
3	0.9	12.1	0.47	16.5	0.18
4	0.7	12.6	0.46	16.5	0.20
Distance measure					
5	0.8	11.8	0.45	17.5	0.06
6	0.9	11.8	0.51	16.7	0.08
7	0.7	11.1	0.84	14.9	0.25
8	1.0	11.3	0.39	13.5	0.02

(c)

Figure 1. Comparison of different weighting schemes and distance measures for both the real and random data sets. Each symbol represents a single cluster; the *x* axis is the number of non-zero dimensions, the *y* axis, the average variance. (a) Comparison between unweighted (triangles) and weighted (scheme 2, circles) data sets. (b) Comparison between the city block metric (triangles) and distance measure *d2* (circles). Open symbols, clusters generated from the real data set; filled symbols, clusters generated from a simulated data set. The clusters for the real set are numbered as in Table 1. Note that the weighting scheme changes the residue frequency distributions such that the within-cluster variance is higher for both real and simulated data sets. (c) Summary of statistics for the different weighting schemes and distance measures. Column 1 describes the trial, column 2 lists the fraction of patterns that were found in the unweighted data set clustered using the city block metric (trial 1), columns 3 and 4 list the mean and variance of the cluster dimensionality for the real data set; columns 5 and 6, the same quantities for the simulated data set. The city block metric was used for the comparison of weighting schemes (trials 1 to 4), and the unweighted data set was used for the comparison of distance measures (trials 5 to 8). The weighting scheme trails are: 1, no weights; 2, tree-based weights; 3, self consistent weights; 4, Voronoi weights (see the text for more description). The distance measure trial 5 utilized *d2* and trials 6 to 8 utilized *d3* with the matrix **M** the PAM(250) substitution matrix, the overall covariance matrix, and within cluster covariance matrices, respectively (see Methods). For trial 8, covariance matrices were calculated for each of the clusters generated using the standard procedure (trial 1) and used for a second round of clustering as described in Methods.

subdivide a set of points into convex subsets and does not depend on the ''clumpiness'' of the data. To investigate this question, random data sets were generated using the individual residue frequency distributions of the HSSP database but lacking the inter-residue correlations (see Methods). The HSSP data set and a simulated data set were subjected to the same clustering procedure and the results are compared in Figure 1.

As described in Methods, no single statistic adequately captures the spread of points within a cluster embedded in a high dimensional space. With two statistics one can do much better. We have used *V*, the within cluster variance per dimension, and *D*, the dimension of the smallest subspace that contains the cluster. Each cluster is represented as a point in Figure 1.

The most striking aspect that distinguishes the results of application of the K-means algorithm to the real (Figure 1(a), open triangles) and simulated (Figure 1(a), closed triangles) data sets is the smaller number of dimensions in the former. There is also significantly greater variation in the number of dimensions per cluster in the real data set. The clusters in the random set appear to have roughly similar shapes and volumes, as expected in a rela-

tively uniform distribution. In contrast, the sizes and shapes of the clusters obtained for the real data set vary considerably, presumably because different sequence patterns in protein families are constrained to different extents.

Comparison of weighting schemes and distance measures

Frequency distributions from multiple sequence alignments can be taken as estimators of the ''true'' probability distributions for substitution of the 20 amino acids at a given position in a protein, but there are two important caveats. First, there are a limited number of sequences in each family, so that observed frequencies may be inaccurate estimates because of small sample size effects. We have dealt with this problem by excluding poorly represented families and positions from the analysis. Second, and perhaps more serious, the different sequences in a family are not independent observations. Rather, they are highly correlated. Frequency distributions derived from sets of evolutionarily related sequences may be heavily biased. A particular amino acid may be highly represented in a particular position simply because it was present in a common ancester, and not because of any underlying structural constraint.

A number of different weighting schemes have been proposed for compensation of the heavily biased sampling in evolutionarily related sequence sets (Vingron & Sibbald, 1993). We experimented with (1) a weighting scheme similar to that described by Altschul *et al*. (1989) and van Ooyen & Hogeweg (1990) in which weights are derived from a tree constructed from pairwise distances between the aligned sequences; (2) the self-consistent weighting scheme of Sander & Schneider (1991), and (3) the Monte Carlo approach to estimating Voronoi volumes described by Sibbald & Argos (1990). Frequency distributions were recalculated for each of the weighting schemes and subjected along with corresponding simulated data sets to the K-means clustering procedure.

Space limitations prohibit the display of scatter plots for each of the weighting schemes. However, the essence of these plots can be roughly captured by the mean and variance of D, the cluster dimensionality (Figure 1). The results obtained with frequency distributions weighted using scheme (1) were very similar to those obtained with the unweighted distri butions (Figure 1, compare circles to triangles).

The average cluster dimensionality was very similar for all the weighted data sets (Figure 1(c), column 3), indicating that the interrelationships among the frequency distributions are not substantially changed by the different weighting schemes. Furthermore, the resulting sequence patterns were not greatly altered by any of the weighting schemes (Figure 1(c), column 2). Since both the relative weight on a particular sequence and the probability of misalignment increase with sequence divergence, attempts at correcting the biased sampling through unequal sequence weighting may increase noise from misalignment errors. Because of the lack of dependence of the results on the weighting scheme, unit weights were used for simplicity in the experiments described in the following sections.

