Definition of the tempo of sequence diversity across an alignment and automatic identification of sequence motifs: Application to protein homologous families and superfamilies

ALEX C.W. MAY

Division of Mathematical Biology, National Institute for Medical Research, The Ridgeway, London NW7 1AA, UK

(RECEIVED April 18, 2002; FINAL REVISION August 6, 2002; ACCEPTED September 23, 2002)

Abstract

It is often possible to identify sequence motifs that characterize a protein family in terms of its fold and/or function from aligned protein sequences. Such motifs can be used to search for new family members. Partitioning of sequence alignments into regions of similar amino acid variability is usually done by hand. Here, I present a completely automatic method for this purpose: one that is guaranteed to produce globally optimal solutions at all levels of partition granularity. The method is used to compare the tempo of sequence diversity across reliable three-dimensional (3D) structure-based alignments of 209 protein families (HOMSTRAD) and that for 69 superfamilies (CAMPASS). (The mean alignment length for HOMSTRAD and CAMPASS are very similar.) Surprisingly, the optimal segmentation distributions for the closely related proteins and distantly related ones are found to be very similar. Also, optimal segmentation identifies an unusual protein superfamily. Finally, protein 3D structure clues from the tempo of sequence diversity across alignments are examined. The method is general, and could be applied to any area of comparative biological sequence and 3D structure analysis where the constraint of the inherent linear organization of the data imposes an ordering on the set of objects to be clustered.

Keywords: Sequence motif; multiple sequence alignment; optimal segmentation; protein homologous family; protein superfamily

Importance of sequence alignment

Multiple alignment of molecular sequences/three-dimensional (3D) structures is a central method in molecular biology (for a recent review, see Smith 1999). By definition, sequence-invariant positions within a set of sequences can only be identified once the relationships within the set have been defined by multiple alignment. Sequence-invariant positions within meaningful multiple protein sequence alignments frequently correspond to residues conserved for func-

tion, for example, an enzyme active site or 3D structure, for example, disulphide bonds. (However, recent work demonstrates that “enzyme function places severe constraints on residue identities at positions showing evolutionary variability, and at exterior nonactive-site positions, in particular” [Axe 2000].) Furthermore, significant sequence identities between aligned proteins allow inference of homology, that is, a shared evolutionary origin. Despite divergence of amino acid sequence, homologous proteins share the same overall 3D structure (Chothia and Lesk 1986). This observation is the basis of comparative protein modeling (Blundell et al. 1987): it is possible to model usefully the 3D structure of a protein given significant sequence identity with protein(s) of known 3D structure. Usually, homologous proteins also share a common function. Again, this is helpful: the function of a protein encoded by a new se-

Requests for reprints to: Alex C.W. May, Division of Mathematical Biology, National Institute for Medical Research, The Ridgeway, Mill Hill, London NW7 1AA, UK; e-mail: amay@nimr.mrc.ac.uk; fax: 44 (0) 20 8816 2460.

Article and publication are at <http://www.proteinscience.org/cgi/doi/10.1110/ps.0211202>.

quence can be suggested by analogy with a similar, previously characterized protein. In fact, this is the most popular approach to genome sequence annotation for currently uncharacterized proteins.

Relationships between protein sequence and 3D structure

The common fold of homologous proteins is usually operationally defined as their common core. The core comprises the main secondary structure elements that share not only the same spatial arrangement but also are related by the same topology. Chothia and Lesk (1986) used rigid-body superposition to delineate core regions for 32 pairs of homologous 3D structures. Their analysis showed that the relationship between the divergence of sequence and 3D structure in the common cores of homologous proteins from eight different families could be best described by an exponential function. That is, the higher the sequence identity, the higher the 3D structure similarity within the core. Also, the higher the sequence identity, the higher the proportion of residues in the common core. These results reflect use of a distance-based definition of topologic equivalence by superposition. As homologous sequences diverge, corresponding secondary structure elements can undergo rigid-body shifts leading to increased deviations between them (May 1996). Thus, the proportion of residues in the common core will concomitantly decrease. Segments connecting secondary structure elements are known as loop regions. There are often changes in the 3D structures of homologous proteins outside helices and/or strands because loop regions accommodate most insertions and deletions (indels) (Johnson et al. 1996). Hence, different numbers of amino acids can be found for proteins descended from a common ancestor that were originally identical in both length and sequence. The relationship between loop regions and indels is often used to infer the location of loops from a multiple sequence alignment.

