A map of the protein space - An automatic hierarchical classification of all protein sequences

Golan Yona, Nathan Linial, Naftali Tishby
Institute of Computer Science
Hebrew University
Jerusalem 91904
Israel
email: golany,nati,tishby@cs.huji.ac.il

Michal Linial
Department of Biological Chemistry
Institute of Life Sciences
Hebrew University
Jerusalem 91904
Israel
email: michall@leonardo.ls.huji.ac.il

Abstract

We investigate the space of all protein sequences. We combine the standard measures of similarity (SW, FASTA, BLAST), to associate with each sequence an exhaustive list of neighboring sequences. These lists induce a (weighted directed) graph whose vertices are the sequences. The weight of an edge connecting two sequences represents their degree of similarity. This graph encodes much of the fundamental properties of the sequence space.

We look for clusters of related proteins in this graph. These clusters correspond to strongly connected sets of vertices. Two main ideas underlie our work: i) Interesting homologies among proteins can be deduced by transitivity. ii) Transitivity should be applied restrictively in order to prevent unrelated proteins from clustering together.

Our analysis starts from a very conservative classification, based on very significant similarities, that has many classes. Subsequently, classes are merged to include less significant similarities. Merging is performed via a novel two phase algorithm. First, the algorithm identifies groups of possibly related clusters (based on transitivity and strong connectivity) using local considerations, and merges them. Then, a global test is applied to identify nuclei of strong relationships within these groups of clusters, and the classification is refined accordingly. This process takes place at varying thresholds of statistical significance, where at each step the algorithm is applied on the classes of the previous classification, to obtain the next one, at the more permissive threshold. Consequently, a hierarchical organization of all proteins is obtained.

The resulting classification splits the space of all protein sequences into well defined groups of proteins. The results show that the automatically induced sets of proteins are closely correlated with natural biological families and super families. The hierarchical organization reveals finer sub-families that make up known families of proteins as well as many interesting relations between protein families. The hierarchical organization proposed may be considered as the first map of the space of all protein sequences.

An interactive web site including the results of our analysis has been constructed, and is now accessible through http://www.protomap.cs.huji.ac.il

Keywords: clustering, protein families, protein classification, sequence alignment, sequence homology.

Introduction

In recent years we have been witnessing a constant flow of new biological data. Large-scale sequencing projects throughout the world turn out new sequences, and create new challenges for investigators. Many sequences that are added to the databases are unannotated and await analysis. Currently, 12 complete genomes (of yeast, E. coli, and other bacteria) are available. Between 35%-50% of their proteins have an unknown function (Pennisi 1997).

In the absence of structural data, analysis necessarily starts by investigating the sequence proper. Sequence analysis has many aspects: composition, hydrophobicity, charge, secondary structure propensity and more. The most effective analyses compare the sequence under study with the whole database, in search for close relatives. Properties of a new protein sequence are extrapolated from those of its neighbors. Since the early 70's, algorithms were developed for the purpose of comparing protein sequences efficiently and reliably (Needleman and Wunsch 1970; Smith and Waterman 1981; Lipman and Pearson 1985; Altschul et al. 1990).

Even with the best alignment of two protein sequences at hand, the basic question remains: Do they share the same biological function or not. It is generally claimed that two sequences with over 30% identity along much of the sequences, are very likely to have the same fold (Sander and Schneider 1991; Flores et al. 1993; Hilbert et al. 1993). Proteins of the same fold usually have similar biological functions. Nevertheless, one encounters many cases of high similarity in fold, despite a low sequence similarity (Murzin 1993; Pearson 1997). Such instances are, unfortunately, missed by simple searches over the database.

Detecting homology may often help in determining the function of new proteins. By definition, homologous proteins evolved from the same ancestor protein. The degree of conservation varies among protein families. However, homologous proteins almost always have the same fold (Pearson 1996). Homology is, by definition, a transitive relation: If A is homologous to B, and B is homologous to C, then A is homologous to C. This simple observation can be very effective in discovering homology. However, when applied simple-mindedly, this observation may also lead to many pitfalls.