A similar approach was used to evaluate alternative distance measures. The Euclidean distance metric gave results very similar to that of the city block metric *d1* (data not shown). Because differences between amino acid frequencies of 0.8 and 0.6 are likely to be less significant than differences between frequencies of 0.2 and 0.0, we experimented with the somewhat *ad hoc* distance measure *d2* which effectively down-weights differences of the former type. Again, the clusters obtained with distance measures *d2* had similar overall properties to those obtained with *d1* (Figure 1(b)). We also experimented with a PAM (250) matrix based distance measure and with the use of the overall covariance matrix as well as individual cluster covariance matrices to adjust for the different frequencies of the different amino acids and to relax the assumption of spherical clusters implicit in the K-means algorithm (see Methods for details).

As summarized in Figure 1(c), the different distance measures gave qualitatively similar results, with the real data set consistently more clustered than the random data set (Figure 1(c), columns 2 and 4). The simplicity of the city block metric and the Euclidean metric makes them preferable over the other distance measures. Because of complications associated with the use of the Euclidean metric for clustering frequency distributions (see Methods), the city block metric was chosen for the studies described in Tables 1 and 3. The lack of sensitivity to the details of the weighting scheme and distance measure argue that the groupings shown in Table 1 are inherent in the data and not simply imposed by the clustering algorithm, a conclusion supported by the degree to which the patterns agree with intuition.

Results of contiguous position classification

The clustering procedure can be readily generalized to treat segments of contiguous positions as described in Methods. To investigate the types of patterns occurring on different length scales, the clustering procedure was repeated for segment lengths ranging from 3 to 15 residues using a fixed number (200) of clusters. Table 2 lists the average cluster dimensionality per position for both the real and simulated data sets. As the window length increased, the variation in the average number of dimensions increased (Table 2, column 4). In contrast, the variation of the simulated data set was relatively constant (Table 2, column 6). Thus, the clusters adopt a wider range of shapes at larger window lengths.

Space limitations preclude the description of the patterns for each segment length. Instead, the following analysis is focused on the results for segment length nine. A detailed description of all patterns for window lengths two to 15 can be obtained from the authors.

Table 2. Results for clustering segments of contiguous positions into 200 classes

Window length	Number of vectors	Real data Cluster dimensionality Mean	Real data Cluster dimensionality Variance	Simulated data Cluster dimensionality Mean	Simulated data Cluster dimensionality Variance
3	21,146	12.07	0.57	14.79	0.16
5	20,812	13.51	0.66	15.76	0.15
7	20,483	14.22	0.85	16.36	0.12
9	20,157	14.50	1.42	16.48	0.11
11	19,833	13.75	2.35	16.72	0.18
13	19,517	14.79	2.61	16.80	0.15
15	19,204	14.69	3.39	16.95	0.17

Note that as the window length increases, the total number of frequency vectors decreases slightly as $N - 1$ positions are lost from each end of each sequence, where N is the segment length.

Patterns for nine consecutive positions

The distribution of clusters obtained for segment length nine is shown in Figure 2 for both the real (open triangles) and simulated (closed triangles) data sets. As in Figure 1, the clusters in the real data set are consistently lower in dimensionality than those in the random data set. The former also have a greater variety of dimensions and shapes.

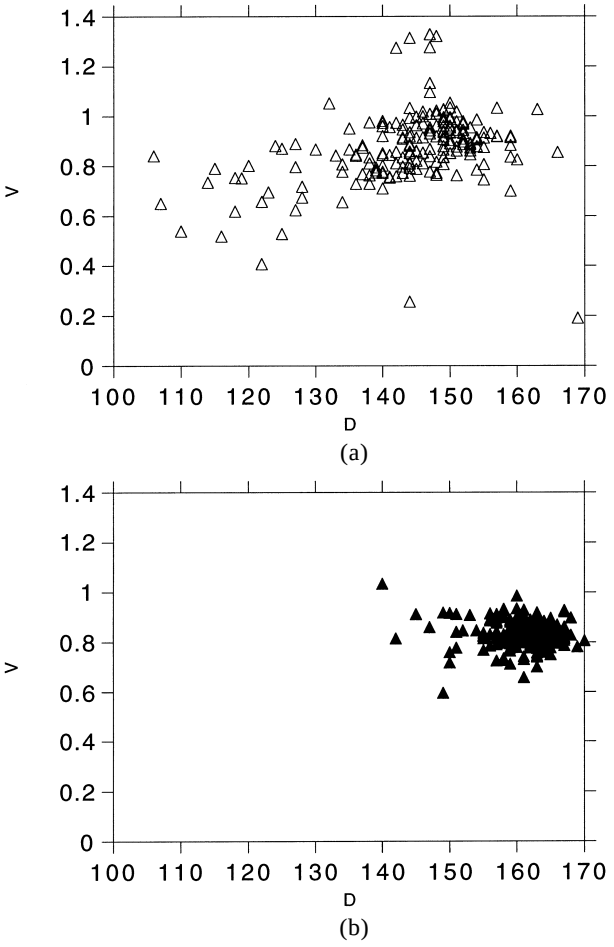


Figure 2. Scatter plots of the number of non-zero dimensions (D) and average variance (V) for each cluster obtained in the nine-consecutive position classification. (a) Real data set; (b) simulated data set.