Earlier methods to partition sequence alignments

Clearly, the patterns of sequence variability across a meaningful multiple sequence alignment will contain clues as to the organization of the common protein fold and its function (for a recent example of the inferences that can be made on the basis of aligned protein sequences, see Heymann and Engel 2000). However, as far as I am aware, there are surprisingly few published methods for automatic identification of variable and conserved regions within a multiple sequence alignment. One of the first seems to be the approach of Herrmann et al. (1996), using algorithmic scale-space filtering. Stojanovic et al. (1999) describe and compare five methods for the purpose; nevertheless, all five suffer from the disadvantage of requiring the user to specify

arbitrary values for a series of parameters, e.g. minimum block length. Similarly, the procedure of Andre et al. (2001) has the same drawback: it is dependent upon four user-defined parameters. Not only is there just a handful of published methods, but also none of them seem to be widely used. It is intellectually unsatisfying then that such delineation is usually performed on the basis of visual inspection with its attendant problems of arbitrariness, lack of consistency and difficulty in "scale-up." Further, this is not just an academic problem: functional inference for a new protein is often carried out by means of searching for the existence of functional motifs (patterns) such as those in the PROSITE database (Hofmann et al. 1999). Similarly, protein motifs can represent 3D structural signatures (for instance, see Jonassen et al. 1999). Clearly, the construction of motifs is an important area. "Sliding window" analysis is often used to smooth sequence(s) into segments containing positions with similar characteristics. A classic example is the hydrophathy plot (for a review, see Doolittle 1986). Here the amino acid hydrophobicity at a position is replaced by the mean value within a window around that position (the hydrophathy index). By definition, the size of the window determines the extent of smoothing. This means that choice of window size is arbitrary, and the results of such an analysis are often highly dependent on this parameter. Another problem is choice of step size for moving the smoothing window. Further, use of a moving mean leads to blurring of edges between regions and gives results that are sensitive to outliers. Most importantly, perhaps, is the failure of sliding window analysis to be able to give an unambiguous answer to the question: what is the optimal segmentation of the sequence(s)? Segmentation is a key concept in protein 3D structure: grouping amino acids into secondary structure elements is a segmentation of the protein chain, for instance.

Results and Discussion

HOMSTRAD and CAMPASS optimal segmentation distributions are very similar

What is the variability of the tempo of sequence diversity across the alignments of homologous families? How does it compare to that observed for the superfamilies (Table 1)? Clearly, the HOMSTRAD and CAMPASS optimal segmentation distributions are very similar (linear correlation coefficient $r = 0.9554$, $P \leq 0.003042$). This is surprising, given that, by definition, the HOMSTRAD families comprise closely related proteins while, of course, the CAMPASS superfamilies consist of distantly related ones.

There is no linear relationship between the size (number of 3D structures) of the families (superfamilies) and the optimal number of segments (HOMSTRAD linear correlation coefficient $r = -0.04201$, $P \leq 0.5459$ (Fig. 1); CAMPASS linear correlation coefficient $r = 0.04516$, $P \leq 0.7125$ (Fig.

Table 1. Distribution of the optimal number of segments for the 209 HOMSTRAD protein families and 69 CAMPASS protein superfamilies

Optimal number of segments	HOMSTRAD frequency	HOMSTRAD %	HOMSTRAD mean ($\pm$ SD) CV of relative segment length ^a	CAMPASS frequency	CAMPASS %	CAMPASS mean ($\pm$ SD) CV or relative segment length ^a
2	87	41.6	84.6 (37.2)	22	31.9	103.5 (23.0)
3	92	44.0	90.7 (31.1)	35	50.7	99.5 (28.5)
4	20	9.6	91.9 (35.1)	4	5.8	99.1 (39.1)
5	6	2.9	91.7 (30.0)	3	4.3	96.0 (39.4)
6	1	0.5	99.94 (—)	4	5.8	91.2 (21.7)
7	3	1.4	89.2 (12.3)	1	1.4	129.38 (—)

The criterion for selection of optimal segment choice is given in Materials and Methods.

^a Mean ($\pm$ SD) coefficient of variation (CV) of relative segment length (normalized with respect to alignment length). The SD is an absolute measure of the spread of a univariate distribution. CV is a measure of relative spread: it is defined as the SD as a percent of the mean. By definition, mean relative segment length for N segments is $1/N$.

2); there is no significant difference between the values of r , $p \leq 0.537$).

Similarly, there is no linear association between alignment length and the optimal number of segments (HOMSTRAD linear correlation coefficient $r = -0.1021$, $P \leq 0.1415$ (Fig. 3); CAMPASS linear correlation coefficient $r = 0.1106$, $p \leq 0.3658$ (Fig. 4); there is no significant difference between the values of r , $P \leq 0.131$).

Optimal segmentation of a contrived sequence alignment and “jumbling” tests suggest that the HOMSTRAD and CAMPASS partition data are meaningful

A concern is that the similarity between the HOMSTRAD and CAMPASS optimal segmentation data (Table 1) may

reflect an inherent bias within the constrained classification method. However, analysis of the optimal partitioning of a contrived sequence alignment and “jumbling” tests suggest that the results here are meaningful.

Consider a pairwise alignment comprising 10 positions of alternating nonidentity and identity. Thus, the information-theoretical entropy (Shenkin et al. 1991) profile consists of alternating 1.00 and 0.00, respectively. According to the criterion for choice of optimal segmentation specified in Materials and Methods, the optimal number of partitions for this contrived sequence alignment is 10 (data not shown). So, it seems likely then that the HOMSTRAD and CAMPASS optimal segmentation data (Table 1) are meaningful.