Though the common evolutionary origin of two proteins is almost never directly observed, we can deduce homology among proteins, with a high statistical confidence, given that the sequence similarity is significant. This is particularly useful in the so called "twilight zone" (Doolittle 1992), where sequences are identical to, say, 10-25%. Transitivity can be used to detect related proteins, beyond the power of a direct search.

The potential value of transitivity has been observed before. In (Watanabe and Otsuka 1995) transitivity and single linkage clustering are employed to extract similarities among 2000 E. coli protein sequences. In (Koonin et al. 1996) a similar analysis is performed on 75% of the proteins encoded in the E. coli genome. The power of transitivity in inferring homology among distantly related proteins (e.g., Streptomyces griseus protease A and protease B) is demonstrated in (Pearson 1997). In (Neuwald et al. 1997) transitivity was combined with a search through the database, for the purpose of modeling a protein family, starting from a single sequence. However, the full power of this idea has not yet been exploited. On the other hand, the perils of transitivity have not been thoroughly investigated either. Here we address this problem as well.

It should be clear that similarity is not transitive, and that it does not necessarily entail homology[1]. Therefore similarity should be carefully used in attempting to deduce homology. The statistical significance level of the similarity should take into account the level of evolutionary divergence within the family, in order to deduce reliable homologies.

Multidomain proteins make the deduction of homology particularly difficult: If protein 1 contains domains A and B, protein 2 contains domains B and C, protein 3 contains domains C and D, then should proteins 1 and 3 be considered homologous? This simple example indicates the difficulty with single-linkage clustering for our purpose.

Expert biologists can distinguish significant from insignificant similarities. However, the sheer size of current databases rules out an exhaustive manual computation of homologies. This is why we developed an automatic method for this task. Our algorithm attempts to discard chance similarities and indirect multiple-domain-based connections.

Our starting point is very strict high resolution classification that employs only connections of very high statistical significance. Henceforth, clusters are merged to form bigger and more diverse clusters. Our algorithm automatically attempts to identify suspicious/problematic connections and to eliminate as well as possible false connections between unrelated proteins. The algorithm operates hierarchically, each time considering weaker connections. Its output is thus a hierarchical organization of all protein sequences. The algorithm incorporates no further biological information and uses only the similarity scores that are provided by standard methods.

This approach leads to the definition of a new metric on the space of all protein sequences. We believe that this emerging metric is more sensitive than existing measures. Such metrics are necessary in the quest of a global self organization of all protein sequences, as discussed in (Linial et al. 1997).

Methods

This section contains a description of our computational procedure. The procedure was carried out on the SWISSPROT database (Bairoch and Boeckman 1992) release 33, with total of 52205 proteins.

Defining the graph

We represent the space of all protein sequences as a directed graph, whose vertices are the protein sequences. Edges between the vertices are weighted with weights that reflect the distance or dissimilarity between the corresponding sequences, i.e. high similarity translates to a small weight (or distance). To compute the weight of the directed edge from A to B, one compares A against all sequences in the SWISSPROT database, and obtains a distribution of scores. The weight is taken as the expectation value (Altschul et al. 1994) of the similarity score between A and B, based on this distribution. This is a statistical estimate for the number of occurrences of the appropriate score at a random setup, assuming the existing amino acid composition. A high expectation value entails a weak connection. Edges of statistically insignificant similarity scores, get discarded

[1]Similarity may be quantified whereas homology is a relation that either holds or does not hold. Significant similarities can be used to infer homology, with a level of confidence that depends on the statistical significance (see Pearson 1996).

(details below). In other words, an edge in the graph between sequence A and B indicates that the corresponding proteins are likely to be related.

This graph has been constructed, using all currently known measures of similarity between protein sequences; Smith Waterman dynamic programming method (SW) (Smith and Waterman 1981), FASTA (Lipman and Pearson 1985) and BLAST (Altschul et al. 1990). These methods are in daily use by biologists, for comparing sequences against the databases. Though SW tends to give the best results on average, it is not uncommon that FASTA or BLAST are more informative (Pearson 1995). Therefore we chose to incorporate all three methods into our graph, to achieve maximum sensitivity [2].