Several of the patterns for window length nine are described in detail in Table 3A along with relevant statistics. Space limitations preclude the description of even a modest number of clusters in this detail; instead we have adopted a more compact representation (Table 3B) to show a number of common patterns found in three separate classifications using different random starting cluster means. Because the distance calculation assumes a one to one correspondence between the positions of the segments being compared, frame shifted patterns are frequently observed in which for example positions 1 to $(N - 1)$ of pattern 1 are very similar to positions 2 to N of pattern 2. To save space such frame shifted patterns are shown only once. Clusters containing less than 25 members are omitted.

As expected, many substitution patterns at individual positions are similar to those observed in the single position clustering (compare the single position substitution patterns in brackets of Table 3B to Table 1). However, because the averaging is also constrained by neighboring sequence patterns, there appear to be more subtle patterns in the contiguous sequence clusters (e.g. compare positions 1 and 3 in cluster 40).

The patterns fall roughly into three broad categories which are illustrated by the examples in Table 3A. The first and largest category consists of patterns with pronounced amphipathicity. The first cluster in Table 3A belongs to this category; a number of additional patterns are shown in more condensed form in Table 3B (section (g)). In some positions, those in which the average hydrophobicity is either very high or very low, but the variability index is high, a simple H / P reduced code is clearly sufficient. For example, most positions in cluster 3 in Table 3A are strongly hydrophobic but eight amino acid residues occur more than average. In contrast, the relative hydrophobicity index in some positions is at one extreme or the other, but only particular residues are allowed. For example, position 4 in cluster 44 tolerates only aromatic residues, while position 1 in the same cluster prefers V and I. In some cases, side-chain size appears to be important, perhaps because of packing effects. Patterns 19, 22 and 23 contain positions in which small (A), medium (L)

Table 3. Recurring patterns for nine consecutive positions

A. *Detailed statistics for several selected clusters*

Cluster number	Size	V	D	Prominent AA	Variability index	Hydrophobicity	Relative cluster volume
1	134	0.94	15.3	VIL	5	0.73	5.3E-5
				TSd	6	0.19	2.1E-5
				paRkDE	8	0.26	4.4E-2
				AskDE	6	0.20	2.4E-3
				QDE	3	0.15	8.4E-2
				AVIL	6	0.72	6.2E-2
				AkdE	8	0.30	1.4E-2
				AqrkE	8	0.30	3.4E-1
				aViL	6	0.68	4.9E-2
2	210	0.95	14.4	aKDE	6	0.21	2.0E-2
				AkDE	8	0.28	5.2E-2
				aVIL	5	0.77	5.1E-2
				ArKE	8	0.32	1.2E-2
				AKDE	9	0.22	3.1E-2
				AL	9	0.55	8.0E-2
				VILf	6	0.83	5.4E-2
				aqRKE	7	0.24	6.3E-2
				ArKdE	9	0.25	3.2E-2
3	148	0.39	11.7	avILFt	8	0.70	3.8E-12
				AILFw	8	0.74	4.1E-10
				GAiLFts	8	0.64	2.7E-13
				GAvILF	9	0.73	1.1E-10
				GAvILFts	8	0.66	2.4E-10
				GAvILFts	8	0.70	2.9E-13
				gAvILF	8	0.76	2.0E-10
				avILF	6	0.79	1.1E-10
				gAiLF	9	0.68	2.2E-11
4	133	1.02	15.6	GAvts	7	0.41	1.7E-3
				GAs	9	0.40	4.2E-4
				GAvs	7	0.44	2.1E-1
				GAvs	7	0.41	1.4E-1
				GAvs	7	0.41	5.3E-1
				GAls	7	0.41	4.9E-1
				GAls	7	0.41	2.6E-1
				GpAs	8	0.39	3.1E-1
				GAvls	8	0.40	4.2E-1
5	107	0.67	16.11	TS	4	0.25	3.5E-7
				avTS	8	0.41	2.1E-8
				gaTSD	7	0.25	3.2E-6
				gpatSnD	8	0.23	1.9E-8
				gtSd	10	0.33	5.1E-5
				paTSnq	8	0.30	8.6E-5
				aTSn	7	0.30	5.5E-5
				PaTSn	7	0.25	1.3E-4
				gptSnd	7	0.24	5.3E-4
6	74	0.67	16.11	Gpavlk	8	0.39	8.4E-5
				Vlf	9	0.58	1.6E-3
				ilTr	9	0.45	1.1E-2
				P	2	0.20	2.0E-4
				VILe	6	0.66	2.4E-3
				ailyFe	8	0.67	3.0E-3
				aVltSn	7	0.45	6.1E-3
				G	1	0.06	9.4E-6
				GalfSk	7	0.39	9.2E-4
7	67	0.88	15.33	AkDE	8	0.30	1.1E-4
				AILfk	7	0.55	2.2E-2
				AL	5	0.59	2.1E-2
				ARKe	7	0.27	7.1E-3
				gaKE	8	0.29	5.1E-3
				aLyf	8	0.56	2.7E-2
				G	1	0.04	7.7E-4
				aVIL	6	0.62	1.0E-2
				tskDe	7	0.28	1.6E-2