The “jumbling” test is a standard approach to estimate the significance of the optimal alignment score for two protein sequences (for a review, see Doolittle 1986). The sequences

Fig. 1. Relationship between size (number of 3D structures) of HOMSTRAD family and optimal number of segments for the 209 HOMSTRAD protein families. Linear correlation coefficient $r = -0.04201$, $P \leq 0.5459$.

Fig. 2. Relationship between size (number of 3D structures) of CAMPASS superfamily and optimal number of segments for the 69 CAMPASS protein superfamilies. Linear correlation coefficient $r = 0.04516$, $P \leq 0.7125$.

are repeatedly randomly reordered (“jumbled”) and then aligned to generate a distribution of scores for the pair. The significance of the score of the real alignment can then be expressed in terms of the familiar Z-score, that is, how far (in units of the SD of the distribution) and in what direction the score of the real alignment is from the mean of the distribution. The “jumbling” test for optimal segmentation

of an alignment consists here of 100 random reorderings of the aligned position entropies and subsequent optimal segmentation thereof. The idea is to test whether there is an intrinsic algorithmic bias for the optimal number of segments for an alignment irrespective of the linear order of aligned positions. The linear association between optimal number of segments for the real alignment and mean

Fig. 3. Relationship between alignment length (number of aligned positions) of HOMSTRAD family and optimal number of segments for the 209 HOMSTRAD protein families. Linear correlation coefficient $r = -0.1021$, $P \leq 0.1415$.

Fig. 4. Relationship between alignment length (number of aligned positions) of CAMPASS superfamily and optimal number of segments for the 69 CAMPASS protein superfamilies. Linear correlation coefficient $r = 0.1106$, $P \leq 0.3658$.

optimal number of segments for the 100 “jumbled” alignments is poor (HOMSTRAD linear correlation coefficient $r = -0.0108$, $P \leq 0.8767$ (Fig. 5); CAMPASS linear correlation coefficient $r = 0.3033$, $P \leq 0.01134$ (Fig. 6); there is no significant difference between the values of r , $P \leq 0.022$). (The smallest mean optimal number of segments for the 100 “jumbled” alignments for a HOMSTRAD

family = 2.8, while that for a CAMPASS superfamily = 3.0. The largest mean optimal number of segments for the 100 “jumbled” alignments for a HOMSTRAD family = 3.5, while that for a CAMPASS superfamily = 3.6. Coefficient of variation (CV) is a measure of relative spread: it is defined as the SD as a percent of the mean. The smallest CV of optimal number of segments for the 100

Fig. 5. Relationship between optimal number of segments for real HOMSTRAD alignment and mean optimal number of segments for 100 “jumbled” alignments for the 209 HOMSTRAD protein families. Linear correlation coefficient $r = -0.0108$, $P \leq 0.8767$.

Fig. 6. Relationship between optimal number of segments for real CAMPASS alignment and mean optimal number of segments for 100 "jumbled" alignments for the 69 CAMPASS protein superfamilies. Linear correlation coefficient $r = 0.3033$, $P \leq 0.01134$.

"jumbled" alignments for a HOMSTRAD family = 17.2, while that for a CAMPASS superfamily = 20.0. The largest CV of optimal number of segments for the 100 "jumbled" alignments for a HOMSTRAD family = 70.6, while that for a CAMPASS superfamily = 83.3. The mean [$\pm$ SD] CV of optimal number of segments for the 100 "jumbled" alignments for a HOMSTRAD family = 35.1 [7.1], while that for a CAMPASS superfamily = 36.1 [11.9]. Clearly, there is usually a wide dispersion in the optimal number of segments for the 100 "jumbled" alignments for both a HOMSTRAD family and a CAMPASS superfamily.) The linear order of aligned positions is indeed found to be important for HOMSTRAD and CAMPASS, so optimal partitioning of their alignments (Table 1) is useful.

Optimal segmentation identifies an unusual protein superfamily

Although the HOMSTRAD and CAMPASS optimal segmentation distributions are very similar, there is a difference for the linear relationship between relative segment length (normalized with respect to number of aligned positions) and mean relative segment sequence variability (normalized with respect to maximum value possible) for a protein family (superfamily) after optimal segmentation. Complete positive correlation ($r = 1.0$) occurs when short segments have low sequence variability and long ones have high variability. Conversely, complete negative correlation ($r = -1.0$) holds when short segments have high sequence variability and long ones have low variability. For HOMSTRAD, 32.1% of

the families (67) have $r > 0.8$, while 29.7% (62) have $r < -0.8$. In contrast, 53.6% of the CAMPASS superfamilies (37) have $r > 0.8$, while only 1.5% (1) have $r < -0.8$. By definition, it is to be expected that protein superfamilies will show a greater tendency towards having short segments with low sequence variability and long ones with high variability after optimal segmentation than homologous families. The exceptional superfamily in this regard is nitrogenase iron protein-like ($N = 3$), where $r = -1.0$. Here, the optimal segmentation of the 437 aligned positions is into 2: an N-terminal partition of 98 aligned positions with mean relative variability = 0.88 and a C-terminal one of 339 aligned positions with mean relative variability = 0.70.

