*J. Mol. Biol.* (1997) **268**, 539–556

JMB

# Global Self-organization of All Known Protein Sequences Reveals Inherent Biological Signatures

**Michal Linial[1], Nathan Linial[2]*, Naftali Tishby[2] and Golan Yona[2]**

[1]*Department of Biological Chemistry, Institute of Life Sciences, Hebrew University Jerusalem, 91904, Israel*

[2]*Institute of Computer Science Hebrew University, Jerusalem 91904, Israel*

**Corresponding author*

A global classification of all currently known protein sequences is performed. Every protein sequence is partitioned into segments of 50 amino acid residues and a dynamic programming distance is calculated between each pair of segments. This space of segments is initially embedded into Euclidean space. The algorithm that we apply embeds every finite metric space into Euclidean space so that (1) the dimension of the host space is small, (2) the metric distortion is small. A novel self-organized, cross-validated clustering algorithm is then applied to the embedded space with Euclidean distances. We monitor the validity of our clustering by randomly splitting the data into two parts and performing an hierarchical clustering algorithm independently on each part. At every level of the hierarchy we cross-validate the clusters in one part with the clusters in the other. The resulting hierarchical tree of clusters offers a new representation of protein sequences and families, which compares favorably with the most updated classifications based on functional and structural data about proteins. Some of the known families clustered into well distinct clusters. Motifs and domains such as the zinc finger, EF hand, homeobox, EGF-like and others are automatically correctly identified, and relations between protein families are revealed by examining the splits along the tree. This clustering leads to a novel representation of protein families, from which functional biological kinship of protein families can be deduced, as demonstrated for the transporter family. Finally, we introduce a new concise representation for complete proteins that is very useful in presenting multiple alignments, and in searching for close relatives in the database. The self-organization method presented is very general and applies to any data with a consistent and computable measure of similarity between data items.



*Keywords:* sequence alignment; database searching; clustering; protein families; protein classification

## Introduction

Ongoing sequencing efforts have already uncovered the sequence of over 50,000 proteins, and the number is growing rapidly. These findings are complemented by many attempts to develop algorithmic/computational tools to analyze and organize the data. There has been considerable progress in the design of algorithms and software for pairwise sequence comparisons (Smith & Waterman, 1981; Lipman & Pearson, 1985; Altschul *et al*., 1990). On a larger scale, tools have been developed for comparisons that involve a small number of sequences (Gribskov *et al*., 1987; Taylor, 1990). Only a few computational studies have considered all, or many, of the known sequences. These studies focus on (1) searching for motifs, signature sequences and domains (Henikoff & Henikoff, 1991; Sheridan & Venkataraghavan, 1992; Harris *et al*., 1992; Sonnhammer & Kahn, 1994; Han & Baker, 1995; Hanke *et al*., 1996), (2) improving mutation matrices (Gonnet *et al*., 1992; Henikoff & Henikoff, 1992), (3) automatic classification of protein sequences into families (Wu *et al*., 1992, Ferran *et al*., 1994), (4) extraction of similarity relationships between protein sequences (van Heel *et al*., 1991; Watanabe & Otsuka, 1995).

Due to our limited understanding of the global organization of protein sequences, actual analyses are currently restricted to local considerations, based on pairwise ''distances'' among sequences. A new sequence is analyzed by extrapolating the properties of its ''neighbors''. From the perspective of computational learning theory, this is a naive

0022–2836/97/170539–18 $25.00/0/mb970948 

''nearest-neighbor classifier'' approach to modeling and to generalization from a model to new sequences (Cover & Hart, 1967).

We seek a globally consistent organization of the sequences that would reveal relationships among protein families and yield deeper insights into the nature of newly discovered sequences. By incorporating several recent developments in the theory of metric embedding, efficient graph algorithms, and unsupervised learning we could, for the first time, deal with the universe of all protein sequences. Here, we present a novel, computationally feasible method that yields a global model of the universe of protein sequences, and generalizes well to new sequences.

A metric derived from the Smith-Waterman (SW) dynamic programming measure of similarity (Smith & Waterman, 1981) turns the space of protein sequences into a finite metric space. Our results are based only on the metric properties of this space, incorporating no further biological information.

We begin by embedding the metric space in hand into Euclidean space, using the embedding algorithm described by Linial *et al*. (1995). Following the steps of this algorithm, we select at random, from the distribution defined in the above algorithm, certain sets of segments. Then each segment is associated with a vector whose components are the distances between the segment and the chosen subsets (where the distance between a segment and a subset is defined to be the minimum distance from segments in the subset). This representation maps the space of all segments to a Euclidean space (the embedding space) with small distortion (for more details, see Theory).

The embedded space is further analyzed and a statistical clustering model of the sequences is constructed. A key aspect of this stage is that the model's generalization power is closely monitored, so as to avoid the common pitfall of overfitting the noise of the original similarity measure. We monitor the model's generalization power by splitting the data into two random subsets and performing an hierarchical clustering algorithm independently on these two randomly chosen subsets of the data. At each level of the hierarchy we cross-validate the results by demanding that the clusters in the two sets perfectly agree (as explained in the next section). This clustering is hierarchical, and thus offers additional insight into the large-scale organization of the space of all protein sequences.

This clustering reveals significant biological signatures. Some families were clustered automatically into a few very specific and distinct clusters. Known motifs within proteins were automatically identified, and were clustered as well into distinct clusters. This tree of clusters provides some insights on relations between protein families, relations that are suggested by examining the splits along the tree. At a higher level of analysis we introduce new tools for representing and analyzing protein familes and their relations, as well as a new concise representation for protein sequences that is very effective in presenting multiple alignments for complex protein families, and can be used in searching for close relatives.

In the next two sections we introduce the theoretical foundation of our work, and the results of the clustering of all protein sequences, as well as their biological significance.

## Theory

A sufficiently large data set of proteins is an obvious prerequisite for a meaningful globally consistent organization of the sequences. The sheer volume of data makes such an undertaking very demanding in terms of computational complexity. There are many further obstacles on the way to self-organizing all protein sequences: (1) no efficient encoding is known for long sequences of amino acid residues; (2) standard measures for sequence similarity do not capture long-range features; (3) it is difficult to evaluate the quality and validity of models in this area, and in particular, their power to predict, or generalize beyond the available training data. Indeed, our results could not be achieved without incorporating several recent developments in the theory of metric embedding, efficient graph algorithms, and unsupervised learning.

### Metric embedding

A metric derived from the Smith-Waterman (SW) dynamic programming measure of similarity (Smith & Waterman, 1981) turns the space of protein sequences into a finite metric space†. Our underlying hypothesis is that the global structure of this metric space encodes much relevant information beyond what is revealed by local considerations that involve only specific pairwise comparisons.

A major tenet of this research is that in exploring metric spaces, there is much to be gained if the metric space under consideration is Euclidean. Consider the problem, encountered by most clustering methods, of selecting a typical representative of a point set. While this problem has no satisfactory general solution, in Euclidean spaces, the set's centroid is an obvious choice. Other basic Euclidean geometric constructs, such as directions are of great help as well. In this view, we begin by embedding the metric space in question into Euclidean space. A recently developed algorithm (Linial *et al*., 1995) yields embeddings where (1) distances in the Euclidean space are in good agreement with the original metric, and (2) the dimension of the host space is relatively small.

† While this measure may fail to satisfy the triangle inequality, such failures occur with frequency below $10^{-7}$, and hardly affect our results.

### Defining the metric space

Distance measures among sequences that are based on dynamic programming are overly sensitive to differences in lengths among proteins. Moreover, they may fail to account for multi-trait phenomena in proteins, since they are derived from local considerations. Therefore, we chose segments of 50 amino acid residues, and not complete proteins, as our basic building blocks. Protein sequences in SWISSPROT (Bairoch & Boeckman, 1992) release 30 (12/94), with fewer than 50 amino acid residues were eliminated. Each of the 38,106 remaining longer sequences was divided into segments of 50 residues with a 50% or higher overlap among consecutive segments, yielding a total of 544,000 segments. This partition into segments maps most functional and structural domains into one or two segments†, and the term segment always refers here to one of the above.

All pairwise similarity scores between segments were computed using Compugene's Bioccelerator (Compugen, 1996), which performs the SW dynamic programming algorithm (Smith & Waterman, 1981), with the Blosum62 mutation matrix (Henikoff & Henikoff 1992)‡. If $s(u, v)$ is the SW similarity score between the segments $u$, $v$, then their distance is defined *via* $d(u, v) = s(u, u) + s(v, v) - 2s(u, v)$.

### Euclidean embedding

To explain the mathematical background of our work, some definitions are in order first. Consider a finite metric space $(X, d)$ where $X$ has $n$ points. An embedding of $X$ associates a vector $f(x)$ with every point $x$ in $X$. The embedding $f$ is said to have distortion $\leqslant C$, if for every two points $x$ and $y$ from $X$:

$$C \parallel f(x) - f(y) \parallel \geq d(x, y) \geq \parallel f(x) - f(y) \parallel$$

A key result in the area of metric embedding was found by Bourgain (1985). According to this results, every metric space $(X, d)$ with $n$ points can be embedded in Euclidean space, so that the embedding has distortion at most $O(\log n)$. These ideas were further expanded by Linial *et al*. (1995), where they developed an algorithm that finds, for every such $X$, an embedding $f$ into Euclidean space. The two main features of this embedding are: (1) the host Euclidean space is only $O(\log^2 n)$-dimensional, (2) the distortion of $f$ is only $O(\log n)$.

The description of the algorithm from Linial *et al*. (1995) follows: we first select uniformly at random $\log^2 n$ subsets of $X$. Specifically: $\log n$ subsets of size 1, $\log n$ subsets of size 2, $\log n$ subsets of size 4, and so on, for every power of 2 until $\log n$ subsets of size $2^{\log(n)-1}$. Then, every element $u$ from $X$ is associated with a real vector of $\log^2 n$ coordinates. Each of these coordinates corresponds to one of these subsets. If $u$ is an element, and $S$ is one of the randomly selected subsets, than the value of the coordinate corresponding to $S$ in the vector associated with $u$ is $d(u, S)$, where $d(u, S)$ denotes the minimum of $d(u, v)$ over all elements $v$ in $S$.

This algorithm is randomized, and makes certain random choices. For any given $X$, almost all random choices lead to an embedding with the above features. In particular, the probability that the distortion exceeds $\log n$ is less than $1/n^2$. Moreover, if in a certain run the algorithm happens to generate an embedding with a worse distortion, then it may be rerun and the new run will amost surely yield an embedding with only a $O(\log n)$ distortion. For a rigorous analysis of this embedding, see Linial *et al*. (1995).

This Euclidean embedding algorithm was applied to the above metric space of all protein segments. Following the steps of this algorithm, we selected at random certain sets of segments. These sets are of varying sizes, as described above. Associated with every segment $u$ is, then, the $\log^2 n$-dimensional vector $(d(u, S))$ with $S$ ranging over all randomly selected sets. The Euclidean embedding of the collection of all segments maps every segment $u$ to its corresponding $\log^2 n$-dimensional vector.