An unusual amino acid composition of the query sequence may strongly bias the results of a search. A case in point are the effects of low complexity segments within sequences (Altschul et al. 1994). Therefore, we also consulted the results of BLAST following a filtering of the query sequence, to exclude low complexity segments, using the SEG program (Wooton and Federhen 1993).

The following sections contain a detailed description of the procedure of assigning weights to edges. The procedure starts by creating a list of neighbors for each sequence, based on all the three methods. In order to place all three methods on comparable numerical scales, a numerical normalization is applied first to all methods. Then, only statistically significant similarities are maintained in these lists. Finally, the weight of an edge is defined as the minimum associated to it by any of the three methods, to capture the apparently strongest relation.

Placing all methods on a common numerical scale

It is relatively easy to compare between scores that a particular method assigns to different comparisons. However, how does one compare between scores that are assigned by different methods? We performed the following calculation: Pick any protein, carry out an exhaustive comparison against the whole database and consider the highest scores in each of the methods. Now plot these values and compare two methods at a time. These scores show a remarkably strong linear relation in log-log scale (not shown), therefore by introducing a (usually small) correction factor, per each protein and per method, the three methods get scaled to a single reference line[3].

Defining the list of neighbors

It is, of course, very difficult to set a clear dividing line between true homologies and chance similarities. Expectation values below 10^{-3} can be safely considered significant and those above 10 reflect almost pure chance similarities. However, the midrange is difficult to characterize, and truly related proteins may have expectation values around 1. An overly strict threshold will miss important similarities within the twilight zone, whereas an excessively liberal criterion will create many false connections. The exact threshold for each method was set to best discriminate among related and unrelated proteins. Our choice is based on the overall distribution of distances over the entire protein space, as given by each of the three methods.

This is illustrated in Fig. 1, which shows the distribution of expectation values over the entire SWISSPROT database, for SW, FASTA, and BLAST. The graphs in Fig. 1 naturally suggest a threshold for each method. The distribution drawn in a log-log scale is nearly linear at low expectation values, but starts a rapid increase at a certain value.

The slope may be viewed as a measure for the geometric entropy ($\cong$ dimensionality) of neighboring sequences (Pisier 1989). The mild slope at low expectation values indicates a low entropy at the short range which seems fairly uniform throughout the space. The steep slope indicates a rapid growth in the number of sequences that are unrelated to the centering sequence (high expectation values). As viewed from the centering sequence, these sequences are essentially random, that indeed reach the entropy of sequences drawn uniformly at random from the space (maximum entropy). The entropy of the neighboring sequences does change from one centering sequence to another. This results in different slopes around each sequence, corresponding to the characteristic evolutionary divergence in the corresponding family.

In view of this discussion we set the threshold to the value where the slope rapidly changes. The thresholds

[2]FASTA is based on a restricted application of the rigorous SW algorithm and is usually being viewed as an approximation of SW, whose main advantage is its speed. However, with the goal of a better identification of remote homologies in mind, we used FASTA with the BLOSUM50 scoring matrix (Henikoff and Henikoff 1992), whereas SW and BLAST were used with the BLOSUM62 scoring matrix. Many similarities were reported exclusively by only a single method - in some cases as many as tens of hits per sequence, which were not detected by the other methods. In the future we intend to incorporate several matrices in order to cover a broader range of evolutionary distances.

[3]The differences between FASTA and SW are mostly due to the different scoring matrices that are being used, and can be corrected by multiplying the original score by the relative entropy of the two matrices (Altschul 1991). The differences between SW and BLAST may be due to approximations in estimating the parameters λ and K (Karlin and Altschul 1990). The underlying assumption in calculating these parameters is that the amino acid composition of the query sequence is close to the overall distribution. This assumption often fails, e.g. for low complexity segments. Moreover, these parameters are based on first order statistics of the sequence, the scoring matrix and the database. The corrections that are required to match SW and BLAST may be due to inaccurate approximations of the estimated parameters, or to higher order statistics of the sequence.