Protein 3D structure clues from the tempo of sequence diversity across alignments

Anfinsen and colleagues first demonstrated that the information needed to specify the 3D structure of a protein resides in its amino acid sequence (the principle of self-assembly; for a review, see Anfinsen 1973). This means that the tempo and organization of sequence variability across a meaningful multiple alignment will hold clues as to the nature of the shared 3D structure. On this basis, it seems useful to cluster protein families (superfamilies) with the same number of segments after optimal partitioning (Table 1) according to similarity in corresponding segment relative length, segment mean relative sequence variability and a measure of the spread of segment relative sequence variability, SD. Each family (superfamily) is represented as a

vector, the number of components of which is (optimal number of segments) $\times$ 3. In other words, the 3 above values represent each segment. The dissimilarity between a pair of families is defined as the Euclidean distance between the respective vectors. The most widely used algorithm for obtaining a hierarchical classification (Romesburg 1984), the unweighted pair-group method using arithmetic averages (UPGMA), is used here. As in May (1999a, 1999b), the support for the clusters defined by a tree is assessed with a jackknife test (Lanyon 1985). Thus, reliable internal nodes of a tree are identified (Table 2). Two questions can be asked about the meaningful family (superfamily) groupings. First, how many of them comprise families (superfamilies) of a common structural class? Second, how many of them consist of families (superfamilies) with a common fold? Class and fold assignments are according to SCOP (Murzin et al. 1995). Twenty of the 209 HOMSTRAD families comprise proteins consisting of more than one SCOP domain while all 69 CAMPASS superfamilies comprise single SCOP domains.

Forty-four of the 86 family clusters are reliable (Lanyon 1985) for the 87 HOMSTRAD families partitioned optimally into two segments (Table 2). Of these 44 groups, four comprise families of the same class. Of these four, three consist of a pair of families (two α and β proteins [α/β]; one all- β proteins); the other consists of three families (all- β proteins). (Twenty-two of the 87 families here are α and β proteins [α/β] while 12 are all- β proteins.) The single reliable family cluster with the same fold is made up of two families: immunoglobulin domain—C1 set: constant nonimmunoglobulin ($N = 5$; number of aligned positions = 102; vector = [0.24, 0.56, 0.21, 0.76, 0.44, 0.24]) and Cu/Zn superoxide dismutase ($N = 7$; number of aligned positions = 169; vector = [0.25, 0.57, 0.21, 0.75, 0.42, 0.26]). The Euclidean distance between the two vectors is 0.03. The common fold between these all β proteins is the immunoglobulin-like β -sandwich (4 of the 87 families here share this fold). The most typical sequence within a family can be

defined as that sequence accorded the lowest weight by a sequence weighting method (May 2001). (A standard weighting scheme—that of Henikoff and Henikoff, 1994—is used here.) A pairwise global sequence alignment between the most typical sequence for each family—1hsaa and 1cbja, respectively (protein names are PDB codes)—yields only 15.8% identity (data not shown). Clearly, there is only a distant sequence relationship between the two families.

The only other instance of a reliable family cluster comprising families with a common fold is that obtained after hierarchical classification of the 92 HOMSTRAD families partitioned optimally into three segments (Table 2). This group comprises two families: FMN-linked oxidoreductases ($N = 2$; number of aligned positions = 378; vector = [0.79, 0.66, 0.48, 0.08, 0.21, 0.41, 0.13, 0.73, 0.45]), and glycosyl hydrolases family 17 ($N = 2$; number of aligned positions = 310; vector = [0.74, 0.54, 0.50, 0.19, 0.24, 0.43, 0.08, 0.67, 0.48]). The Euclidean distance between the two vectors is 0.19. The shared fold between these α and β proteins (α/β) is the TIM β/α -barrel (4 of the 92 families here share this fold). By definition, a most typical sequence cannot be identified for a pair of sequences. However, the maximum identity obtained for the four interfamily pairwise global sequence alignments is only 17.3% (data not shown). Again then, there is only a remote sequence relationship between the two families.

Seventeen of the 21 superfamily clusters are reliable for the 22 CAMPASS superfamilies partitioned optimally into two segments (Table 2). Of these 17 groups, one consists of superfamilies of the same class (α and β proteins [α/β]; 10 of the 22 superfamilies here are α and β proteins [α/β]). The single reliable superfamily cluster with the same class comprises two superfamilies: flavodoxin-like ($N = 7$; number of aligned positions = 192) and periplasmic binding type I (domain 2) ($N = 6$; number of aligned positions = 196). Although the folds of the two superfamilies do not coincide

Table 2. UPGMA hierarchical classification of HOMSTRAD protein families (CAMPASS protein superfamilies) with the same number of segments after optimal partitioning (Table 1)

Optimal number of segments	Number of reliable HOMSTRAD family clusters ^a	Number of reliable HOMSTRAD family clusters with same class (fold) ^b	Number of reliable CAMPASS superfamily clusters ^a	Number of reliable CAMPASS superfamily clusters with same class (fold) ^b
2	44 (51.2)	4 (1)	17 (81.0)	1 (0)
3	37 (40.7)	6 (1)	22 (64.7)	2 (0)
4	10 (52.6)	1 (0)	2 (66.7)	1 (0)
5	4 (80.0)	0 (0)	— (—)	— (—)
6	— (—)	— (—)	3 (100.0)	1 (0)
7	— (—)	— (—)	— (—)	— (—)

Reliable clusters (internal nodes) within trees are identified according to Lanyon (1985); a tree for N objects comprises $N - 1$ internal nodes. The tree reliability of Lanyon (1985) can only be applied to hierarchical classifications of ≥ 4 objects.