Table 3—*continued*
B. Condensed representation of selected patterns

	Size	V	D
(a) Conserved glycine			
1. [GAv][G]. .[G]. . .[gaV]	146	0.87	15.2
2. . [G][G]. . .π.π	100	0.86	14.8
3. [Yf].π. .[G].[LsD].	69	0.72	13.8
4. π. .π[G].[sDe][IYF]π	37	0.91	15.4
5. . . .[VIL]. .[G][gAS].	228	0.85	15.7
6. . . .[P][VIL]ϕ.[G].	74	0.85	14.3
7. .[AvL][Ark][π[ALy][G].π[VIL]	161	0.89	14.6
8. [ARK].[AVI][AL]ππ.[G][AVI]	167	0.95	15.0
(b) Conserved proline			
9. [VIL][P]. . . ππ. .	147	0.76	15.2
10. πϕ.[VIL][P]. . . .	101	0.72	14.3
11. . ϕ. .[Ats][P]. . [aVI]	152	0.81	16.1
12. [PLF].[GAV]ϕ[P]π. . .	54	0.75	12.1
(c) Conserved polar residues			
13. [VIL]. .[N]. .π. .	61	0.81	14.0
14. . .[D]. . .[iln][TSh]π	54	0.76	14.4
15. . .[VIL[[D]. .πππ	138	0.91	15.7
16. .[AVt]. .[D]. . .π	149	0.84	15.4
17. .[T]π. .[Avt]. . .	90	0.76	14.9
18. PP[qDe][LYF][vLF]XX[D]X	76	0.81	13.8
(d) Conserved non-polar residues			
19. [iLF].π[A]ϕ2ππ.π	136	0.89	14.6
20. . .π.[iLm]. .[A][ViL]	126	0.73	14.3
21. .[AtS][A]ϕ.[ALY][AVF][LmQ]π	69	0.76	14.6
22. [AvL]ππ[L].π.[vIL].	228	0.92	15.9
23. ππ[F]ππ.ππ.	93	0.84	14.4
(e) Conserved arginine/lysine			
24. . . .[Rk][LFw]. . .[IF]			
25. [ASD].[AVT]π.[RK].[vIL].	74	0.81	12.7
26. [ThR].[RK][LFK][VIL][VIL][Avl][AY].	63	0.81	12.2
27. .π[RK][gPA]π[Hde][AVI].[Avl]	35	0.79	13.1
(f) Threonine and serine			
28. .[aiT][PTS]ππ.π.[iLm]	59	0.70	14.1
29. [TSd]ππ[QDE]ϕ.π[AVL].	145	0.91	14.6
30. π[TSQ].[aTS].[GLN].ϕ[VIw]	26	0.62	13.0
31. .[iTS][TS]πππ[aTS][ATS]π	65	0.76	14.0
32. .[gAS]π[ATS]ϕ[iL][gAT]. .	95	0.76	15.2
33. [vwT].[tSq]π. .[vIL][Re]π	54	0.93	13.8
(g) Alternating hydrophobic-polar			
34. [VIL][TSd]ππ[QDE][AVIL]ππϕ	134	0.94	15.3
35. [aKD]π[aVI]ππ.[VIL]ππ	210	0.95	14.4
36. πϕ.[vIL][aVT][VIL][PAS].[VIL]	69	0.65	12.2
37. .[ViL]π[viL].π.[VIL].	63	0.78	15.0
38. [GND].π[VIL].[VIL]. .π	99	0.87	15.2
39. .πππ[VIL] . [VIL] .	122	0.84	15.0
40. [VIT]π[VIL][Vil].πππ	111	0.86	14.4
41. . .[Vlt].[aVI][VIT].[Pa]π	70	0.81	14.3
(h) Miscellaneous			
42. ϕ[GAs].ϕ. . . .[gAS]	29	0.76	14.2
43. [PVL][VIL][ViL][gAl][AVY].[PNH]. .	58	0.70	12.2
44. [VI]π.[YFW]. .[WTR]π.	43	0.77	14.5
45. πϕ.[GYs][NHR][PYF][iLF].[gAR]	33	0.89	11.2
46. ϕϕ.ϕϕϕϕ[avl]ϕ	148	0.39	11.7

Positions with variability indices less than six are described by amino acids in brackets (upper and lower cases are as in Table 1). The remaining positions are represented by ϕ (average hydrophobicity greater than 0.65), π (average hydrophobicity less than 0.35) or . (average hydrophobicity between 0.35 and 0.65).

and large (F) hydrophobic side-chains are conserved. The hydrophobicity patterns neighboring these conserved positions are in many cases quite distinctive.

The second category consists of patterns with highly conserved residues (Table 3B, sections (a) to (e)). Interestingly, only a subset of the amino acids are absolutely conserved in any of the patterns. Clusters with conserved glycine residues are particularly

common (20% of all clusters). Because of the conformational flexibility of glycine residues, these patterns may be advantageous in local structures with unusual backbone torsion angles. Several clusters have more than one conserved glycine. For example, pattern 2 (Table 3A) contains two consecutive conserved glycine residues, and pattern 1 has a GXXG motif. In pattern 6, there is a proline residue four positions prior to a conserved glycine, with preferences for hydrophobic residues in the two positions following the proline. In pattern 3 the aromatic residues Y and F are favored five residues prior to a conserved glycine. Other clusters containing conserved glycine residues and highly constrained neighboring substitution or hydrophobicity patterns are listed in Table 3B section (a).