## Cross-validated hierarchical clustering

Data clustering has been the method of choice for self-organizing point sets in Euclidean spaces for many years (Duda & Hart, 1973). Yet only recently has a clear distinction been made between the two different roles of clustering. When concise representation of data is sought (compression), data should be clustered so as to minimize certain global distortion measures, regardless of the actual meaning and significance of the cluster centroids (this procedure is often called Vector Quantization). In contrast, when clusters should serve as a reliable model for generalization, great care should be taken not to overfit the model to the randomness, noise and bias in sampling the training points. It has been a major goal of computational learning theory to provide conditions under which good generalization can be derived from small samples (e.g. see Kearns & Vazirani, 1994).

It is well known that overfitting to the training data can be avoided *via* cross-validation, i.e. testing the parameters of clusters against independent validation data. Generalization in high-dimensional spaces is a notoriously problematic task. A major difficulty is that representing an $n$-dimensional object to a desired precision may require a

† This choice of length is elaborated on in the Discussion.

‡ An alternative to the SW algorithm is the faster, but less accurate edit distance metric. The two disagree at about 20% of the cases, where SW yields improved alignments (with gaps) and scores that differ by factors up to 2. Fortunately, Compugene's Bioccelerator hardware (Compugen, 1996) made the SW algorithm computationally feasible.

sample set of size exponential in $n$ (''the curse of dimensionality''). It is important to understand that the drastic reduction in dimensions achieved by the embedding algorithm does not automatically guarantee the ability to properly generalize: while our embedding algorithm maintains, with small distortion, the distances among datum points, nothing is proven so far about other points from the original, high-dimensional distribution. In other words, there is no guarantee that the whole distribution is smoothly embedded in the lower-dimensional space. This would follow only from stronger assumptions on the sample set.

The validity of our clustering is thus monitored by splitting the data into two random subsets and the requirement that the clusters in the two sets perfectly agree at every level of the cluster hierarchy. By Vapnik's theory (Vapnik, 1982), this perfect correspondence between two independent samples implies a tight upper bound on the probability that these two independent cluster sets disagree on the classification of new independent points. Likewise, a bound is obtained on the generalization error of the model. Similar techniques were used by Pereira *et al*. (1993) for distributional clustering of English words, and for other studies in statistical modeling.

Our clustering approach resembles the familiar hierarchical vector quantization (VQ) algorithm (Gray *et al*., 1980; Gray, 1984): each datum point is associated with the nearest centroid, and then the centroids are re-estimated to minimize the distortion within each cluster. This process is repeated until convergence to a (usually local) minimum of the distortion. To reduce the dependency on initial conditions, the process begins with a single cluster. Subsequently, at each iteration, the cluster of highest aspect ratio is split†.

The model's generalization power is monitored by performing the algorithm on two randomly chose subsets of the data and by aborting every split on which the two processes ''disagree''‡ . This criterion is clearly very strict, and more relaxed criteria for the matching are examined as well.

† Singular value decomposition (SVD; Press *et al*., 1988) is applied to the covariance matrix of each current cluster. The cluster of largest SVD principal component gets split along the corresponding direction.

‡ Agreement entails a one-to-one correspondence between clusters of the first set and those of the second set, where corresponding clusters have (1) nearly equal sizes, (2) nearly identical centroids, and (3) similar singular value decompositions. A split that fails these criteria gets aborted, and a split is performed on the cluster of the next-longest principal axis. The process terminates when all attempted splits get aborted.

§ Although the PROSITE catalog is the most extensive classification of proteins in the SWISSPROT Data Bank, it should be noted that the number of functionally or structurally defined families in PROSITE is actually smaller, since some PROSITE patterns match motifs such as glycosylation sites, subcellular localization signals, etc.

This clustering protocol is computationally intensive and was performed using the MOSIX distributed system (Barak *et al*., 1995).

## Results

The above algorithms were applied to the space of 544,000 segments derived from those 38,106 proteins of the SWISSPROT databank, with 50 amino acid residues or more. The whole computational process was fully automatic, without any human interventions or biological consideration. On termination, when the cross-validation criteria allowed no further splitting, the process yielded a tree of 106 clusters. We feel safe to say that this constitutes a genuine ''self-organization'' of all those protein segments. All the biochemical, evolutionary and functional background that is used in this procedure is reduced to the definition of the SW (Smith & Waterman, 1981) measure of similarity.

In order to evaluate the quality of this clustering, we made various comparisons with a known partial classification of proteins, namely the protein families in PROSITE (release 12.1 October 1994 (Bairoch, 1993). This list of about 700 groups of related proteins comprises 46% of the proteins in the databank§. Henceforth, a ''family'' of proteins always refers to a class on this list and the nomenclature of PROSITE is adopted.

We will start by evaluating the hierarchical tree in terms of the family composition within the clusters. We will focus on several clusters that match interesting motifs, or suggest the existence of biological features common to different families. In the second part, we will incorporate the data obtained from the distribution of protein segments among the 106 clusters to create a new representation of families (referred to as fingerprints), which induces quantitative indices of similarity between protein families. In the last part, we will introduce a new method for representing full-length proteins, based on the order of segments within a protein, and the clusters into which these segments were classified. Thus, by incorporating the detailed information from our clustering, a natural measure of similarity emerges for complete proteins as well. Moreover, this new representation is highly effective in visualization of domains shared by a group of related proteins. The three levels of analysis rely on the initial tree created by our algorithms, using the properties of the tree such as the relative position of a cluster in the tree, size of clusters, geomtry of a cluster, and the Euclidean distance between the centroids of the clusters.

### Clusters of protein sequences

The tree of 106 clusters, generated by the hierarchical clustering algorithm, is shown in Figure 1. Inspection of the tree shows that while most clusters were generated by a series of splits, corresponding to a deeper level in the hierarchical tree,

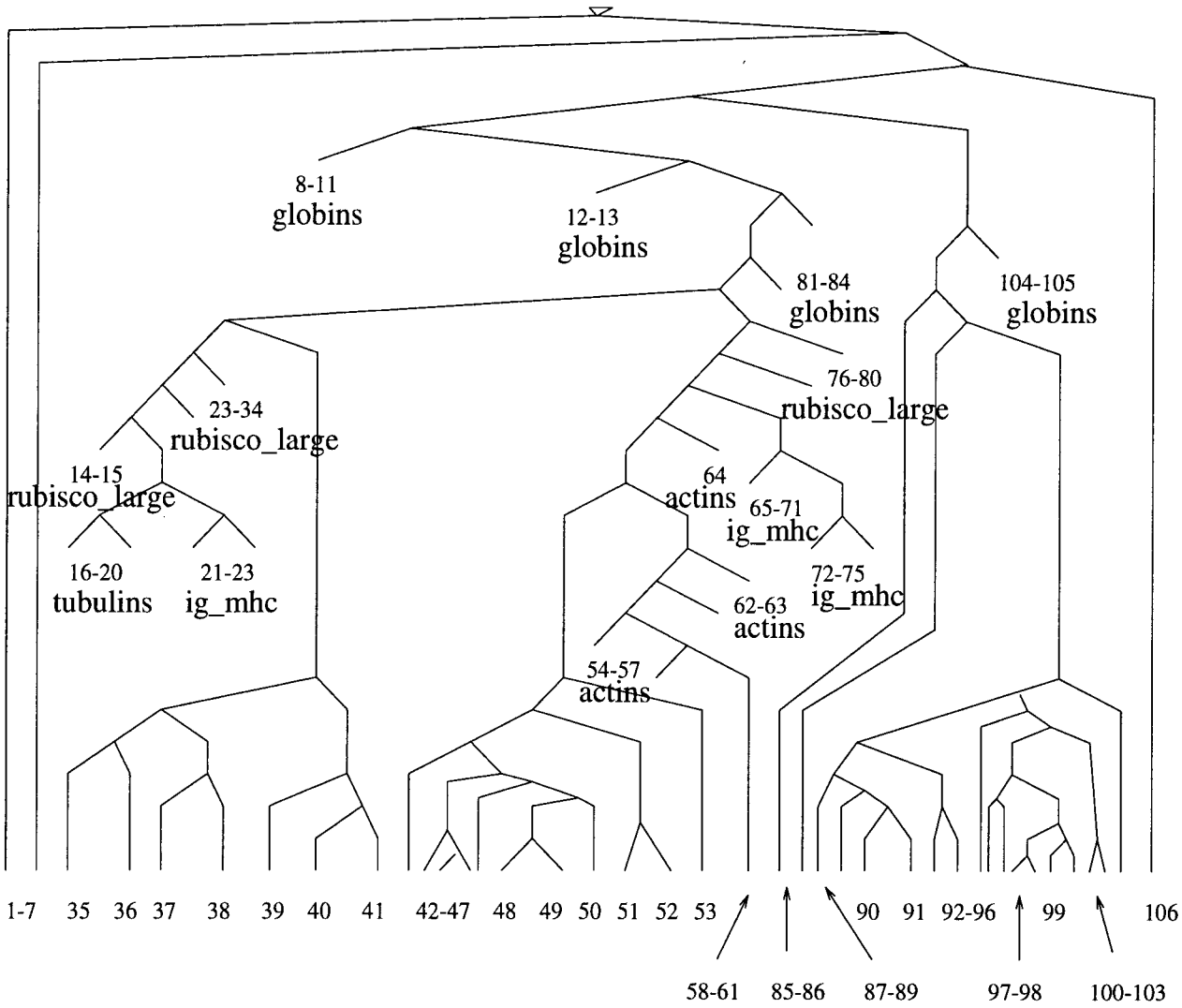


**Figure 1.** Hierarchical clustering of protein segments. Only major splits are shown, with the appropriate cluster numbers 1 to 106. Certain clusters, e.g. 8 to 13 are created already early in the process, but most clusters correspond to deeper, more involved series of splits. Some of the conserved families that split from the rest at early stages are shown. Subclassification within family (e.g. hemoglobin alpha chain, hemoglobin beta chain, myoglobin, etc.) is not indicated.

a substantial number of clusters were created and stabilized already after a few splits. The most extreme example is in the case of globins, which comprise clusters 8 to 13, 81 to 84, 104 and 105. All evolved very early during the clustering process.

Figure 2 offers a general view of the clusters' complexity and the distribution of data among the clusters. A cluster's complexity is measured by the number of PROSITE families that contribute at least one segment to it. About 56% of our clusters correspond to a single family and another 12% of the clusters are still of low complexity, with up to 20 families per cluster. At the high-complexity end, over 200 families appear in 22% of the clusters. On the other hand, the vast majority (90%) of the segments belong to highly complex clusters (over 200 families/cluster). Therefore, while most clusters are small and have low complexity, they comprise only a small fraction of the data. Several large clusters may need further splitting (see Discussion). However, despite their complex nature, some of the large clusters are very informative (see below).