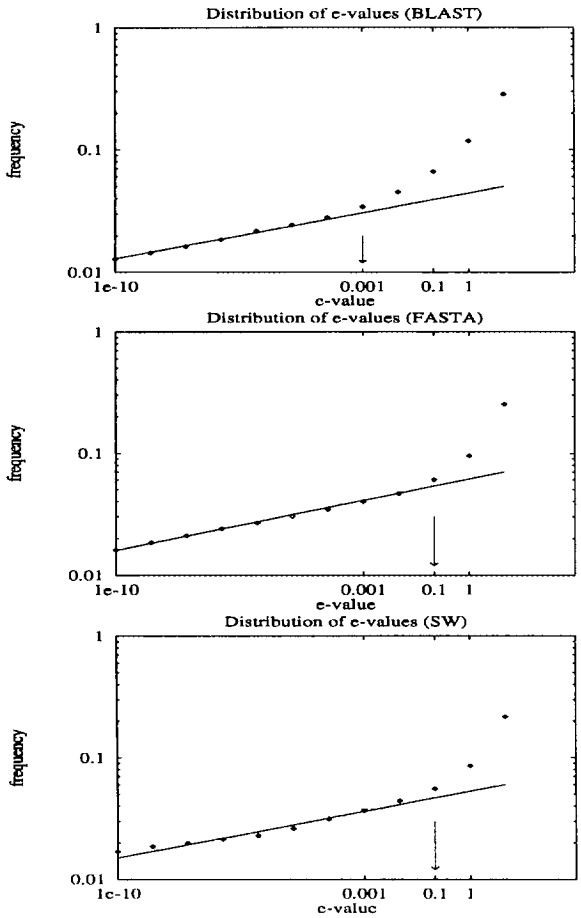

Figure 1: **Overall distribution of e-values according to the three main algorithms for sequence comparison.** a) BLAST b) FASTA c) SW. The distributions are plotted in a log-log scale. Note that the deviation from straight line starts earlier in BLAST, around 10^{-3}, whereas in FASTA and SW it starts only around 10^{-1}.

for SW, FASTA and BLAST are set at 0.1, 0.1 and 10^{-3} respectively[4]. An edge from vertex A to vertex B is maintained only if a significant score is obtained on comparing the corresponding proteins. Namely, if either SW or FASTA yield an expectation value ≤ 0.1 or BLAST's expectation value is $\leq 10^{-3}$.

A major difference between BLAST and SW/FASTA is that BLAST does not allow gaps[5] and therefore tends to overestimate the statistical significance of alignments. We counter this behavior of BLAST by the above asymmetry in selecting the edges. While this property may help BLAST reveal significant similarities that the other methods miss (e.g. Pearson 1995), we have to beware of highly fragmentary alignments that cannot be considered biologically meaningful. Therefore, we ignore those BLAST scores that come from a large number of HSPs (high scoring pairs), whereas the MSP (maximal segment pair) is insignificant[6].

Finally, even if the comparisons between proteins A and B fail to satisfy the previous criteria, the edge from A to B is maintained when all three methods yield an expectation value ≤ 1.

Obviously, it may happen that some of the similarities which lie around the thresholds are chance similarities. Our goal in designing the algorithm that is described next was to detect such similarities and eliminate them.

Exploring the connectivity

Having created this graph, we turn to explore it. We seek clusters of related sequences which can be assigned a characteristic biological function.

There are two major obstacles which should be considered: i) Multidomain proteins can connect two or more unrelated groups; ii) Overestimates of the statistical significance of the similarity scores. Naturally, chance similarities become more abundant as significance levels decrease.

Therefore, transitivity should be applied restrictively. By analogy, if transitivity is to be viewed as a force that attracts sequences, then it should be countered by some "rejecting forces" so that unrelated clusters be kept apart and prevent a collapse in the space of all protein sequences.

Our approach

Our starting point is reached by eliminating all edges of weight below a certain, very high, significance threshold (i.e. low expectation value). This operation splits our graph to many small components of strong connectivity. In biological terms, we split the set of all proteins into small groups of closely related proteins, which correspond to highly conserved subfamilies.