^a Figures in brackets are number of reliable clusters as a percentage of all clusters for that optimal number of segments.

^b Class and fold assignments are according to SCOP (Murzin et al. 1995).

in SCOP, both are three-layer ($\alpha/\beta/\alpha$) sandwiches, that is, α -helices on both sides of a β -sheet. Both β -sheets are parallel, consisting of five (order 21345) and six strands (order 213456), respectively. A similar situation occurs for one of the two reliable superfamily clusters of the same class for the 35 CAMPASS superfamilies segmented optimally into three segments (Table 2): β -glucosyltransferase–glycogen phosphorylase ($N = 2$; number of aligned positions = 830) and cytidine deaminase domains ($N = 2$; number of aligned positions = 120). The β -glucosyltransferase–glycogen phosphorylase fold comprises two non-similar domains with three layers ($\alpha/\beta/\alpha$) each, while that for the cytidine deaminase domains is a single three-layer ($\alpha/\beta/\alpha$) sandwich (10 of the 35 superfamilies here are α and β proteins [α/β]).

A further test of the significance of the two HOMSTRAD reliable family clusters with a common fold is to combine the HOMSTRAD and CAMPASS data for each optimal number of segments (Table 1). One of the 22 superfamilies with optimal segmentation into two has the immunoglobulin-like β -sandwich fold: the immunoglobulins. Thus, the combined data for optimal number of segments = 2 comprises 109 alignments (Table 1), of which five have the immunoglobulin-like β -sandwich fold. Fifty-three of the 108 family/superfamily clusters here are reliable (Lanyon 1985) including that described above comprising two HOMSTRAD families sharing the immunoglobulin-like β -sandwich fold (data not shown). Two of the 35 superfamilies with optimal partitioning into three has the TIM β/α -barrel fold. Thus, the combined data for optimal number of segments = 3 consists of 127 alignments (Table 1), of which six have the TIM β/α -barrel fold. Fifty-four of the 126 family/superfamily clusters here are reliable (Lanyon 1985) including that described above consisting of two HOMSTRAD families sharing the TIM β/α -barrel fold (data not shown).

Conclusion

This article covers two issues. First, it introduces a new method to segment a sequence alignment into regions of similar aligned position variability so identifying motifs. Through its use of dynamic programming, the method is guaranteed to identify a global optimum for all levels of partition granularity. Furthermore, there is no need for a priori thresholds—a clear advantage of the approach over sliding window analysis. Second, it applies the method to compare the tempo of sequence diversity across reliable 3D structure-based alignments of 209 protein families (HOMSTRAD) and that for 69 superfamilies (CAMPASS). (The mean alignment length for HOMSTRAD and CAMPASS are very similar.) Surprisingly, the optimal segmentation distributions for the closely related proteins and

distantly related ones are found to be very similar (Table 1; linear correlation coefficient $r = 0.9554$, $P \leq 0.003042$). Further, optimal segmentation identifies an unusual protein superfamily: nitrogenase iron protein-like ($N = 3$). Here, the optimal segmentation of the 437 aligned positions is into two: a short N-terminal partition with high mean relative variability, and a C-terminal long one with lower mean relative variability.

Of course, it must be remembered that selection of members for protein families within HOMSTRAD and that for protein superfamilies within CAMPASS is by hand (for a discussion of this problem, see May 1999b). So, there is sampling bias due to human choice of proteins. Sequence weighting (for a review, see Durbin et al. 1998) could be used to address this problem. In fact, lack of sequence weighting might explain why there are only two reliable HOMSTRAD family clusters with the same class and same fold (Table 2). After all, a guiding principle of comparative sequence analysis is that it is possible to obtain protein 3D structure clues from the tempo of sequence diversity across alignments. (Similarly, it will be useful to apply this method to many more sequence alignments of multi-domain proteins to examine relationships between sequence and domain organization for different numbers of domains.) A related issue is the stability of the optimal segmentations (Table 1): a bootstrap or jackknife method could be used to assess this.

This article highlights the need to look further than a “headline” figure of percentage sequence identity and to ask the question: What is the underlying distribution of the identical amino acids within a given sequence alignment?