Table 1 provides a closer view of the clusters, including their size and family composition. Many conserved familes get classified into a few distinct, low-complexity clusters. Such families include globins (clusters 8 to 13, 81 to 84, 104 and 105), ribonucleotide reductases (rubisco_large, clusters 14-15, 24 to 34, 76 to 80), immunoglobulins (ig_mhc, clusters 20 to 23, 65 to 74), actins (clusters 54 to 57, 62 to 64, 75 and 99) and tubulins (clusters 16 to 19, 48 to 51). Certain families have almost all their segments classified to low-complexity clusters. For example, 98% of the actin segments and 96% of the rubisco_large segments fell into such clusters. Other families, such as metallothionein, kazal

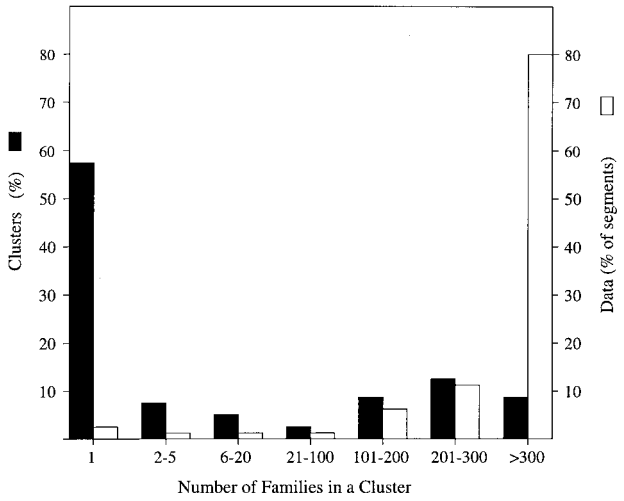


**Figure 2.** Distribution of clusters according to their complexity and distribution of data. Filled bars show the distribution of clusters by the number of families represented in them (''cluster complexity''). Open bars show the distribution of segments according to the complexity of the clusters containing them. For example, the left open bar indicates that about 2.5% of the segments are in clusters that represent only one family.

serine protease inhibitor (kazal), and phospholipase a2 (pa2_asp) have all, or almost all, their segments in only one or two clusters. These clustering patterns illustrate that our method is sensitive as well as selective.

Some of the families are composed of different subfamilies. In most cases our clustering method distinguishes between these subfamilies. For example, hemoglobin alpha chains were clustered to clusters 9, 10, 11, 81, 82 and 83, while hemoglobin beta chains were clustered to clusters 8, 13, 84, 104 and 105, and myoglobins to clusters 39 and 50. For clarity, we refer to all of these as globins. Such subclassification was resolved for other families as well. In the frame of this work we will not pursue this issue further.

Note that the number of segments in a cluster may differ from the number of proteins from which they are derived. A high ratio between these two parameters reflects the existence of repeats, or redundancy, in these proteins. For example, in cluster 35 this redundancy ratio is about 5.5. All the proteins that have segments in this cluster are classified as zinc finger proteins. This high ratio results from four to eight repetitions of the signature specific to zinc finger domains, all of which were clustered into cluster 35 (see below). Another example is cluster 88, where the segments to proteins ratio is even higher, about 13. Though only 15% of the proteins within these clusters have a PROSITE classification, all of those segments are repeated domains of structural proteins (mostly collagens).

It should be emphasized that the process that created the tree of 106 clusters (Figure 1 and Table 1) is fully automatic, and no biological consideration was made. Yet, the global organization reveals many clusters that correspond to significant biological patterns.

## Amino acid composition

The amino acid distribution was calculated for each of the 106 clusters. In certain clusters, the amino acid distribution hardly differs from their distribution over the whole data bank, while other clusters show marked variations. Both cases are observed in large as well as in small clusters (Figure 3(a) and (b), respectively). Certain pairs of clusters have similar amino acid distributions, although they represent distinct protein families (not shown). Likewise, differences in the distribution of amino acids account for certain clusters, but certainly not for all of them. Consequently, this distribution alone does not necessarily determine biological properties. Only a few clusters exhibit degenerate amino acid distributions. For example, in clusters 87 and 88 glycine and proline are relatively prevalent while all other amino acids are underrepresented (Figure 3(c)), reflecting the degeneracy of the proteins from which the relevant segments are derived (Wootton, 1994).

## Motifs and domains

Some clusters exclusively match well-defined motifs within proteins. That is, segments that correspond to a specific biological pattern were grouped together to form a well-separated and distinct cluster. Two specific examples are the zinc finger motif and the homeobox domain.

### *The zinc finger motif*

The zinc finger motif is found in many DNA-binding proteins (like transcription factors) in which the zinc finger is the DNA-binding domain, but also in certain proteins in which the role of the zinc finger is unknown (see Cukierman *et al.* (1995)). These proteins are characterized by 2 to 30 finger-like sub-structures, each centered around a zinc ion. Each finger is about 30 residues long, with only a few highly conserved amino acids within it.

Cluster 35 corresponds to this motif. All the segments classified to it belong to proteins from the zinc_finger_c2h2 family (one of the two major zinc finger families). Moreover, these segments are exactly the segments that contain the zinc finger pattern (as defined by PROSITE), thus corresponding to the zinc finger motifs in each protein (Figure 4(a)).

### *The homeobox domain*

The homeobox domain is a 60 amino acid residue polypeptide sequence, found in nuclear, DNA-binding proteins. This domain binds DNA through a helix-turn-helix structure. Proteins that contain the homeobox domain are likely to act as regulators of transcription. Cluster 41 in our classification matches this domain. Out of 304 proteins in the homeobox family, 194 are represented in this cluster. Note that the segments within this cluster are exactly those that contain the homeobox signature (Figure 4(b)). The motif was extracted from the complete proteins without any *a priori* information.

**Table 1.** Detailed description of clusters

| Cluster no. | No. of segments | No. of proteins | No. of proteins with PROSITE label | No. of families | Main families (PROSITE) |
|---|---|---|---|---|---|
| 1 | 5206 | 3419 | 1993 | 249 | |
| 2 | 67538 | 21897 | 9430 | 637 | |
| 3 | 55403 | 19560 | 8398 | 613 | |
| 4 | 38109 | 15286 | 6535 | 558 | |
| 5 | 5935 | 2827 | 1217 | 202 | |
| 6 | 19364 | 9256 | 4046 | 438 | |
| 7 | 74082 | 23147 | 9999 | 650 | |
| 8–13 | 915 | 464 | 464 | 1 | Globin |
| 14–15 | 380 | 198 | 198 | 1 | Rubisco_Large |
| 16–19 | 609 | 154 | 150 | 1 | Tubulin |
| 20–23 | 600 | 155 | 155 | 1 | Ig_Mhc |
| 24–34 | 2375 | 223 | 222 | 1 | Rubisco_Large |
| 35 | 895 | 167 | 159 | 1 | Zinc_Finger_C2h2 |
| 36 | 129 | 67 | 67 | 1 | Kazal |
| 37 | 189 | 97 | 96 | 1 | Cytochrome_C |
| 38 | 337 | 209 | 203 | 3 | Snake_Toxin, Tubulin, Metallothionein |
| 39 | 369 | 151 | 150 | 2 | Cytochrome_C, Globin |
| 40 | 269 | 121 | 119 | 1 | Cytochrome_B_Heme |
| 41 | 780 | 361 | 194 | 2 | Homeobox |
| 42 | 3047 | 2006 | 1299 | 134 | |
| 43 | 4119 | 2869 | 155 | 243 | |
| 44 | 3838 | 2508 | 1288 | 193 | |
| 45 | 4824 | 3289 | 1953 | 261 | |
| 46 | 4104 | 2769 | 1746 | 167 | |
| 47 | 4415 | 3179 | 1734 | 252 | |
| 48 | 538 | 291 | 208 | 4 | Tubulin, Insulin, Hsp20, Sasp_1 |
| 49 | 466 | 242 | 211 | 3 | Gapdh, Cyto_B_Heme, Cooper_Blue |
| 50 | 1202 | 611 | 356 | 10 | Tubulin, Globin, Cyto_B_Heme, Rnase_Pancreatic |
| 51 | 1170 | 617 | 345 | 8 | Tubulin, 2fe2s_Ferredoxin, Histones, Globin |
| 52 | 2380 | 1344 | 936 | 30 | |
| 53 | 10830 | 5336 | 2529 | 345 | |
| 54–56 | 500 | 103 | 103 | 1 | Actin |
| 57 | 230 | 126 | 125 | 2 | Actin, Ribosomal_S12 |
| 58 | 4739 | 3192 | 1592 | 238 | |
| 59 | 3278 | 2629 | 1197 | 216 | |
| 60 | 3364 | 2434 | 1393 | 186 | |
| 61 | 3285 | 2415 | 1498 | 167 | |
| 62 | 172 | 101 | 100 | 1 | Actin |
| 63 | 155 | 128 | 127 | 2 | Actin, Histones_H3_2 |
| 64 | 208 | 105 | 103 | 1 | Actin |
| 65–74 | 1479 | 156 | 156 | 1 | Ig_Mhc |
| 75 | 201 | 102 | 102 | 1 | Actin |
| 76–80 | 1104 | 247 | 222 | 1 | Rubisco_Large |
| 81–84 | 866 | 445 | 445 | 1 | Globin |
| 82 | 205 | 205 | 205 | 1 | |
| 83 | 228 | 228 | 228 | 1 | |
| 84 | 217 | 217 | 217 | 1 | |
| 85 | 57893 | 20694 | 9231 | 637 | |
| 86 | 34771 | 14752 | 6685 | 581 | |
| 87 | 497 | 106 | 16 | 7 | Collagens |
| 88 | 1363 | 103 | 16 | 3 | Collagens, C_Type_Lectin |
| 89 | 4908 | 3497 | 2048 | 243 | |
| 90 | 301 | 105 | 105 | 1 | Pa2_Asp |
| 91 | 2325 | 1373 | 899 | 64 | |
| 92 | 4124 | 2814 | 1502 | 189 | |
| 93 | 3592 | 2486 | 1451 | 185 | |
| 94 | 4517 | 3048 | 1691 | 238 | |
| 95 | 3498 | 2328 | 1260 | 203 | |

**Table 1**—*Continued*

| Cluster no. | No. of segments | No. of proteins | No. of proteins with PROSITE label | No. of familes | Main families (PROSITE) |
|---|---|---|---|---|---|
| 96 | 3908 | 2735 | 1621 | 205 | |
| 97 | 205 | 101 | 85 | 1 | Cytochrome_B_Heme |
| 98 | 140 | 61 | 53 | 1 | Lactalbumin_Lysozyme |
| 99 | 513 | 423 | 278 | 9 | Actin, Cox2, Cyto_B_Heme, Chaperonins_Cpn10 |
| 100 | 3984 | 2758 | 1541 | 216 | |
| 101 | 3305 | 2287 | 1236 | 199 | |
| 102 | 2357 | 1537 | 985 | 109 | |
| 103 | 4530 | 2963 | 1779 | 244 | |
| 104–105 | 453 | 230 | 230 | 1 | Globin |
| 106 | 72461 | 23270 | 10250 | 652 | |

Each cluster is specified by its number (first column), the number of segments within it (second column) and the number of distinct proteins from which these segments originate (third column). The other three columns (partially) characterize clusters in terms of the PROSITE classification of the proteins. The fourth column gives the number of proteins that have a PROSITE label. The complexity of the cluster, i.e. the number of families that contribute these proteins, and major representative families are in columns 5 and 6, respectively. Notes: (1) A protein that contributes a segment to some cluster is considered a ''member'' in this cluster. (2) A ''family'' of proteins is always one of the classes in the PROSITE list, and the PROSITE nomenclature is adhered to. Only 46% of the proteins are classified in PROSITE. Multi-trait behaviors of proteins are not accounted for. For family definition and biological significance, refer to the PROSITE dictionary. (3) Where consecutive clusters represent only one and the same family, these are presented in a single record. (4) The number of segments in a cluster may differ from the number of proteins from which they are derived. A high ratio between these two parameters reflects the existence of repeats, or redundancy, in these proteins (see clusters 35 and 88). (5) Some families have almost all their segments in well-characterized, low-complexity clusters. Families with over 50% of their segments found in low-complexity clusters are underlined. (6) Subdivision within families was resolved but is not indicated.