To proceed from this basic classification, we lower the threshold, in a stepwise manner, and take into account more relaxed statistical significant similarities. Doing this, several clusters of a given threshold may merge at a more permissive threshold. However, we closely

[4]While the self-normalized statistical estimates of FASTA and SW (Pearson 1998) are quite reliable, the statistical estimates of BLAST may be biased for low complexity sequences (see previous footnote). Consequently, filtering may significantly reduce the number of high scoring hits reported by BLAST. If we acknowledge only sequences that pass the filter, we may miss many relations of biological significance. Instead, a more stringent threshold is set in this case for BLAST at 10^{-6}.

[5]We haven't yet tested the new version of BLAST which does incorporate gaps in the alignment.

[6]Specifically, denote the number of HSPs and the MSP score by N_{HSP} and S_{MSP} respectively. The average and the standard deviation of N_{HSP} and S_{MSP} are calculated for high scoring sequences (μ_{HSP} , σ_{HSP} , μ_{MSP} and σ_{MSP} respectively). Those hits that are based on N_{HSP} which satisfy $N_{HSP} > \mu_{HSP} + \sigma_{HSP}$, with MSP score $S_{MSP} < \mu_{MSP} - \sigma_{MSP}$, and are not significant according to SW and FASTA, are ignored.

monitor this process and allow a merge only in view of strong statistical evidence for a true connection among the proteins in the resulting set. We turn to a detailed description of these two main steps:

Basic classification If all edges of weight below a certain significance threshold are eliminated, the transitive closure of the similarity relation among proteins splits the space of all protein sequences into connected components or **clusters**. These are proper subsets of the whole database wherein every two members are either directly or transitively related. These sets are maximal in this respect and cannot be expanded. Thus they offer a self-organized classification of all protein sequences in the database. We set the threshold at the very stringent significance level of 10^{-100}. Similarities which are reported as significant above the level of 10^{-100} are very conserved and stretch along at least 150 aa. Thus, no chance similarities occur at this level. We also do not encounter connections based on a chain of distinct common domains in multi-domain proteins. The resulting connected components can be safely expected to correlate with known conserved biological subfamilies.

Note that this is a *directed* graph, and hence is not necessarily symmetric. Specifically, it may (and does) happen that there is an edge from protein A to protein B, but none in the reverse direction. Furthermore, even if both edges exist, their weights usually differ. Therefore, our notion of a component is that of a *strongly connected component*[7]. The partition into strongly connected components is clearly more refined that the partition into connected components.

The clustering algorithm Our procedure is recursive. That is, given the classification at threshold T, we should give a method for deriving the classification at the next more permissive level ($T1 = 10^5 T$).

The algorithm runs in two phases. First we identify and mark groups ("pools") of clusters that are considered as candidates for merging (see Fig. 2a). A local test is performed where each candidate cluster is tested with respect to the cluster which "dragged" it to the pool.

To quantify the similarity of two clusters P and Q, we calculate the geometric mean of all pairwise scores of pairs one of which is in P and the other is in Q. Unrelated pairs are assigned the default evalue of 1. When the geometrical mean of the evalues is below $\sqrt{T}$ our interpretation is that P and Q are indeed related and that their connection does not reflect chance similarities or chain of domain-based connections[8]. We define the *Quality* of the $P-Q$ connection as the minus log of the geometric mean. It ranges between 0 and 100, and the higher it is the more significant is the connection.

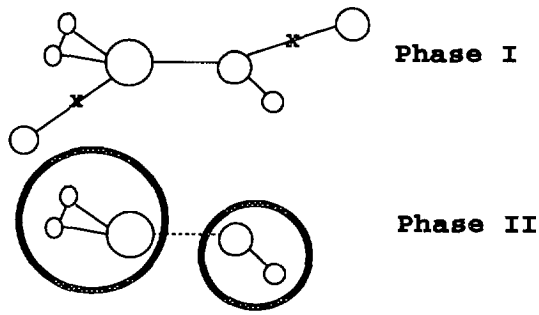

Figure 2: **The clustering algorithm.** Phase I) Identify pairs of clusters that are considered as candidates for merging. Decisions are made based on the geometric mean ("quality") of the pairwise scores of the connections between the two clusters. If this mean exceeds a specific threshold then the cluster is accepted as a candidate, and enters into the pool. Otherwise it is rejected (denoted by "X"). Phase II) Pairwise clustering is applied to identify **groups** of clusters which are strongly connected. At each step the two closest **groups** are chosen and merged provided that the geometrical mean exceeds the threshold. Otherwise they stay apart (denoted by dashed line). When merging group of clusters s with group of clusters s' we take into account the weighted geometric mean between all clusters in s and in s'.