Proline residues also have unique structural characteristics. Again, there are a number of patterns with conserved proline residues (Table 3B section (b)) and these have additional positions with distinctive substitution and hydrophobicity profiles.

Conserved charged amino acids may be involved in metal chelation, salt bridges, or catalysis. Interestingly, patterns with conserved charged residues often have strong preferences at additional positions. For example, patterns 14 and 18 contain conserved aspartic acid residues with strongly hydrophobic substitution patterns at different relative positions. In pattern 13, a position rich in V, I and L occurs three residues prior to a conserved asparagine.

The third category consists of patterns which have similar substitution patterns at all positions. For example, in Table 3A pattern 3 has preference for I, L and F, pattern 4 is glycine-rich and pattern 5 is dominated by T and S. Fairly strong structural constraints such as the requirement for flexibility may give rise to these repetitive patterns.

It is instructive to compare the patterns described in Table 3 to the patterns in the Prosite database. There are a number of key differences. First, patterns listed in Table 3 are common to multiple protein families: the proteins around which the different multiple sequence alignments in the starting dataset are based have less than 25% sequence identity. Families with particularly divergent sequences are represented several times (there are four globin chains and three immunoglobulin chains in the set), but since most of the clusters have of the order of 50 members, a particular pattern would have to occur in more than ten different places within a single protein for a single family to contain the majority of instances of the pattern. In contrast, Prosite patterns most often characterize single protein families. Second, the patterns in Table 3 contain no gaps (perhaps the major shortcoming of the current approach), while Prosite patterns can extend for substantially longer stretches. Third, the patterns are obtained in quite a different way. The patterns in Table 3 are generated completely automatically without any information other than the amino acid sequence, while the patterns of Prosite depend on the prior classification of sequences into functional or structural groups.

Primarily because of the first factor, there is not a large overlap between the patterns contained in the two sets. This reflects a more fundamental difference: the conserved patterns in Prosite reflect either functional constraints or quite specific structural constraints, while the patterns in Table 3 probably arise from more general structural constraints or properties common to many different classes of proteins.

Variation patterns and substitution matrices

It is interesting to compare the association of amino acids in clusters with conventional substitution matrices which estimate the cost of substituting one residue type for another. One of the most powerful current substitution matrices is the BLOSUM62 matrix which was generated from the Blocks database of multiple sequence alignments (Henikoff & Henikoff, 1992). The relationship between the BLOSUM substitution matrix and the clusters of Table 1 is simple: the value of a particular cross term in the substitution matrix is a function of the (weighted) average probability that two residues will be in the same cluster. It should be pointed out that our analysis relies on the alignments contained within the HSSP database, which were generated using a conventional substitution matrix (McLachlan, 1971).

There are instructive differences in the performance of the PAM(250) (Gonnet *et al*., 1992) based distance measure *d3* (see Methods) for single positions *versus* strings of contiguous positions. As shown in Figure 1(c), use of the PAM matrix in clustering of single positions gives results quite similar to those of the simple distance measure (1). However, for segment length nine, many of the patterns which contain highly conserved residues were not present when the PAM matrix was used and there were many fewer patterns overall (data not shown). The averaging involved in the use of a substitution matrix, although not detrimental for the patterns in Table 1, which in any event are averages over large numbers of different local contexts, results in considerable loss of sensitivity for comparisons between segments of contiguous positions.

It is clear from Table 3 that substitution patterns are position-dependent. There have been numerous proposals for grouping the 20 amino acids into smaller numbers of sets in order to make the analysis of sequence to structure mapping more manageable. One approach groups amino acids according to their similarity based on standard substitution matrices. For example, the sub-groupings (1) I, L, M, V; (2) F, Y; (3) H, K, R; (4) A, P, S, T; and (5) D, E, N, Q were derived from analysis of the Dayhoff substitution matrix (Risler *et al*., 1988). Mixed codes based on chemical properties of the amino acids have also been proposed (French & Robson, 1983); the suggested groupings were (1) L, M, I, V, F; (2) R, K, E, D; (3) Q, N, T, S. The wide variety of groupings shown in Table 3 suggests that any reduced code will have limited generality.

Discussion

We have described a completely automated approach to identifying recurring sequence motifs in protein families. The patterns identified here (see Tables 1 and 3) probably include most of the local motifs which transcend protein family boundaries for proteins of known structure. Because of the numerous factors which enter into the determination of protein structures, the data set is probably somewhat biased and there may well be additional patterns in the large number of protein families for which structures are not available.

The clustering procedure used here, although simple, appears to be quite adequate for modeling the data: the local covariances of residue occurrences found in multiple sequence alignments. First, the independence of the results from the choice of starting cluster centers required for the K-means algorithm attests to the numerical stability of the procedure. Second, the results are surprisingly robust to changes in the distance metric and sequence weighting schemes (Table 1 and Figure 2). Third, most of the patterns obtained for individual positions (Table 1) and many of the patterns obtained for segments of contiguous positions (Table 3) are consistent with expectation (the division between hydrophobic and polar patterns in Table 1 is perhaps the simplest example).