## Clusters that match heterogeneous biological signatures

Some clusters represent biological signatures that are more heterogeneous but still very distinctive. These clusters are of medium size (1000 to 6000 segments) and medium complexity (about 10 to 200 families represented in each cluster). Some of them suggest a possible relation between the contributing families. In other instances, a finer resolution might be attained through further splitting. Two such examples are cluster 52 and cluster 5, each of which predominantly represents one known family.

### *Cluster 52; the EF hand motif and its relatives*

There are 2380 segments in cluster 52, originating from 30 different families. Despite this relatively large number of families, the amino acid composition in this cluster deviates from the overall values (Figure 5(a)). Predominant in this cluster is the family of proteins containing the EF hand motif. The EF hand is a short domain of about 30 amino acid residues that coordinates calcium ions. The motif is present in parvalbumin, calmodulin, troponin-c and others, all of which are involved in $Ca^{2+}$ signaling, thus regulating cellular activities. These proteins often carry several EF domains. Segments that correspond to these domains are classified to cluster 52 (see Figure 5(b)). In such domains, glutamic and aspartic acids prevail (they participate in coordinating the calcium ions) and indeed these amino acids occur in cluster 52 above the average (Figure 5(a)). In addition, the frequency of amino acids that are absent from classical EF hands, such as proline and cysteine, is low.

### *Cluster 5; the EGF domain*

Cluster 5 has about 6000 segments, which come from over 200 families. It has a distinct amino acids distribution (Figure 6(a)), where the representation of hydrophobic amino acids is low and histidine, proline, tryptophan and cysteine abound. Of 1217 proteins that are classified by PROSITE and contribute to this cluster, 8% are EGF-like proteins. However, these proteins contribute more than 30% of the corresponding segments, and are thus the predominant family. The EGF (epidermal growth factor) domain is a small polypeptide chain of 53 amino acid residues. The EGF domain includes six cysteine residues, which have been shown (in EGF) to be involved in disulfide bonds. This amino acid is highly abundant in this cluster (over three times the overall frequency, see Figure 6(a)). In proteins containing an EGF-like domain, only the EGF-like domains are classified to cluster 5, yet, each protein contributes about ten segments on average, in accord with the repeated nature of this domain (see Figure 6(b)).

Other families that are represented in this cluster are protein kinase, c-type lectin, homeobox, chitin binding, kringle, wnt1, and more. In most of them cysteine abounds and is involved in disulfide bonds. On the other hand, families of proteins rich in cysteine are not necessarily classified to this cluster, e.g. kazal proteases, snake toxins, etc. (for details see Table 1).

## Kinship of protein families as inferred from the clustering tree

Further information can be extracted from the tree of 106 clusters, by examining the splits as they

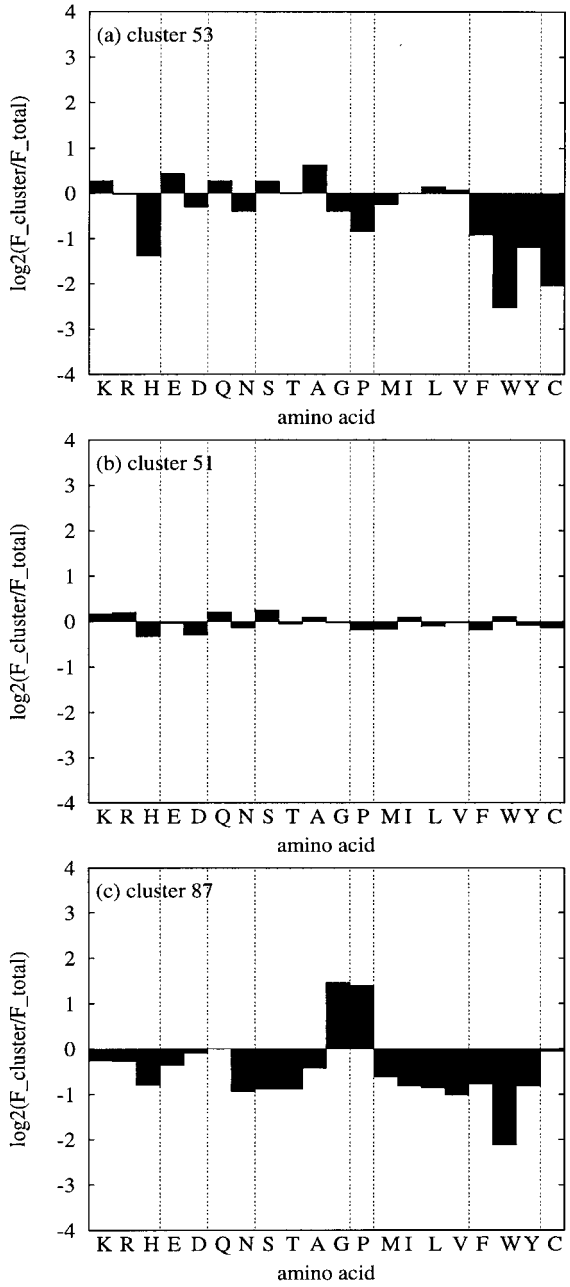


**Figure 3.** Amino acid distribution in selected clusters. Amino acids are marked by the single-letter code, and are grouped into biochemically related groups, separated by broken vertical lines. From left to right: amino acid residues that are basic (K,R,H), acidic (E,D), polar and uncharged(Q,N), small (S,T,A,G), proline (P), nonpolar hydrophobic (M,I,L,V), aromatic (F,W,Y), and cysteine (C). The variance in frequency is quantified as the logarithm of the frequency of a given amino acid within a cluster divided by its frequency throughout the entire databank (no difference between the expected value and the observed yields zero on this logarithmic scale). (a) Many of the clusters display a unique amino acid distribution. One example is cluster 53 with 10,830 segments. It is relatively rich in glutamine (Q), glutamic acid (E) and alanine (A) and is underrepresented in all aromatic residues (F,W,Y), histidine (H), proline (P) and cysteine (C). (b) Some other clusters show a smooth distribution close to the overall amino acid distribution.

occur along the tree. As clusters that split from the same root cluster may be biologically related, unknown relations between families may be revealed by examining the "evolutionary" process in the final tree. Likewise, clusters that represent a small number of families may hint at a connection between the families that they represent (e.g. clusters 48 to 51, 99 and more). We focus on only two cluster groups. Yet, other possible connections extracted from the junctions in the tree are open to interpretations, and may require further experimental data.

### *Cytochromes and globins*

Clusters 39 and 40 (together with cluster 41) split from their ancestor cluster quite late in our clustering process (Figure 1). Cluster 40 totally matches the cytochrome *b/b*6 family (cytochrome *_b_*heme), while cluster 39 is composed of cytochrome *c* (45% of the segments) and globins (54.5% of the segments), mostly myoglobins. Obviously, cytochromes of the two types are related, but the connection between globins and cytochromes is more interesting and suggests an intrinsic link. Indeed, an evolutionary relation between globins and cytochromes was recently proposed (Hardison, 1996).

### *Metal and DNA-binding proteins*

Clusters 39 and 40 are only part of a more complex structure. Figure 1 and Table 1 suggest an interesting and complex relation that ties clusters 35 to 41. The most common feature is that almost all families represented in those clusters bind metal ions (zinc finger, cytochrome *c*, metallothionein, cytochrome *b/b*6 and globins), or heme (cytochrome *b/b*6, globins) or DNA (homeobox, zinc finger). These families differ in their biological role (enzymes, transcription factors, etc.). Some of them use cysteine to stabilize their 3-D structure, e.g. zinc finger, snake toxins and kazal proteases. The high frequency of cysteine in those families is reflected in the amino acid composition of these clusters, but does not account for all of them (compare Figure 7(a), (b) and (d) with (c). Note that other clusters that are rich in cysteine (e.g. cluster 5, Figure 6(a)) are not part of this super-structure. Thus, no simple relation of amino acids composition ties all these clusters together. Rather, the

This is, for example, the case with cluster 51, even though it has only 1170 segments and represents only a small number of families. Note that the cluster size does not indicate the amino acid distribution profile (compare (a) and (b)). (c) Few of the clusters (e.g. cluster 87) consist of segments with very low compositional complexity (predominantly G and P). Most segments in this cluster are part of proteins that play a structural role (see Table 1) and have numerous repetitions.

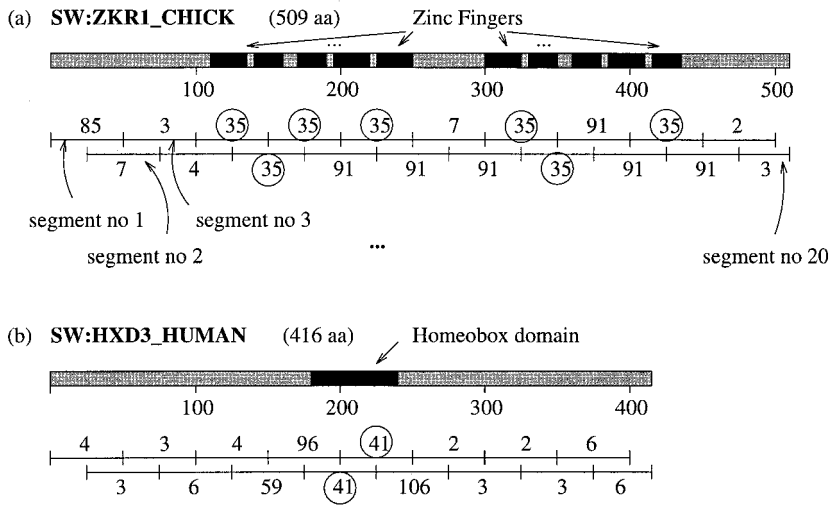


**Figure 4.** (a) The zinc finger motif. Cluster 35 matches the zinc finger domain. All segments in this cluster are part of proteins that are classified as zinc_finger_c2h2 according to PROSITE. Moreover, these segments match exactly the zinc finger domains in each such protein. Out of 241 proteins in this family, 159 contribute at least one segment to this cluster. One such example is sw:zkr1_chick (509 amino acid residues). In the representation, the zinc finger domains are denoted by filled boxes. The segment boundaries are indicated below. The number near each segment is the number of the cluster to which this segment was classified. The ten zinc fingers in this protein are divided roughly into two blocks. Note that only the zinc finger domains are classified to cluster 35 (circled). Part of the second block was classified to cluster 91, which is also rich with zinc finger proteins. (b) The homeobox domain. Cluster 41 matches the homeobox domain. Most segments classified to this cluster correspond to the homeobox domain in 193 different proteins. One of them is sw:hxd3_human (416 amino acid residues). Cluster 41 contains exclusively the homeobox signature (marked in black).

connection is complex and leaves some open questions.