It is proposed that use of the automatic method described here to partition optimally sequence alignments will add value to comparative sequence analysis. Further, given a protein that undergoes a conformational change (e.g., that which might occur on ligand binding) and the availability of 3D structures of the relevant states, the approach here could be used to decompose the protein into elements that move essentially as rigid bodies during the conformational change. Here the data for optimal zonation would be the inter-CA atom distances after rigid-body superposition of the 3D structures. Of course, such an approach might also be useful to monitor the dynamics of a molecule during a computer simulation. Another application lies in the analysis of the mobility of the atoms within a protein 3D structure determined by X-ray crystallography. Regions of a protein molecule with low/high mobility are usually identified by eye from a plot of mean temperature factor for all atoms in each amino acid versus residue number. Automation of such an approach in this way would have obvious advantages. Evidently, there are many possible areas of comparative sequence and 3D structure analysis where this method could be helpful.

The results of this analysis are available on the Web at <http://mathbio.nimr.mrc.ac.uk/~amay>.

Materials and methods

A new method to partition sequence alignments

A protein sequence comprises amino acids related by a linear ordering. It is possible to make good use of this organization for the purpose of segmentation: it allows use of a simple dynamic programming algorithm (Hawkins and Merriam 1973; Bement and Waterman 1977; for a review of constrained classification, see Gordon 1996). In particular, use of dynamic programming means that such a clustering is globally optimal for all predefined numbers of subsets (Gordon 1996). (Use of the term “dynamic programming” here does not refer to the familiar biologic sequence alignment algorithms: it refers to a general algorithmic technique for optimization of which the alignment algorithms are a specific instance.) Also, an a priori threshold is not necessary to delineate clusters.

The problem is this: how to partition optimally the N aligned positions within a sequence alignment into all predefined numbers of subsets k where $1 \leq k \leq N$. A measure of the variability of a segment is the sum of squared deviations of the constituent points about their mean. Thus, a homogeneous segment has a low sum of squared deviations about its mean. Defining the total within-segment variability W by summation for k segments, the globally optimal partitioning into k segments is that which minimizes W . The Hawkins and Merriam (1973) method identifies all optimal sets of k segments covering all N points. For a segment comprising points $i, i+1, \dots, j$ let $r(i, j)$ denote the sum of squared deviations of these points about their mean. Let $F_j(m)$ be the within segment sum of squared deviations when an optimal j segment covering of point 1 to m is made. Of course, the 1 segment coverings are unique so:

$$F_1(m) = r(1, m) \quad m = 1, 2, \dots, N$$

This means that the optimal two-segment coverings can be deduced from the one-segment coverings. In turn, the optimal three-segment coverings can be deduced from the optimal two-segment coverings, etc. In other words, this is a recursive solution:

$$F_j(m) = \min_{j-1 \leq n < m} [F_{j-1}(n) + r(n+1, m)] \quad j = 2, 3, \dots, N$$

It is important to note that the above equation for $F_j(m)$ is slightly different from the original in Hawkins and Merriam (1973) where $1 \leq n \leq m$. Their formula is inconsistent with the description of their algorithm. In fact, Hawkins agrees, and has confirmed that the correct equation is as above (personal communication).

Thus, the value of W is deduced as $F_k(N)$. The boundaries of the optimal segments are obtained from a “trace-back”. Of course, the most appropriate value of k for an alignment is not known a priori. However, k can be inferred automatically in the following way. The improvement in W on addition of another segment is expressed in terms of “explained” variation: the percent reduction in W with respect to W when $k = 1$, that is, no segmentation. The optimal value of k is taken as that value of k that shows the largest increase in “explained” variation with respect to $k - 1$ (Everitt 1993). By this criterion, the optimal value of k for the alignment in Figure 7 is $k = 3$ (Table 3).

I have recently described use of this constrained classification algorithm for the optimal classification of protein sequences (May 2001). A measure of amino acid variability at an aligned position is needed. As in May (2001), the information-theoretical entropy (Shenkin et al. 1991) is used here to represent the variability of an aligned column of residues (Fig. 7, Table 3; note: the simplicity of

1	V	K	1.00
2	K	G	1.00
3	C	C	0.00
4	F	W	1.00
5	N	K	1.00
6	C	C	0.00
7	G	G	0.00
8	K	K	0.00
9	E	E	0.00
10	G	G	0.00
11	H	H	0.00
12	T	Q	1.00
13	A	M	1.00
14	R	K	1.00
15	N	D	1.00
16	C	C	0.00
17	R	T	1.00
18	-	E	1.00

Fig. 7. Sequence alignment of the HOMSTRAD (Mizuguchi et al. 1998) family with the shortest alignment length (number of aligned positions = 18) out of the 209: the zinc finger-CCHC-type family ($N = 2$). The first column denotes alignment position. The second and third columns contain the sequences of Incpn and Incpc, respectively (protein names are PDB codes.). The last column shows a measure of aligned position sequence variability: the information-theoretical entropy (Shenkin et al. 1991). Statistics for optimal segment choice of the alignment are given in Table 3.

the sequence alignment in Fig. 7 greatly facilitates the “working through” of the method given in Table 3). According to this formulation, the more diverse the position is, the higher the entropy. Thus, an aligned position comprising only a single residue type is accorded entropy 0. In May (2001), only multiply aligned positions, that is, those aligned positions without a gap in any sequence were considered. Here, because we are interested in the tempo of sequence diversity across an alignment, it seems sensible to take into account all aligned positions. Thus, the gap character is treated as an extra residue type.