Our results permit limited but significant generalizations about the distribution of protein amino acid sequences in sequence space. The robustness of the results suggests that the majority of the patterns are reasonably well separated from one another. Furthermore, the distribution of sequences in protein families appears to be considerably more ''clumpy'' than random distributions. The clusters obtained for the real protein sequence data are consistently lower in dimensionality than those identified in applications of the same clustering procedure to random datasets (Figures 1 and 2, Tables 1 to 3).

The classification of positions into different clusters provides a simple yet potentially powerful means to abstract the information contained in multiple sequence alignments into a higher level representation. A multiple sequence alignment can be replaced by a sequence of cluster numbers with relatively little loss of information. The resulting higher level sequences can be subjected to much the same types of analysis as normal amino acid sequences in efforts to correlate sequence with structure (Rost & Sander, 1993).

Our results may have useful applications for sequence comparisons, in particular for the identification of distant homologs for newly determined sequences. It is well established that searches with profiles constructed from sets of aligned sequences are considerably more sensitive to distant homologs than searches with single sequences. The reason for this is simple: a sequence profile contains at each position family-specific information about the likelihood of different amino acids to substitute at that position, while a search with a single sequence typically uses the same global substitution matrix at each position. As mentioned in the Introduction, the use of a single substitution matrix may average out weak but important similarities, whereas our clusters are in fact strings of distinct substitution matrices. One can imagine using the clusters as ''generalization rules'' whereby the substitution matrices generated from the closest cluster or clusters to each segment of a query sequence are used for scoring sequence alignments.

A similar strategy may facilitate extrapolating from a small number of aligned sequences. The idea is that given a small sample of the variation possible at a given position, the closest clusters can be identified to predict the variation likely to be observed in new members of the same family. Generalization in this fashion may permit the power of profile-based searching to be employed with only a few examples from the sequence family (or perhaps from only one example).

One way to implement the strategy described in the previous paragraph would be to use the variation patterns of Table 3 to generate a rough profile or sets of profiles for new sequences which have no close relatives: for each segment of nine residues in the sequence, select the closest pattern (or a weighted average of nearby patterns) and build a profile by splicing together the variation patterns for the different segments. Next, search the database with this inferred profile. This procedure potentially circumvents the limitations associated with using the same substitution matrix at each position of a sequence. The method may also be useful for generalizing from a small number of aligned sequences, but once there are more than five to ten, the substitution patterns are probably better inferred directly from the aligned sequence set.

There are also potential applications to protein structure prediction. There is a significant correlation between the local structures adopted by members of a given cluster, although the extent of correlation varies from cluster to cluster. For example, more than 80% of the occurrences of the first two patterns in Table 3A in known protein structures are in α-helices. Intriguingly, the conserved charged residues in patterns 13, 15 and 16 in Table 3B are buried in more than 70% of the occurrences of the patterns. Pattern 7 in Table 3A is very similar to the Schellman helix C-terminal capping motif (Aurora *et al*., 1994) and as expected occurs frequently in helix caps. A more extensive analysis of the structural correlates of the sequence patterns will be presented elsewhere. The tracing of the structural correlates of sequence patterns is essentially the inverse of the more standard (and very powerful) procedure of tabulating the frequencies of occurrence of the 20 amino acids in different structural environments (Bowie *et al*., 1991; Chou *et al*., 1978).

Finally, we should note that the results described here are highly dependent on the quality of the starting multiple sequence alignments. As the amount of sequence data increases and multiple sequence alignment algorithms are improved, approaches

similar to the one described here should become increasingly powerful.

Methods

The data

Multiple sequence alignments for proteins of known structure were taken from a non-redundant subset (PDB select 25; Rost & Sander, 1993) of the HSSP database (Sander & Schneider, 1991). No two multiple sequence alignments in this subset have parent sequences with greater than 25% identity. Because of the wide degree of sequence variation in families such as the globins and the immunoglobulins, the PDB select 25 list does include more than one chain per family in several cases (there are four globin chains and three immunoglobulin chains, for example). To reduce the problems associated with small sample size, families with fewer than 20 members were excluded from the analysis. Insertions common to less than 20 members of larger families were also excluded (the HSSP database consists of global sequence alignments). The final data set included 154 protein families with an average of 98 sequences per family.

Distance measure

Cluster analysis requires a metric on the space to be clustered. An advantage of using multiple sequence alignments is that there is a natural choice of metrics: the difference in the frequency distributions. A particularly simple choice is the ''city block'' metric:

$$d1(i, j) = \sum_{k=1}^{20} |F(i, k) - F(j, k)| \quad (1)$$

where $d1(i, j)$ is the distance between frequency distributions i and j and $F(i, k)$ is the frequency of occurrence of the kth amino acid at position i, $\Sigma_{k=1}^{20} F(i, k) = 1$. A distance measure for comparing single positions can be readily generalized to treat strings of contiguous positions. The distance between one segment of a multiple sequence alignment and a second segment of the same length is conveniently defined to be the sum of the distances between each of the corresponding positions:

$$d_N(i, j) = \sum_{n=0}^{N-1} d(i + n, j + n) \quad (2)$$

where N is the length of the window, i and j are the starting positions of the first and second segments, and $d(i + n, j + n)$ is for example distance measure $d1$ above.