### "Fingerprints" of biological families based on cluster membership

It is not easy to characterize biological families, say by a single consensus sequence or pattern. Consequently, most families are very diverse and populate many of our clusters. Therefore, the nature of a family cannot be deduced by inspection of a specific cluster. However, the distribution of segments from proteins in a family among the various clusters is more revealing. This broader view leads to an interesting novel representation of families, that distinguishes well different families. For example, families such as globin and gapdh (glyceraldehyde-3-phosphate dehydrogenase) exhibit a complex, yet well-defined distribution over clusters (Figure 8(a) and (b)). The distribution of segments from a family among clusters can be viewed as a fingerprint of the family. The statistical significance of this representation is guaranteed, again, by the cross-validation in the clustering procedure. Thus, not only membership in small clusters is informative. Membership in large and complex clusters may play a significant role in characterizing biological families.

Fingerprints of protein families allow quantitative comparisons among families: pick any distance measure among probability distributions, e.g. KL-divergence (Cover & Thomas, 1991) or variational distance. The similarity between two protein families is quantified by the distance of their fingerprints. In this way we can find, for each family, its proximal families.

It should be noted that the kinship of protein families, which is directly inferred from the tree structure (as was suggested for the globins and the cytochromes in the last section), is based on a local common motif, while the new representation reflects a global nature of all domains within a family, and suggests a more thorough kind of similarity, which projects to the biological function of the family.

The power of this method is demonstrated on several families of membrane proteins and transporters, whose mutual distances turn out to be the smallest (Figure 9). The four families (three transporter subfamilies, and a family of membrane proteins) share almost the same fingerprint, an evidence for the close biological function they all serve. Other transporters (e.g. antiporters†) and ion channels (e.g. neurotransmitter-gated ion-channels) resemble the fingerprints of the four families mentioned above to varying degrees. Thus, a connection is established among superfamilies within many of the membranous proteins. Fingerprints can be further analyzed by considering subfamilies and their fingerprints, as well as by inspecting superfamilies (unpublished results).

### Higher-level measures of similarity between sequences

Fingerprints capture the distribution of segments in a family among the different clusters, but fail to account for the order of segments within proteins. Significant information can be extracted for full-length proteins as well, by mapping each protein

† This family is not part of PROSITE list: 30 proteins in SWISSPROT that are defined as antiporters were grouped, and the corresponding fingerprint was created, based on the distribution of the proteins' segments among the clusters.

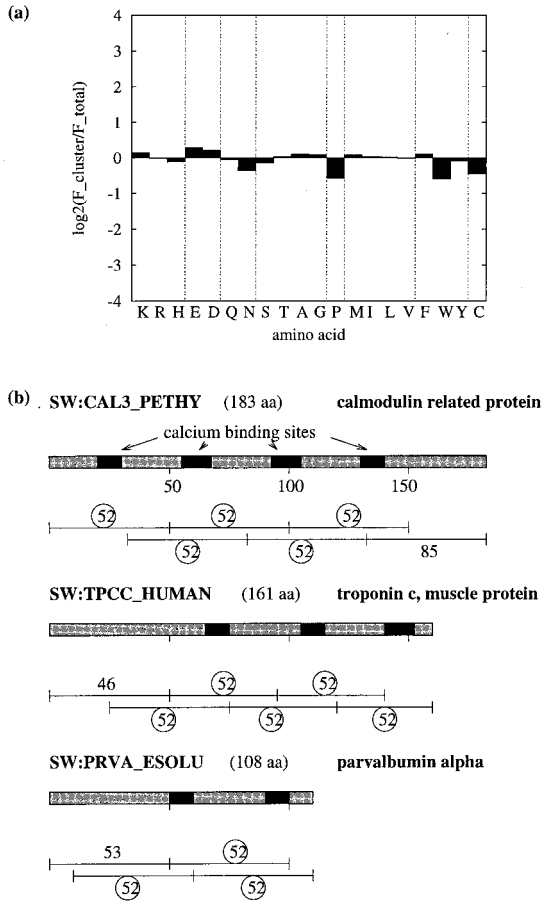


**Figure 5.** (a) Amino acid distribution in cluster 52. Acidic amino acid residues (E and D) are more frequent than their average frequency over all the database, while N, P, W and C are underrepresented. For details on the representation see Figure 3. (b) Cluster 52, the EF hand domain. The predominant family in cluster 52 is the family of proteins containing the EF hand motif. The motif is present in parvalbumin, calmodulin and troponin-c (one protein is shown from each subfamily). These proteins often carry several EF domains (denoted by filled boxes). Note that all the segments that correspond to these domains are classified to this cluster. For details on the representation, see Figure 4.

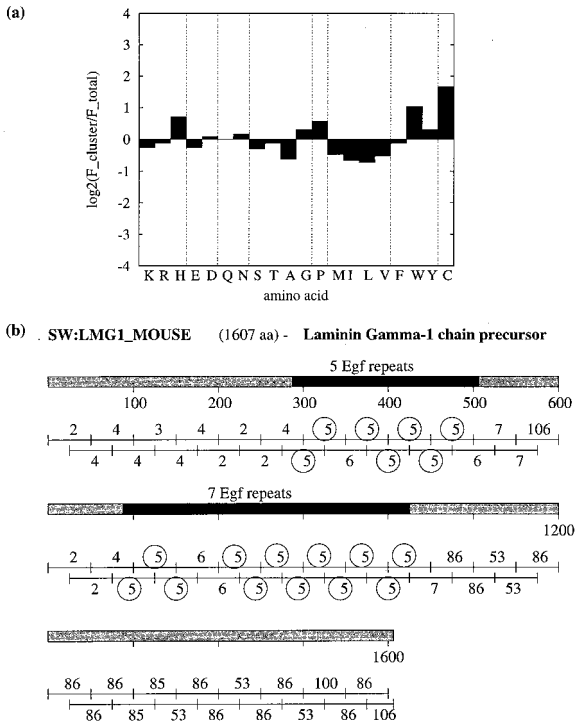


**Figure 6.** (a) Amino-acids' distribution in cluster 5. Despite the large size of this cluster (5935 segments) and its high complexity (over 200 families), it shows a unique amino acid distribution, with prevalent cysteine, histidine, tryptophan, proline and glycine. All these residues impose major structural features. All hydrophobic amino acids are underrepresented. For details on the representation, see Figure 3. (b) Cluster 5, the EGF motif and related patterns. Cluster 5 is rich with proteins containing the EGF-like domain. One such example is sw:lmg1_mouse–laminin gamma-1 chain precursor (laminin b2 chain). Laminin is a complex glycoprotein, consisting of three different polypeptide chains, which are bound to each other by disulfide bonds into a cross-shaped molecule. The protein contains EFG-like domains, denoted by the filled boxes. Each box matches several tandem repeats. Note that all the segments classified to cluster 5 correspond to the EGF domain. Four segments were classified to cluster 6 (which is the closest cluster to cluster 5). This is due to the fact that some of the EGF repeats are less conserved, and may exhibit a slightly different composition (see PROSITE documentation on the EGF-like motif). For details on the representation, see Figure 4.

to the sequence of clusters in to which its segments fall. In other words, every protein is encoded by a ''word'' over an alphabet of 106 ''characters'' (the clusters). A natural similarity measure on full-length proteins emerges. Namely, apply dynamic programming, where the similarity score between characters depends only on the distance among the clusters' centroids and the clusters' sizes†. Penalties for gaps are high, since even a single omission entails a gap of 25 amino acid residues. We denote this new way of comparison between complete proteins as BMR, ''best matching route''.

A multiple alignment of the different members of the acetylcholine receptor (AChR) family using this representation is shown in Figure 10. An examination of the string of clusters' numbers already shows a subdivision within the acetylcholine receptors, dividing alpha, beta, delta, epsilon and gamma subunits into small related groups.

† The charge for switching from cluster $i$ to cluster $j$ is taken as $c \log(p(i)p(j)) + d(i, j)$. Here $p(i)$ and $p(j)$ are the clusters' relative sizes, and $d(i, j)$ is their Euclidean distance. The constant $c$ is optimized for best performance. This measure accounts for the difference in clusters' sizes and the fact that the presence of a small cluster in the clusters' sequence is more significant than the presence of a large cluster.

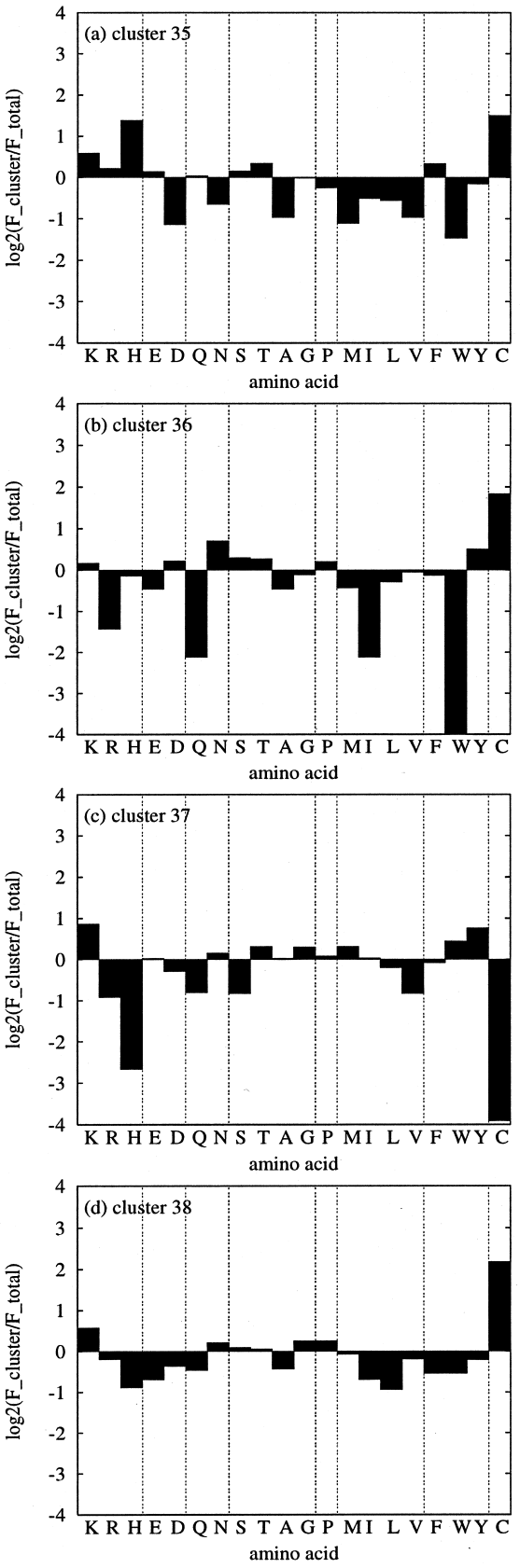


**Figure 7.** Clusters 35 to 38. These four clusters have a common ancestor in the hierarchical tree (Figure 1). The four clusters are closely related despite the large differences in their amino acid distribution. For details on the representation, see Figure 3.