At the second phase we carry out a variant of a pairwise clustering algorithm. This algorithm successively merges only pairs of clusters that pass the above test and are not suspected as representing chance or domain-based similarities. A detailed description of this algorithm appears in Fig. 2b.

This analysis is performed at different thresholds, or **confidence levels**, to obtain an hierarchical organization. The analysis starts at the 10^{-100} threshold. Subsequent runs are carried out at levels 10^{-95},10^{-90},10^{-85}, ... $10^{-0} = 1$.

Results

Nearly all of the clusters we found are biologically meaningful. Some of them correspond to well known families, but many others represent less studied families. Some clusters consist exclusively of unknown proteins or hypothetical proteins.

Needless to say, this overwhelming body of information cannot be properly surveyed in a single paper. We have constructed an interactive web site that contains the results of our analysis (http://www.protomap.cs.huji.ac.il), where users can get acquainted with this new map of the protein space. Here we can only offer a glimpse of this new map. Rather than discuss novel connections and relationships through specific examples, we highlight the novel tools and the main features of this map.

[7] A directed graph is *strongly connected* if for every two vertices there is a directed path from x to y as well as from y to x.

[8] Other thresholds were investigated as well. The present choice was supported by the logarithmic distribution of scores (Fig. 1).

Confidence level	Cluster size over 100	51-100	21-50	11-20	6-10	2-5	1	Total no of clusters
10^{-100}	8	18	90	234	528	3727	29870	34475
10^{-95}	8	19	100	240	537	3806	29086	33796
10^{-90}	8	20	111	256	545	3871	28224	33035
10^{-85}	8	23	119	262	563	4004	27189	32168
10^{-80}	8	25	133	264	594	4071	26140	31235
10^{-75}	9	31	132	275	623	4131	25051	30252
10^{-70}	10	34	138	293	653	4136	23943	29207
10^{-65}	11	32	156	309	660	4180	22911	28259
10^{-60}	13	34	171	319	677	4170	21772	27156
10^{-55}	15	40	178	334	676	4194	20646	26083
10^{-50}	15	51	184	350	676	4188	19463	24927
10^{-45}	17	53	197	362	696	4181	18282	23788
10^{-40}	21	54	203	383	714	4109	17129	22613
10^{-35}	23	53	213	393	760	4101	15801	21344
10^{-30}	26	53	232	415	774	4014	14428	19942
10^{-25}	29	57	252	421	788	3897	13191	18635
10^{-20}	32	64	263	436	779	3775	11839	17188
10^{-15}	35	64	270	464	808	3645	10620	15906
10^{-10}	38	76	293	457	802	3231	9112	14009
10^{-5}	51	92	315	431	684	2655	7169	11397
10^{-0}	51	94	315	456	703	2816	6167	10602

Table 1: **Distribution of clusters by their size at each confidence level.**

General information

Table 1 shows the distribution of cluster sizes at various confidence levels. At each level, the universe of all proteins splits into clusters, which merge to form larger and coarser sets as the confidence level decreases. Clearly, the number of isolated proteins (clusters of size 1) diminishes as well.

In contrast, an application of a single-linkage clustering algorithm resulted in an avalanche (not shown). Already at confidence level (10^{-20}) most of the space is divided by the single-linkage algorithm to a small number of very large clusters. This is the result of chance similarities and chains of domain-based connections that lead unrelated families to merge into few giant clusters. A major ingredient of our new algorithm is the choice of rules for avoiding such undesirable connections and preventing the collapse.

Table 2 shows the 40 largest clusters at the lowest confidence level (10^{-0}) that we consider. The description of each cluster is based mainly on the SWISSPROT annotations of its members [9]. This should be viewed only as a sample. For further information the reader is referred to our site.

[9]We do not have a fully automatic way for annotating all the clusters. The biological interpretation may require a substantial degree of biological insight. However, a simple census of proteins based on SWISSPROT definition/characterization usually gives a good indication of the cluster's type.