Cluster analysis

The data set consists of roughly 20,000 frequency distributions. Most clustering algorithms become extremely time consuming with data sets of over 1000 members. The K-means algorithm is one of the few that can be used with extremely large data sets. In brief, a set of K initial cluster centers are chosen at random and each datum point is assigned to the closest center. New cluster centers are then determined by taking the mean of all of the data points in each cluster, and each datum point is re-assigned to the closest center in another pass through the data set (Everitt, 1993). This simple iterative scheme of recalculating cluster means and re-assigning data points to clusters is repeated until no data points are moved from cluster to cluster.

For technical reasons, the city block metric is somewhat preferable to the Euclidean metric for clustering frequency distributions using the K-means algorithm. Viewed as vectors in a $20N$ dimensional space, the frequency distributions vary widely in absolute magnitude (for window length one, a position in which only one amino acid occurs is represented by a vector of length one, while a position in which all 20 amino acids occur with equal probabilities is represented by a vector with length $[20 \times (1/20)^2]^{1/2} = 0.22$). The Euclidean distance between a position in which ten of the amino acids occur with equal frequencies and a position in which the other ten amino acids occur with equal frequencies is 0.45, while the distance between two positions in which different residues are absolutely conserved is 1.4. The city block distance between the two sets of positions is the same (1.0) in both cases, a more satisfactory result since no residues are in common in either pair. To avoid the problems associated with the use of the Euclidean metric with variable magnitude frequency vectors, the frequency vectors can be normalized to unit magnitude. However, the updating procedure basic to the K-means algorithm also changes the absolute magnitude of the cluster centers. The latter can be kept fixed, but this requires a somewhat awkward renormalization step after each re-assignment of vectors to clusters in the K-means procedure.

Error measures

How is the extent of clustering best evaluated? An explicit example illustrates the difficulties with evaluating different clustering strategies in high dimensional spaces, and in particular with data of the type involved here. Consider a position which can tolerate either of two amino acids, for example valine and isoleucine. With a small and possibly biased sample, the frequency of occurrences of the two residues may range from 0.0 to 1.0; the constraint being that the variation is contained within a two-dimensional subspace of the entire 20-dimensional space (only valine and isoleucine are allowed). The maximum distance between two points in this subspace is the same as the maximum distance between two points in the entire 20-dimensional space (two in both cases). The mean distance of the members of a cluster from the cluster mean is clearly a poor measure of the dimensionality of the cluster.

Two statistics which have proved useful for capturing the distribution of points within a given cluster are D, the number of dimensions for which the cluster mean exceeds 0.02 (chosen empirically), and V, the average variance in these dimensions. D clearly indicates the dimensionality of the subspace in which the cluster lies, and V, the average spread within this subspace.

To assess the extent of clustering of the sequence data, parallel experiments were carried out on simulated data sets. To construct these sets, the frequency distributions for each of the 20 amino acids were evaluated and then used to generate randomized versions of the HSSP database. The statistics of the simulated data sets are essentially those of the HSSP database with all covariances between substitutions at particular positions or between nearby positions set to zero. For each weighting scheme, a separate simulated dataset was generated based on the amino acid frequency distributions of the corresponding weighted dataset. We note that the more standard procedure of randomization by shuffling does not apply here since we are not seeking family-specific patterns.

A single composite statistic, the product of the variances of the individual residue frequencies, is also given in

several of the Tables to facilitate comparison between different positions within the same cluster. This crude volume measure is normalized by division by the corresponding quantity for the whole data set:

$$Volume\ (l) = \frac{\prod_{k=1}^{20} \frac{1}{M_l} \sum_{j=1}^{M_l} |F_l(j,k) - \langle F_l(j,k)\rangle|}{\prod_{k=1}^{20} \frac{1}{S} \sum_{j=1}^{S} |F(j,k) - \langle F(j,k)\rangle|}$$

where $F_l(j, k)$ is the frequency of the kth amino acid in the jth distribution in cluster I, $\langle F_l(j, k)\rangle$ is the center of the Ith cluster, and $\langle F(j, k)\rangle$ is the center of the entire data set. M_l is the number of vectors (or distributions) in cluster I, and S, the number in the whole dataset. To reduce the effects of small sample size artifacts, 0.001 is added to the terms in the product in the numerator (again, the value of 0.001 was determined empirically).

Numerical stability, alternative distance measures and the K-means algorithm

A disadvantage of the K-means algorithm is that both the number of clusters and the starting cluster centers must be specified in advance. In practice, use of more than the natural number of groupings results in the subdivision of several of the larger clusters. This is easily recognized, and each pattern is shown only once in Tables 1 to 3. The numerical stability of the algorithm and the dependence of the results on the starting cluster centers were assessed by carrying out multiple independent calculations using different sets of starting centers. Only the recurrent clusters are reported in the Tables.