Comparison of protein homologous families and superfamilies

Protein sequences are usually grouped into homologous families (i.e., proteins related by divergent evolution from a common ancestor) on the basis of “significant” sequence identity. For example, the SCOP classification (Murzin et al. 1995) defines family membership operationally on the basis of sequence identity $\geq 30\%$. Of course, homologous proteins also share a common fold and usually function. Superfamilies consist of proteins that despite having “low” sequence identities are probably distantly related by descent with modification. Similar 3D structure and function suggest common ancestry here. An advantage of sequence identity as a similarity measure between proteins is its transparency. For instance, % sequence similarity values are often quoted without stating the amino acid classification on which they are based. Sequence identity is unambiguous in that way. However, it must be remembered that sequence identity is only a count of residue identities in two sequences (cf. the Hamming distance), and does not say anything about the distribution of the matches as specified by an alignment.

Table 3. Optimal segmentation of the sequence alignment shown in Figure 7

Number of segments	W	"Explained" variation (%)	Improvement in W (%)	Segments
1	4.44	—	—	—
2	3.40	23.44	23.44	1–11; 12–18
3	1.66	62.71	39.27	1–5; 6–11; 12–18
4	1.47	67.00	4.29	1–5; 6–11; 12–15; 16–18
5	0.80	82.00	15.00	1–5; 6–11; 12–15; 16; 17–18
6	0.67	85.00	3.00	1–2; 3; 4–5; 6–11; 12–15; 16–18
7	0.00	100.00	15.00	1–2; 3; 4–5; 6–11; 12–15; 16; 17–18

For a description of the algorithm, see main text.

Reliable 3D structure-based sequence alignments: HOMSTRAD and CAMPASS

The Blundell group has compiled two databases of protein 3D structure alignments. The first, HOMSTRAD (Mizuguchi et al. 1998), is a database of aligned 3D structures of homologous proteins. The HOMSTRAD families comprise what the authors describe as "representative members." The second, CAMPASS (Sowdhamini et al. 1998), is a database of aligned 3D structures of superfamilies. An advantage of these databases is that not only are the sequence alignments based on 3D structure but also they are of very high quality. The availability of these alignments allows investigation of the tempo of sequence diversity across meaningful alignments of proteins related at different levels of sequence identity. Here I use the November 1999 full-release of HOMSTRAD (comprises 1193 proteins organized into 209 families; the mean [$\pm$ SD] number of structures per family is 5.7 [4.9]), and the currently available release of CAMPASS (consists of 288 protein domains grouped into 69 superfamilies; the mean [$\pm$ SD] number of structures per superfamily is 4.2 [2.9]). Interestingly, the HOMSTRAD and CAMPASS family (superfamily) size distributions are very similar (linear correlation coefficient $r = 0.9363$, $P \leq 0.0000001$; data not shown). The mean ($\pm$ SD) number of aligned positions for the 209 HOMSTRAD families is 249.7 (178.3), while that for the 69 CAMPASS superfamilies is 248.8 (160.5). One assumption made by the Student's *t*-test for a significant difference between two means is that the two distributions share the same variance. Clearly, the variances here are not identical and so the Student's *t*-test is not appropriate. A nonparametric alternative is the Wilcoxon signed rank test that, by definition, makes no assumptions about the two distributions. The Wilcoxon test tests two distributions for equality. According to the Wilcoxon test, there is not a significant difference between the HOMSTRAD and CAMPASS size distributions ($p \leq 0.6572$). Two hundred of the 209 HOMSTRAD families have a mean sequence identity of $>30\%$ (the mean [$\pm$ SD] mean sequence identity is 41.18% [9.51]) while all 69 CAMPASS superfamilies have a mean sequence identity of $<25\%$ (the mean [$\pm$ SD] mean sequence identity is 15.00 [3.21]). (Sequence identity is defined here over the multiply aligned positions only. This is because if a gap were considered as a different residue type then it is not clear how to treat matching gap characters in terms of sequence identity for a given pair; May 2001.) However defined, sequence identity does not indicate the nature of the gaps within an alignment. The ratio sequence length: alignment length reflects the relative number of gaps included in a sequence for alignment. As expected, although the mean numbers of aligned positions for HOMSTRAD and CAMPASS are very similar, there is a difference between the two in terms of gap density within the alignments. In general, the CAMPASS align-

ments contain more gaps: only 5.3% (11) of the 209 HOMSTRAD families have a mean value of the ratio sequence length:alignment length <0.7 while 33.3% (23) of the 69 CAMPASS superfamilies do.