Possibly related clusters

The clustering algorithm automatically rejects many possible connections among clusters (see Fig. 2). Connections with quality below the threshold get rejected. However, many of these rejected connections are still meaningful and reflect genuine though distant homologies. We refer to the rejected mergers as *possibly related clusters.*

In examining a given cluster, much insight can be gained by observing other clusters which are possibly related to it. Even though some of these connections are justifiably rejected, in particular at the lowest level of confidence 10^{-0}, many others suggest structural or functional similarity, despite a weak sequence similarity. At this stage it is hard to give exact rules for evaluating these relations, and one's judgment must be used. Such judgment can also take into account the pairwise alignments of protein pairs, one from the cluster under study and the other in the possibly related cluster. The alignments can be found in the web site.

Table 3 is an illustrative example. It shows the clusters which are possibly related to cluster 4 (Immunoglobulin V region), ordered by quality value.

Table 4 shows the clusters which are possibly related to cluster 5 (Immunoglobulins and major histocompatibility complex). Clusters which we suspect to be unrelated are marked (in italic). One can validate the significance of "possibly related clusters" according their quality and to the alignments, and insignificant connections can be easily traced and ignored by a manual examination of the given alignments.

The clusters which are possibly related to cluster 5 consist mostly of proteins that adopted the Im-

Cluster number	size	family
1	718	Protein kinases
2	593	Globins
3	514	G-protein coupled receptors
4	330	Immunoglobulin V region
5	326	Immunoglobulins and major histocompatibility complex
6	318	Homeobox
7	315	Ribulose bisphosphate carboxylase large chain
8	284	ABC transporters
9	260	Zinc-finger C2H2 type
10	256	Calcium-binding proteins
11	252	Serine proteases, trypsin family
12	229	GTP-binding proteins - ras/ras-like family
13	221	Myosin heavy chain, tropomyosin, kinesins
14	208	Collagens, structural proteins
15	206	Cytochrome p450
16	198	GTP-binding elongation factors
17	196	Tubulins
18	190	Cytochrome b/b6
19	187	ATP synthases
20	172	Heat shock proteins
21	171	Alcohol dehydrogenases (short-chain)
22	171	Snake toxins
23	152	NADH-ubiquinone oxidoreductase
24	142	Bacterial regulatory components of signal transduction
25	141	DNA-binding proteins of HMG
26	140	Nuclear hormones receptors
27	139	Actins
28	139	Intermediate filaments
29	138	GTP-binding, ADP-ribosylation factors family
30	136	Neurotransmitter-gated ion-channels
31	133	Zinc-containing alcohol dehydrogenases
32	133	Cellular receptors, EGF-family
33	130	Amylases
34	130	Hemagglutinin
35	129	RNA-directed DNA polymerase
36	125	Chaperones, chaperonins
37	122	Phospholipase A2
38	120	Insulins
39	115	Cytochrome c
40	115	Ketoacyl synthase

Table 2: **Largest clusters at the lowest confidence level (10^{-0}).**

munoglobulin fold of the IG constant region. These clusters are disjoint from the clusters possibly related to cluster 4 that are involved in aspects of recognition in the immune system (via the variable regions). Cluster 1796, where both regions occur, is related to both. However, different parts of the proteins account for the different relationships. This information gives a view of the geometry of the protein space in the vicinity of the Immunoglobulin super-family (work in progress).

Possibly related clusters are also informative for the study of domains. By navigating through these clusters one finds many sequences which belong to different families, all sharing this domain. The list of related clusters can be viewed as a "soft clustering", where the same protein can participate in several different clusters.

The above examples involve only related clusters at the lowest level of confidence. We should note, however, that throughout the clustering process a pending pair of related clusters may later join upon lowering the level of confidence.

Tracing the formation of clusters

A major aspect of the hierarchical organization is that clusters of a given threshold may merge at a more permissive threshold. This reflects the existence of subfamilies within a family, or families within a superfamily.