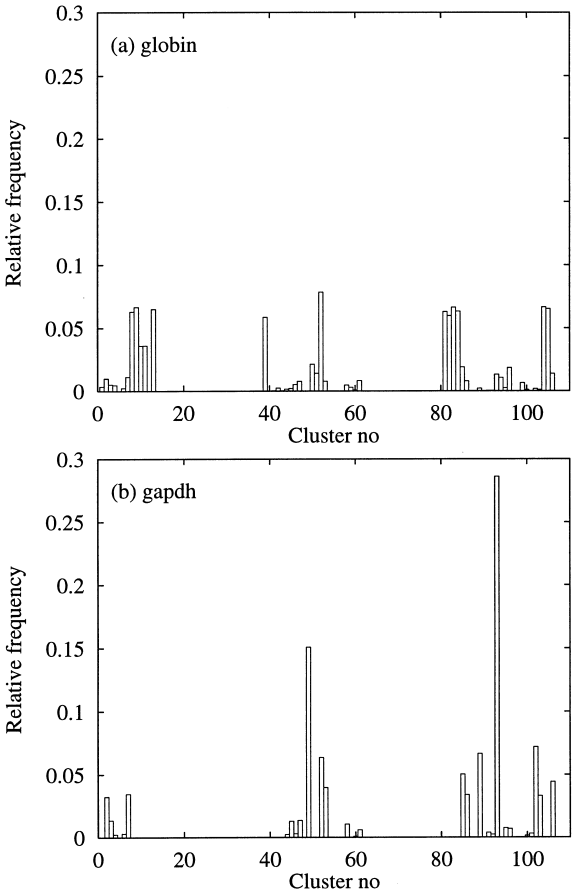


**Figure 8.** The representation of several biological families by their fingerprints. Every biological family is fitted with the distribution of segments of proteins from the family among the clusters, thus obtaining a new representation of the family (fingerprints). The relative frequency of a cluster is defined as the number of segments within the family classified to this cluster, divided by the total number of segments within the family. (a) and (b) Globins and gapdh (glyceraldehyde-3-phosphate dehydrogenase) are two well-studied families with 666(3434) and 103(1332) proteins(segments), respectively. The characteristic and complex fingerprints of each of these families is shown in (a) and (b), respectively.

While few of the segments are common to all subunits, most of them are common to subsets of these different subunits. Note that while the visualization of the multiple alignment of the complete proteins is not practical in this case (the average length of proteins in this family is 500 amino acid residues), and lacks the clarity needed to understand the complicated connections that reside between the different subunits, this new representation of complete proteins reduces significantly the details, while maintaining the important information within. The BMR algorithm can be applied to generate a quantitative measure of similarity among the AChR subfamilies (unpublished results).

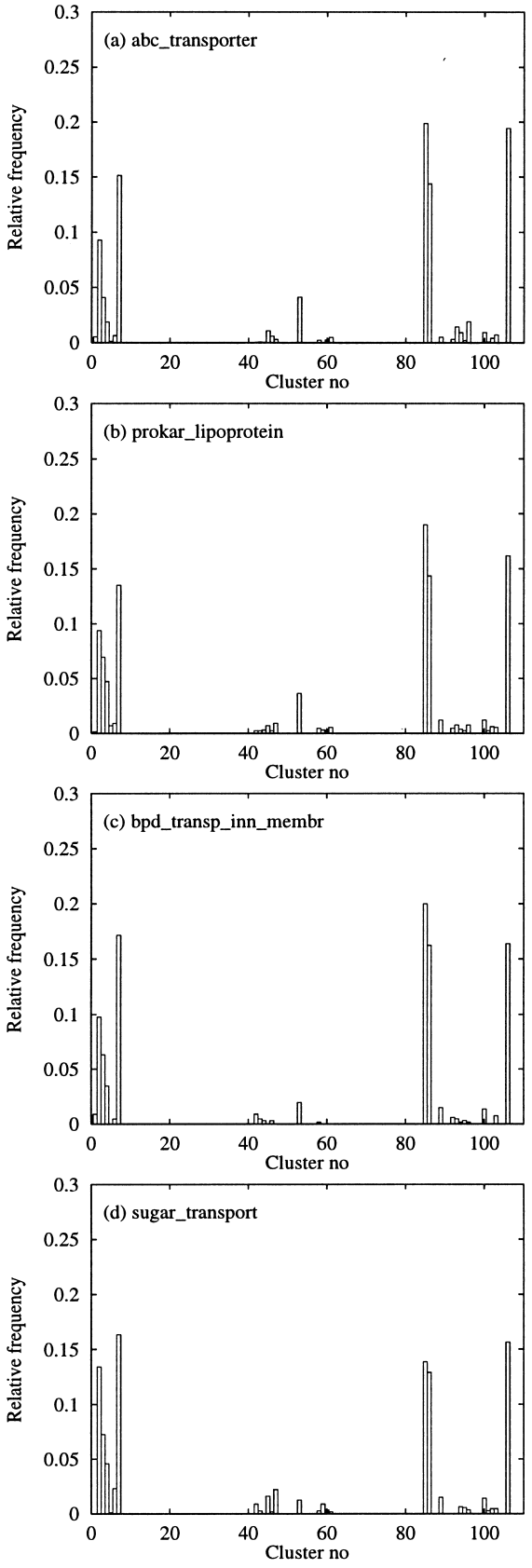


**Figure 9.** Families with similar clustering profiles (fingerprints). ABC transporter (179 proteins, 3898 segments), prokaryotic lipoprotein (131 proteins, 1311 segments), binding protein-dependent transporter of the inner membrane (60 proteins, 665 segments) and the bacterial sugar transporter (57 proteins, 1132 segments), all have similar (but non-identical) clustering profiles in (a) to (d), respectively. Their mutual distances are very small. All the transporters are proteins with multiple membrane-spanning domains. Prokaryotic lipoprotein consists of proteins with a membrane-attached domain. Note that many of the clusters that prevail in the distributions of these four families (clusters 2, 7, 85, 86 and 106) are very large (see Table 1). So, while membership in individual clusters is not very informative in this case, the complete fingerprint does provide a very useful characterization common to these families, which distinguishes them from the rest. Additional families of membranous proteins, including neurotransmitter-gated ion-channels and G-protein receptors have fingerprints that resemble, to varying degrees, the fingerprints shown in (a) to (d).

While it is obvious that this representation maintains the original nature of full-length proteins, and may be used towards a more refined classification of families (Figure 10), it is intriguing to find out whether it reveals other interesting features of proteins. We tested this new method on full-length proteins in comparison with the SW algorithm (Table 2). We translated all proteins in the database into sequences of characters in the alphabet of 106 characters and compared each protein against the database, using the BMR algorithm, in search of related proteins. The quality of performance was estimated by taking a single member from each family in PROSITE, comparing it against all the database, and identifying its related proteins in the family. Identification was based on the following ''equivalence number'' identification criterion (Pearson, 1995): define the cutoff score as the similarity score that balances the number of related sequences below it and the number of unrelated sequences with score above it (i.e. the score where the number of false positives equals the number of false negatives). Only proteins with score at or above the cutoff score are considered as identified. The results were compared against the SW algorithm with the blosum62 scoring matrix and values $-10, -1$ for gap penalties, currently considered the best method known (Pearson, 1995).

Already with the BMR's simplistic approach, it competed successfully with SW on about 80 families of varying sizes (see Table 2). The performance of the BMR method is superior for families that are well characterized in terms of structure or function, since these families fall into small clusters that receive a high score (see the footnote to page 549). The BMR method and its biological consequence will be described in more detail elsewhere.