Acknowledgments

I thank Allan D. Gordon and Douglas M. Hawkins for helpful discussions via e-mail. I am also grateful to colleagues at NIMR: Jaap Heringa and Willie Taylor for critical reading of the manuscript and Nigel Douglas for computational assistance.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

References

- Andre, C., Vincens, P., Boisvieux, J.F., and Hazout, S. 2001. MOSAIC: Segmenting multiple aligned DNA sequences. *Bioinformatics* **17**: 196–197.
- Anfinsen, C.B. 1973. Principles that govern the folding of protein chains. *Science* **181**: 223–230.
- Axe, D.D. 2000. Extreme functional sensitivity to conservative amino acid changes on enzyme exteriors. *J. Mol. Biol.* **301**: 585–595.
- Bement, T.R. and Waterman, M.S. 1977. Locating maximum variance segments in sequential data. *Math. Geol.* **9**: 55–61.
- Blundell, T.L., Sibanda, B.L., Sternberg, M.J.E., and Thornton, J.M. 1987. Knowledge-based prediction of protein structures and the design of novel molecules. *Nature* **326**: 347–352.
- Chothia, C. and Lesk, A.M. 1986. The relation between the divergence of sequence and structure in proteins. *EMBO J.* **5**: 823–826.
- Doolittle, R.F. 1986. *Of urfs and orfs: A primer on how to analyze derived amino acid sequences*. University Science Books, Mill Valley, CA.
- Durbin, R., Eddy, S.R., Krogh, A., and Mitchison, G. 1998. *Biological sequence analysis: Probabilistic models of proteins and nucleic acids*. Cambridge University Press, Cambridge, MA.
- Everitt, B. 1993. *Cluster analysis*. E. Arnold, London.
- Gordon, A.D. 1996. A survey of constrained classification. *Comput. Stat. Data Anal.* **21**: 17–29.
- Hawkins, D.M. and Merriam, D.F. 1973. Optimal zonation of digitized sequential data. *Math. Geol.* **5**: 389–395.
- Henikoff, S. and Henikoff, J.G. 1994. Position-based sequence weights. *J. Mol. Biol.* **243**: 574–578.
- Herrmann, G., Schon, A., Brack-Werner, R., and Werner, T. 1996. CONRAD: A method for identification of variable and conserved regions within proteins by scale-space filtering. *Comput. Appl. Biosci.* **12**: 197–203.
- Heymann, J.B. and Engel, A. 2000. Structural clues in the sequences of the aquaporins. *J. Mol. Biol.* **295**: 1039–1053.
- Hofmann, K., Bucher, P., Falquet, L., and Bairoch, A. 1999. The PROSITE database, its status in 1999. *Nucleic Acids Res.* **27**: 215–219.
- Johnson, M.S., May, A.C.W., Rodionov, M.A., and Overington, J.P. 1996. Discrimination of common protein folds: Application of protein structure to sequence/structure comparisons. *Methods Enzymol.* **266**: 575–598.

- Jonassen, I., Eidhammer, I., and Taylor, W.R. 1999. Discovery of local packing motifs in protein structures. *Proteins* **34**: 206–219.
- Lanyon, S.M. 1985. Detecting internal inconsistencies in distance data. *Syst. Zool.* **34**: 397–403.
- May, A.C.W. 1996. Pairwise iterative superposition of distantly related proteins and assessment of the significance of 3-D structural similarity. *Protein Eng.* **9**: 1093–1101.
- . 1999a. A cautionary note on interpretation of hierarchical classifications of protein folds. *Struct. Fold. Des.* **7**: R213.
- . 1999b. Toward more meaningful hierarchical classification of protein three-dimensional structures. *Proteins* **37**: 20–29.
- . 2001. Optimal classification of protein sequences and selection of representative sets from multiple alignments: Application to homologous families and lessons for structural genomics. *Protein Eng.* **14**: 209–217.
- Mizuguchi, K., Deane, C.M., Blundell, T.L., and Overington, J.P. 1998. HOMSTRAD: A database of protein structure alignments for homologous families. *Protein Sci.* **7**: 2469–2471.
- Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. 1995. SCOP: A structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.* **247**: 536–540.
- Romesburg, H.C. 1984. *Cluster analysis for researchers*. Lifetime Learning Publications, Belmont, CA.
- Shenkin, P.S., Erman, B., and Mastrandrea, L.D. 1991. Information-theoretical entropy as a measure of sequence variability. *Proteins* **11**: 297–313.
- Smith, T.F. 1999. The art of matchmaking: Sequence alignment methods and their structural implications. *Struct. Fold. Des.* **7**: R7–R12.
- Sowdhamini, R., Burke, D.F., Huang, J.F., Mizuguchi, K., Nagarajaram, H.A., Srinivasan, N., Steward, R.E., and Blundell, T.L. 1998. CAMPASS: A database of structurally aligned protein superfamilies. *Structure* **6**: 1087–1094.
- Stojanovic, N., Florea, L., Riemer, C., Gumucio, D., Slightom, J., Goodman, M., Miller, W., and Hardison, R. 1999. Comparison of five methods for finding conserved sequences in multiple alignments of gene regulatory regions. *Nucleic Acids Res.* **27**: 3899–3910.