A potential disadvantage of distance measure *d1* (equation (1)) is that a difference in frequency of 0.1 is treated similarly regardless of whether the difference is between 0.7 and 0.6 or between 0.1 and 0.0. Because of lineage effects, the former is likely to be less informative than the latter. A simple exponential scaling was used to emphasize differences of the latter type:

$$d2(i, j) = \sum_{k=1}^{20} |\exp[-F(i, k)] - \exp[-F(j, k)]| \qquad (3)$$

The K-means algorithm implicitly assumes the clusters to be spherical. If several variables are highly correlated or have significantly different variances, clusters may resemble prolate ellipsoids more closely than spheres. Non-spherical clusters can be accommodated by calculating the within-cluster covariance matrix and using the generalized Mahalonobus distance given by equation (4) when assigning data points to clusters (Everitt, 1993):

$$d3(i, j) = [|\mathbf{F}_i - \mathbf{F}_j|]\mathbf{M}[|\mathbf{F}_i - \mathbf{F}_j|] \qquad (4)$$

where $\mathbf{F}_i = F(i, k)$ and $\mathbf{M}$ is the inverse of a covariance matrix.

If the number of dimensions is of the same order as the number of data points in individual clusters, the matrix inversion required is not possible. In this case the inverse of the covariance matrix can be approximated by inverting the diagonal elements (the variances) and setting off-diagonal elements to zero. The modified K-means method in this case leads to minimization of the effective volume of the clusters rather than the average within-cluster distances.

Distance measure *d3*, with $\mathbf{M}$ equal to an amino acid substitution matrix such as a PAM matrix, weights differences according to the likelihood of substitution of one residue type for another (Dayhoff *et al*., 1972). This is a simple generalization of the similarity measure used in comparing single sequences that are distantly related. This measure, essentially a return to the single substitution matrix approach mentioned in the Introduction, is clearly only useful in the limit of small numbers of sequences per family.

Acknowledgements

We thank H. Schneider for the HSSP database; D. A. Agard for encouragement and computational resources; N. Hunt for compiling the PDB/ϕ-ψ dataset; S. Henikoff, J. Henikoff, S. Pietrokovski, D. Yee, K. Zhang, N. Hunt, T. Defay, M. Robinson, D. Teller, S. Karlin, L. Brocchieri, and members of the Agard laboratory for critical reading of the manuscript. K.F.H. is supported by the Howard Hughes Medical Institute Predoctoral Fellowship. This work was partially supported by the National Science Foundation, Science and Technology Center Cooperative Agreement BIR-9214821 and young investigator awards to D.B. from the NSF and the Packard Foundation.

References

Altschul, S. F., Carroll, R. J. & Lipman, D. J. (1989). Weights for data related by a tree. *J. Mol. Biol.* **20**, 647–653.

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

Bairoch, A. & Bucher, P. (1994). PROSITE: recent developments. *Nucl. Acids. Res.* **22**, 3583–3589.

Bowie, J. U., Luthy, R. & Eisenberg, D. (1991). A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, **253**, 164–170.

Brown, M., Hughey, R., Krogh, A., Mian, I., Sjolander, K. & Haussler, D. (1993). Using Direchlet mixture priors to derive hidden Markov models for protein families. In *First International Conference on Intelligent Systems for Molecular Biology* (Hunter, L., Searls, D. & Shavit, J., eds), pp. 47–55, AAAI Press, Washington, DC.

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

Dayhoff, M. O., Eck, R. V. & Park, C. M. (1972). A model of evolutionary change in proteins. In *Atlas of Protein Sequence and Structure*. National Biomedical Research Foundation, Washington, DC.

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

French, S. & Robson, B. (1983). What is a conservative substitution? *J. Mol. Evol.* **19**, 171–175.

Gonnet, G., Cohen, M. & Benner, S. (1992). Exhaustive matching of the entire protein sequence database. *Science* **256**, 1443–1445.

Gonnet, G., Cohen, M. & Benner, S. (1994). Analysis of amino acid substitution during divergent evolution: the 400 by 400 dipeptide substitution matrix. *Biochem. Biophys. Res. Commun.* **199**, 496–498.

Gribskov, M., Luthy, R. & Eisenberg, D. (1990). Profile analysis. *Methods Enzymol.* **183**, 146–159.

Henikoff, S. & Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proc. Natl Acad. Sci. USA*, **89**, 10915–10919.

Johnson, M., Overington, J. & Blundell, T. (1993). Alignment and searching for common protein folds using a data bank of structural templates. *J. Mol. Biol.* **231**, 735–752.

McLachlan, A. D. (1971). Identification of common molecular subsequences. *J. Mol. Biol.* **61**, 409–424.

Risler, J. L., Delorme, H. D. & Henaut, A. (1988). Amino acid substitutions in structurally related proteins, a pattern recognition approach. *J. Mol. Biol.* **204**, 1019–1029.

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

Sander, C. & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins: Struct. Funct. Genet.* **9**, 56–68.

Sibbald, P. & Argos, P. (1990). Weighting aligned protein or nucleic acid sequences to correct for unequal representation. *J. Mol. Biol.* **216**, 813–818.

van Ooyen, A. & Hogeweg, P. (1990). Iterative character weighting based on mutation frequency: a new method for constructing phyletic trees. *J. Mol. Evol.* **31**, 330–342.

Vingron, M. & Sibbald, P. R. (1993). Weighting in Sequence Space: A comparison of methods in terms of generalized sequences. *Proc. Natl Acad. Sci. USA*, **90**, 8777–8781.

Edited by F. Cohen

(Received 3 March 1995; accepted in revised form 23 May 1995)