| Access code | Length | No of segments | Column no 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 | 16 | 17 | 18 | 19 | 20 | 21 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| acha_bovin. | 457 | ( 18) | 45 | 86 | 106 | . | 60 | 89 | 4 | 3 | 6 | 6 | 1 | 53 | 53 | 94 | 94 | 2 | 2 | 106 | 103 | | |
| acha_chick. | 456 | ( 18) | 45 | 86 | 106 | . | 60 | 89 | 4 | 3 | 5 | 5 | 1 | 53 | 53 | 94 | 94 | 7 | 106 | 85 | 102 | | |
| acha_human. | 482 | ( 19) | 45 | 86 | 85 | 4 | 60 | 89 | 4 | 3 | 6 | 4 | 1 | 53 | 53 | 59 | 94 | 3 | 2 | 85 | 103 | | |
| acha_mouse. | 457 | ( 18) | 45 | 86 | 106 | . | 60 | 89 | 4 | 3 | 6 | 6 | 1 | 53 | 53 | 94 | 94 | 2 | 2 | 103 | 103 | | |
| acha_najna. | 104 | ( 4) | | | | | | | 4 | 6 | 6 | 6 | | | | | | | | | | | |
| acha_natte. | 104 | ( 4) | | | | | | | 3 | 6 | 6 | 6 | | | | | | | | | | | |
| acha_rat. | 457 | ( 18) | 45 | 86 | 106 | . | 60 | 89 | 4 | 3 | 6 | 6 | 1 | 53 | 53 | 94 | 94 | 2 | 2 | 103 | 103 | | |
| acha_torca. | 461 | ( 18) | 92 | 93 | 106 | . | 60 | 89 | 4 | 3 | 5 | 5 | 1 | 53 | 53 | 59 | 94 | 106 | 85 | 106 | 103 | | |
| acha_torma. | 461 | ( 18) | 92 | 93 | 7 | . | 60 | 89 | 4 | 3 | 5 | 5 | 1 | 53 | 53 | 59 | 94 | 106 | 85 | 106 | 103 | | |
| achb_bovin. | 505 | ( 20) | 85 | 85 | 7 | . | 60 | 7 | 4 | 3 | 2 | 3 | 106 | 86 | 86 | 53 | 94 | 4 | 6 | 4 | 106 | 58 | 58 |
| achb_human. | 501 | ( 20) | 85 | 86 | 106 | . | 60 | 106 | 3 | 3 | 3 | 4 | 106 | 86 | 86 | 53 | 94 | 6 | 6 | 4 | 85 | 58 | 3 |
| achb_mouse. | 501 | ( 20) | 86 | 86 | 106 | . | 60 | 7 | 4 | 3 | 2 | 3 | 106 | 86 | 86 | 53 | 94 | 6 | 6 | 4 | 58 | 58 | 58 |
| achb_rat. | 501 | ( 20) | 86 | 86 | 106 | . | 60 | 106 | 4 | 3 | 3 | 3 | 106 | 86 | 86 | 53 | 94 | 5 | 6 | 6 | 58 | 58 | 58 |
| achb_torca. | 493 | ( 19) | 86 | 86 | 106 | . | 60 | 43 | 4 | 2 | 2 | 3 | . | 85 | 53 | 53 | 94 | 4 | 3 | 3 | 85 | 58 | 3 |
| achd_bovin. | 516 | ( 20) | 106 | 86 | 42 | . | 60 | 43 | 4 | 2 | 103 | 3 | 1 | 86 | 53 | 46 | 44 | 106 | 85 | 106 | 3 | 3 | 4 |
| achd_chick. | 513 | ( 20) | 106 | 85 | 43 | . | 60 | 4 | 6 | 3 | 4 | 4 | 1 | 53 | 53 | 46 | 44 | 3 | 7 | 106 | 2 | 2 | 4 |
| achd_human. | 517 | ( 20) | 7 | 86 | 43 | . | 60 | 4 | 6 | 3 | 103 | 3 | 1 | 53 | 53 | 46 | 106 | 106 | 85 | 85 | 2 | 3 | 4 |
| achd_mouse. | 520 | ( 20) | 7 | 86 | 43 | . | 60 | 3 | 4 | 3 | 103 | 2 | 1 | 86 | 53 | 46 | 44 | 7 | 106 | 106 | 2 | 3 | 3 |
| achd_rat. | 517 | ( 20) | 7 | 86 | 42 | . | 60 | 89 | 4 | 3 | 103 | 3 | 1 | 86 | 53 | 46 | 44 | 106 | 85 | 106 | 2 | 3 | 4 |
| achd_torca. | 522 | ( 20) | 3 | 86 | 43 | . | 6 | 4 | 6 | 4 | 103 | 4 | 1 | 53 | 53 | 46 | 3 | 7 | 106 | 7 | 7 | 106 | 3 |
| achd_xenla. | 521 | ( 20) | 3 | 86 | 43 | . | 60 | 3 | 4 | 3 | 103 | 2 | 85 | 53 | 53 | 46 | 2 | 7 | 85 | 106 | 3 | 3 | 3 |
| ache_bovin. | 491 | ( 19) | 2 | 103 | 106 | . | 60 | 7 | 3 | 2 | 3 | 3 | 85 | 86 | 53 | 46 | 106 | . | 86 | 106 | 3 | 89 | 92 |
| ache_human. | 493 | ( 19) | 7 | 103 | 106 | . | 60 | 43 | 3 | 3 | 4 | 3 | 85 | 86 | 53 | 46 | 106 | . | 85 | 106 | 2 | 92 | 92 |
| ache_mouse. | 493 | ( 19) | 2 | 103 | 106 | . | 60 | 43 | 4 | 3 | 3 | 3 | 85 | 86 | 53 | 46 | 85 | . | 86 | 106 | 2 | 89 | 92 |
| ache_rat. | 494 | ( 19) | 2 | 103 | 85 | . | 60 | 43 | 4 | 2 | 3 | 3 | 85 | 86 | 53 | 46 | 85 | . | 86 | 106 | 2 | 89 | 92 |
| achg_bovin. | 519 | ( 20) | 2 | 102 | 43 | . | 60 | 60 | 6 | 3 | 103 | 7 | 85 | 86 | 53 | 53 | 106 | 2 | 106 | 7 | 4 | 2 | 2 |
| achg_chick. | 514 | ( 20) | 7 | 102 | 43 | . | 60 | 43 | 6 | 3 | 2 | 2 | 85 | 86 | 53 | 53 | 2 | 3 | 85 | 106 | 2 | 89 | 3 |
| achg_human. | 517 | ( 20) | 106 | 102 | 43 | . | 60 | 60 | 6 | 2 | 103 | 7 | 85 | 86 | 53 | 53 | 85 | 106 | 7 | 106 | 4 | 3 | 3 |
| achg_mouse. | 519 | ( 20) | 103 | 102 | 43 | . | 60 | 60 | 6 | 2 | 103 | 106 | 85 | 86 | 53 | 53 | 106 | 7 | 7 | 106 | 2 | 7 | 3 |
| achg_rat. | 519 | ( 20) | 103 | 102 | 43 | . | 60 | 3 | 6 | 2 | 103 | 7 | 85 | 86 | 53 | 53 | 106 | 2 | 7 | 106 | 2 | 2 | 3 |
| achg_torca. | 506 | ( 20) | 86 | 86 | 43 | . | 60 | 43 | 6 | 4 | 4 | 4 | 85 | 53 | 86 | 46 | 106 | 7 | 7 | 7 | 89 | 89 | 92 |
| achg_xenla. | 510 | ( 20) | 85 | 102 | 43 | . | 60 | 2 | 3 | 2 | 2 | 3 | 85 | 86 | 53 | 53 | 106 | 7 | 85 | 106 | 3 | 3 | 3 |

**Figure 10.** Multiple alignment of the acetylcholine receptor family. Thirty-two members of the acetylcholine receptor family are shown, divided into alpha, beta, delta, epsilon and gamma subunits. The proteins are represented by the sequence of clusters into which their segments fall, and aligned using the BMR algorithm for pairwise comparison (see the text for details). In columns where there are only a few predominant clusters, these clusters are colored. The color intensity is proportional to the degree of conservation of a cluster within an aligned column in these 32 proteins. Only a few columns are colored for illustration. This representation emphasizes the regions that are shared by all subunits (see the fifth column), and those that are shared by some of the subunits. For example, in the third column, alpha, beta and epsilon share almost the same segment, while delta and gamma subunits share a slightly different one. However, clusters 43 and 106 are geometrically close, an evidence for the strong relation between these segments. Likewise, in the 14th column, beta and gamma share a common segment, while delta and epsilon share a different one. Still, these segments are close (the distance between clusters 46 and 53 is small), indicating the common nature they all share as members of the same ''mother'' family.

**Table 2.** Performance of BMR compared with that of SW

| Family | No. of proteins | BMR | SW | Query |
|---|---|---|---|---|
| Rubisco_Large | 224 | 222 | 212 | P35214 |
| Tubulin | 164 | 160 | 140 | P02568 |
| Egf | 119 | 54 | 53 | P07246 |
| Actins_1 | 106 | 106 | 94 | P25160 |
| Gapdh | 102 | 97 | 95 | P17336 |
| Hsp70_1 | 101 | 86 | 85 | P22879 |
| Histone_H2a | 59 | 59 | 54 | P19140 |
| Chaperonins_Cpn60 | 55 | 52 | 43 | P19866 |
| Lactalbumin_Lysozyme | 54 | 51 | 43 | P08992 |
| Histone_H2b | 53 | 52 | 51 | P16868 |
| Metallothionein_Cl1 | 46 | 43 | 42 | P02303 |
| Reca | 42 | 36 | 29 | P29843 |
| Pglycerate_Kinase | 39 | 31 | 31 | P18564 |
| Tropomyosin | 36 | 33 | 26 | P05697 |
| Chalcone_Synth | 36 | 36 | 35 | P00705 |
| Ribosomal_S12 | 34 | 25 | 24 | P25336 |
| Histone_H3_2 | 31 | 27 | 27 | P09862 |
| Catalase_1 | 30 | 28 | 25 | P14717 |
| C2_Domain | 27 | 18 | 18 | P27362 |
| Pal_Histidase | 25 | 17 | 13 | P14714 |
| 1433_1 | 25 | 21 | 18 | P29254 |
| Photosystem_I_Psaab | 24 | 23 | 22 | P11383 |
| Enolase | 24 | 22 | 17 | P26348 |
| Arf | 24 | 19 | 18 | P10851 |
| Trp_Synthase_Beta | 19 | 19 | 19 | P33421 |
| Phytochrome | 19 | 17 | 16 | P08562 |
| Tfiid | 15 | 15 | 14 | P13393 |
| Biopterin_Hydroxyl | 15 | 14 | 14 | P20077 |

Some of the families on which BMR did at least as well as SW are shown. For each such family, we show the number of proteins in the family longer than 50 amino acid residues (second column), and the number of proteins from the family that were identified using the ''equivalence number'' identification criterion (Pearson, 1995), in either method (third and fourth columns). Accession numbers for the queries are given in the last column. See the text for further details.

## Discussion

We present a novel method for self-organizing complex data and demonstrate its performance by globally organizing all known proteins. Our method employs the current best sequence comparison algorithm, namely, SW dynamic programming with the blosum62 similarity matrix and the matching parameters for penalizing sequence gaps (Pearson, 1995). Evaluations of this algorithm on full-length proteins showed excessive dependence on protein lengths. Furthermore, this algorithm usually fails to detect multi-trait features in proteins. Consequently, we chose to normalize the lengths, and computed this metric on segments of 50 amino acid residues (see the footnote to page 540 for reservations). This choice of fragment length is made according to the length of patterns in the PROSITE classification, most of which are between 5 and 40 amino acid residues long. The choice of 50-mer fragments is consistent with structural features in proteins, since many folds consist of 20 to 80 amino acid residues. Still, performing our procedure at other segment lengths may yield different granularity and eventually new interesting insights on other classes of proteins.

Our procedure allows, for the first time, a full-scale comparison of nearly 40,000 proteins. Note that the only biological information utilized by our method comes in the form of a reliable and computable local metric. Given the pairwise distances among protein segments, all segments are carefully clustered into statistically significant families.

Our approach has to overcome two major obstacles: (1) the data are inherently high-dimensional; (2) it is hard to organize it accordingly to a provably valid model. We deal with the first issue by using a novel geometric embedding of the segments in a lower-dimensional Euclidean space, with small distortion. The second problem is handled through a careful cross-validated hierarchical clustering of the segments in this lower-dimensional space.

So far, our work has yielded a classification into only 106 classes (Figure 1, Table 1). Yet, even with this small number of 106 clusters we found many significant biological signatures. Some known families of proteins were clustered into well-distinguished clusters, and other clusters match well-known motifs and domains within proteins. Kinship of protein families could be inferred from the clustering tree: different families that were clustered into the same cluster, or split from the same ancestor cluster may share some biological features. A similarity measure emerges for full-length proteins as well. Proteins can be characterized by their clusters sequence. This representation leads to a quantitative comparison measures between full-length proteins, based on the best matching route (BMR) of clusters. Indeed, in many instances, our

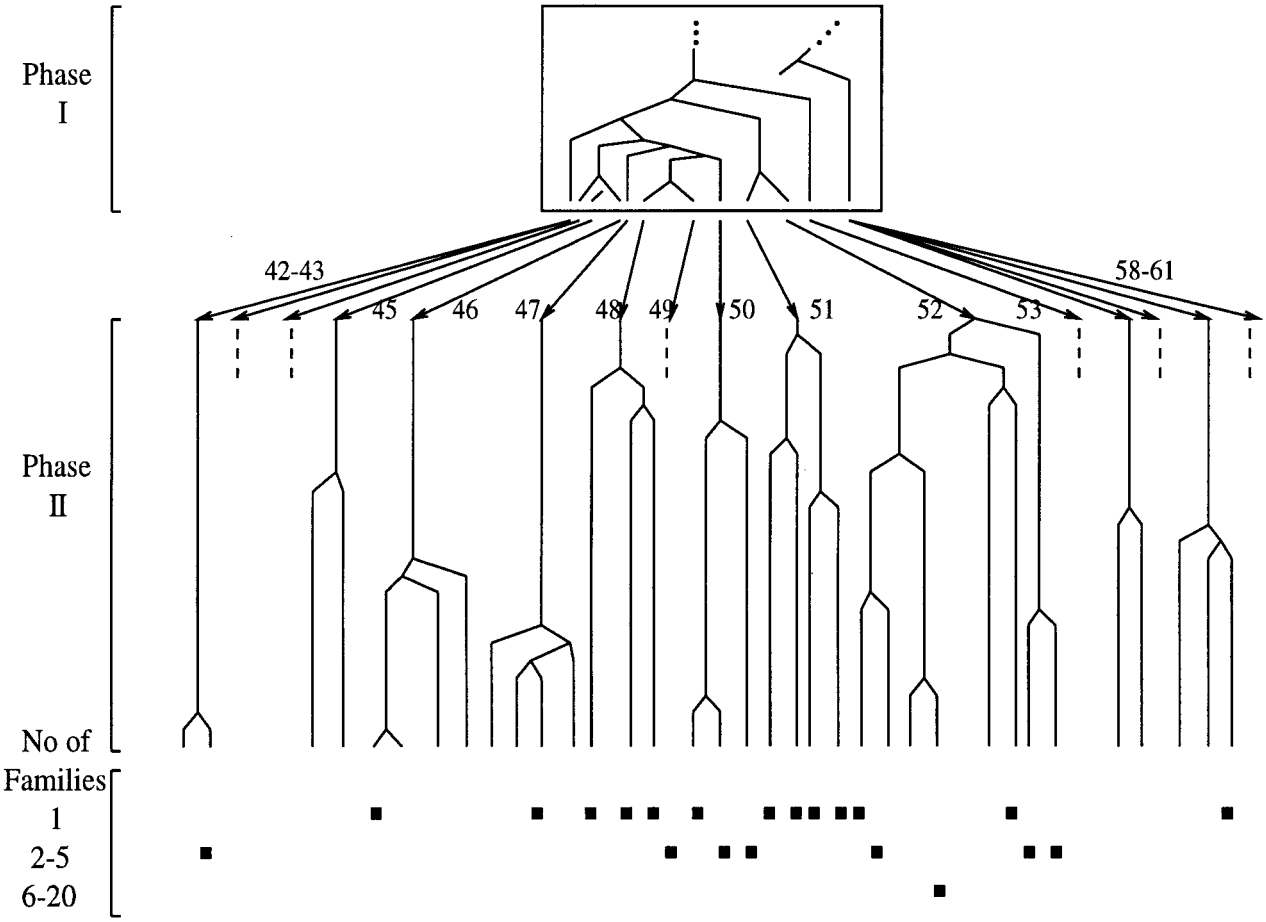


**Figure 11.** Second phase splitting with delayed cross-validation. Further splitting is performed, under the same strict criteria for stable splits (see the text). However, at this phase, these criteria are not verified at every step, so cross-validation is carried out only after all splittings are performed. The original (phase I) 106 clusters yield 146 clusters. This Figure shows the more refined tree structure for the clusters numbered 42 to 53 and 58 to 61 in the first phase (total of 55,599 segments). Only six clusters (numbers 43, 44, 49, 53, 59 and 61) remained intact. The other ten split into 35 subclusters. The rectangular box shows these clusters at the end of phase I, and the resulting subclusters are below. The leaves of the tree show 35 of the 146 clusters in phase II. Clusters that represent only a few families are denoted by a small filled box at the leaves. Note that for the clusters of the second phase, this splitting resulted in clusters of only one family (13 cases), and two to five families (seven cases). The other clusters are still very large. Clearly, both small and large clusters were affected. For instance, the subclusters originating from cluster 48 each represent a single PROSITE family. Some of the highly complex clusters are affected, e.g. clusters 45 to 47 and 58 to 61, each with 150 to 250 families.

comparison method (BMR) outperformed the currently best sequence comparison method (SW).

However, in view of the 700 PROSITE families, a more refined classification seems desirable. We are currently testing versions of the clustering process, where cross-validation is applied only once in a number of splitting phases. More permissive cross-validation procedures may still yield meaningful, more refined classifications. The outcome of one such procedure is shown in Figure 11. Starting from the above 106 clusters, clusters with high aspect-ratios were split and cross-validation was performed only subsequently, when the number of clusters reached 150. Four clusters failed the cross-validation test, and their segments were returned to the general pool. Thus 146 clusters were obtained, all satisfying the same cross-validation criteria: 16 of the original clusters that underwent further splitting are shown, resulting in 41 subclusters. Clearly, both small and large clusters were affected (compare with Table 1). This procedure also verifies the stability of relations between protein families that can be suggested on the basis of the tree of 106 clusters. It indicates a weak relation in cases where the participating families were set apart, and a strong relation when they remained together. When applied to the 146 clusters, BMR did better than SW on 11 additional families, indicating further potential for this method.

Our standard yardstick here is the PROSITE classification. While this is a major reference against which results such as ours ought to be checked, certain shortcomings of the PROSITE classification should be kept in mind. Only 46% of

the proteins are currently classified in PROSITE. Moreover, the classification is often determined on the basis of very short subsequences, less than ten residues in some cases, which often represent various signals or very local, small sites, and not necessarily structural or functional domains. Besides, most of the families are small, containing only a few members each (over 80% of the families have less than 30 members in each). Our strict criteria for validity stops the clustering process short of complete resolution, thus many small families are ''lost'' in bigger clusters. We can expect further progress when more proteins from small families are sequenced.

Besides the immediate information about biological patterns that can be derived from the clusters, they yield additional insight into the classification of protein families. Protein families have characteristic distributions among the clusters, which we call fingerprints. While most of the 106 obtained clusters correspond to a single functional protein family, most segments belong to very few large, non-specific, clusters. Still, the fingerprints of families that do not correspond to a single cluster are characteristic enough to distinguish important functional protein families. Comparisons between fingerprints of distinct PROSITE families yield similarity indices of both statistical and biological significance, where families of similar biological roles tend to have similar fingerprints. Such indices can be helpful in defining families and superfamilies.

Our segment clustering approach provides an elegant, higher-level, representation of protein sequences. We believe that these tools can be refined and extended to larger protein databases, and provide more accurate predictions on the relationships among protein families and the nature of new sequences.

## Acknowledgments

We thank Amnon Barak, and Compugen Ltd. for generously making their computational facilities available to us, and H. Margalit and D. Parnas for valuable discussions.

## References

Altschul, S. F., Carrol, R. J. & Lipman, D. J. (1990). Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.

Bairoch, A. (1993). The PROSITE dictionary of sites and patterns in proteins, its current status. *Nucl. Acids Res.* **21**, 3097–3103.

Bairoch, A. & Boeckman, B. (1992). The SWISS-PROT protein sequence data bank. *Nucl. Acids Res.* **20**, 2019–2022.

Barak, A., Laden, O. & Yarom, Y. (1995). The NOW MOSIX and its preemptive process migration scheme. *Bull. IEEE Tech. Comm. Operat. Syst. Appl. Environ.* **7**, 5–11.

Bourgain, J. (1985). On Lipschitz embedding of finite metric spaces in Hilbert space. *Israel J. Math.* **52**, 46–52.

Compugen Ltd., BIOCCELERATOR Manual, 1996 (http://www.compugen-us.com).

Cover, T. M. & Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Trans. Info. Theory,* **IT-13**, 21–27.

Cover, T. M. & Thomas, J. A. (1991). In *Elements of Information Theory,* pp. 12–33, John Wiley and Sons, New York.

Cukierman, E., Huber, I., Rotman, M. & Cassel, D. (1995). The ARF1 GTPase-activating protein: zinc finger motif and Golgi complex localization. *Science,* **270**, 1999–2002.

Duda, R. O. & Hart, P. E. (1973). In *Pattern Classification and Scene Analysis,* pp. 189–252, John Wiley and Sons, New York.

Ferran, E. A., Pflugfelder, B. & Ferrara, P. (1994). Self-organized neural maps of human protein sequences. *Protein Sci.* **3**, 507–521.

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

Gray, R. M. (1984). Vector quantization. *IEEE ASSP Mag.* 4–29.

Gray, R. M., Kieffer, J. C. & Linde, Y. (1980). Locally optimal block quantizier design. *Inform. Control,* **45**, 178–198.

Gribskov, M., Mclachlen, A. D. & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proc. Natl Acad. Sci. USA,* **84**, 4355–4358.

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

Hanke, J., Beckmann, G., Bork, P. & Reich, J. G. (1996). Self-organizing hierarchic networks for pattern recognition in protein sequence. *Protein Sci.* **5**, 72–82.

Hardison, R. C. (1996). A brief history of hemoglobins: plant, animal, protist, and bacteria. *Proc. Natl Acad. Sci. USA,* **93**, 5675–5679.

Harris, N. L., Hunter, L. & States, D. J. (1992). Mega-classification: discovering motifs in massive datastreams. In *Proc. 10th Natl Conf. on AI*, pp. 837–842, AAAI press, The MIT Press, Menlo Park, Cambridge.

Henikoff, S. & Henikoff, J. G. (1991). Automated assembly of protein blocks for database searching. *Nucl. Acids Res.* **19**, 6565–6572.

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

Kearns, M. J. & Vazirani, U. V. (1994). In *An Introduction to Computational Learning Theory,* pp. 1–48, MIT Press, Cambridge, MA.

Linial, N., London, E. & Rabinovich, Yu (1995). The geometry of graphs and some of its algorithmic applications. *Combinatorica,* **15**, 215–245.

Lipman, D. J. & Pearson, W. R. (1985). Rapid and sensitive protein similarity. *Science,* **227**, 1435–1441.

Pearson, W. R. (1995). Comparison of methods for searching protein sequence databases. *Protein Sci.* **4**, 1145–1160.

Pereira, F., Tishby, N. & Lee, L. (1993). Distributional clustering of English words. In *Proc. 30th Annual Meeting of the Association for Computational Linguistics,* pp. 183–190, ACL Press.

Press, W. H., Teukolsky, S. A., Vetterling, W. T. & Flannery, B. P. (1988). In *Numerical Recipes in C,* pp. 60–72. Cambridge University Press, Cambridge.

Sheridan, R. P. & Venkataraghavan, R. (1992). A systematic search for protein signature sequences. *Proteins: Struct. Funct. Genet.* **14**, 16–28.

Smith, T. F. & Waterman, M. S. (1981). Comparison of biosequences. *Advan. Appl. Math.* **2**, 482–489.

Sonnhammer, E. L. L. & Kahn, D. (1994). Modular arrangement of proteins as inferred from analysis of homology. *Protein Sci.* **3**, 482–492.

Taylor, W. R. (1990). Hierarchical method to align large numbers of biological sequences. *Methods Enzymol.* **183**, 456–474.

van Heel, M. (1991). A new family of powerful multivariate statistical sequence anaysis techniques. *J. Mol. Biol.* **220**, 877–887.

Vapnik, V. N. (1982). In *Estimation of Dependences Based on Empirical Data,* pp. 162–176, Springer-Verlag, New York.

Watanabe, H. & Otsuka, J. (1995). A comprehensive representation of extensive similarity linkage between large numbers of proteins. *CABIOS,* **11**, 159–166.

Wootton, J. C. (1994). Sequences with 'unusual' amino acid compositions. *Curr. Opin. Struct. Biol.* **4**, 413–421.

Wu, C., Whitson, G., Mclarty, J., Ermongkonchai, A. & Chang, T. (1992). Protein classification artificial neural system. *Protein Sci.* **1**, 667–677.

***Edited by F. E. Cohen***

*(Received 12 November 1996; received in revised form January 1997; accepted 17 January 1997)*