We thus obtain a subdivision of clusters into smaller subsets as we proceed from one level to another. This is illustrated in Fig. 3 for the 2Fe-2S ferredoxins family. As we move from level 10^{-5} to level 10^{-10}, this cluster (consisting of 102 proteins) splits into flavin reductases, ferredoxins, flavohemoproteins and other sub-

Cluster number	Size	Quality	Number of connections	Family
1643	5	0.29	219	B-cell antigen receptor complex associated protein
927	10	0.11	193	T-cell surface glycoprotein CD4
2613	3	0.03	20	Polymeric-Immunoglobulin receptor
5	326	0.01	226	Immunoglobulins and major histocompatibility complex
1137	8	0.01	18	T-cell-specific surface glycoprotein CD28, cytotoxic T-lymphocyte protein
1189	8	0.01	9	Myelin P0 protein precursor
1796	5	0.01	9	Poliovirus receptor precursor

Table 3: **Clusters possibly related to cluster 4 (Level: 1e-0).** Clusters are sorted by quality (i.e. the minus log of the geometric average of connections). Note that all clusters belong to the super family of Immunoglobulins

Cluster number	Size	Quality	Number of connections	Family
1831	5	0.38	248	T-cell receptor gamma chain c region
4	330	0.01	226	Immunoglobulin V region
104	66	0.01	64	Cell adhesion molecules, myelin-associated glycoprotein precursor axonin-1 precursor, B-cell receptor CD22-beta precursor, and more
578	16	0.01	28	High affinity Immunoglobulin gamma FC receptor
596	*16*	*0.01*	*33*	*Recombination activating proteins, zinc finger, c3hc4 type*
856	*11*	*0.01*	*11*	*Cornifin (small proline-rich protein)*
1262	7	0.01	21	T lymphocyte activation antigen
1636	5	0.01	8	Basigin precursor
1796	5	0.01	7	Poliovirus receptor precursor

Table 4: **Clusters possibly related to cluster 5 (Level: 1e-0).** Only clusters with more than one member are shown. Clusters are sorted by quality as in table 3. Almost all clusters belong to the super family of Immunoglobulins. Probably unrelated clusters are 596 and 856 (in italic).

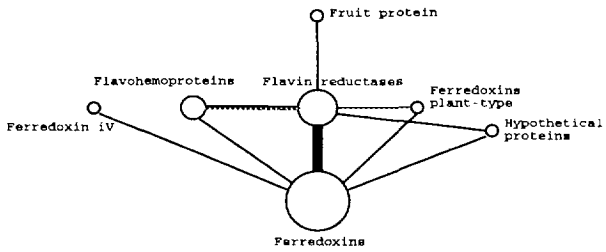

Figure 3: **The Ferredoxins family (cluster 47 at level 10^{-0}).** Each circle stands for a cluster at threshold = 10^{-10}. Circles' radii are proportionate to the cluster's size. The drawn edges appeared upon lowering the threshold to 10^{-5}. Edge widths are proportionate to the number of connections between the corresponding clusters.

families. Note that some clusters form cliques - where each cluster is connected to each other, thus validating the connection between each pair of clusters. To fully evaluate such subclassification, the graphical and sequence alignments of sequence pairs within clusters are available (at the web site).

Hierarchical organization within protein families and super families

Our hierarchical organization suggests a refined classification within known families. This classification is based on the information extracted while moving across the different levels of the tree (Scanning the hierarchy over **all** levels). This can be illustrated by the following simple example.

The small G-protein/Ras super family The ras gene is a member of a family of genes that have been found in tumor virus genomes and are responsible for the ability of the viruses to cause tumors in the cells they infect. In most cases this viral oncogene is closely related to a cellular counterpart (called proto-oncogene). Infection by a retrovirus that carries a mutant form of the ras gene (ras oncogene), or mutations, can cause cell transformation. Indeed, mutations in ras gene are linked to many human cancers.

The cellular ras protein binds guanine nucleotide and exhibits a GTPase activity. It participates in the regulation of cellular metabolism, survival and differentiation. In the last decade many additional proteins that are related to ras were discovered. They all share the guanine nucleotide binding site and are of 21-30 KDa in length. They are referred to as the small-G-protein super-family (Nuoffer and Balch 1994).

This family of proteins has several sub-families: ras, rab, ran, rho, ral, and smaller sub-families. Like to ras, these proteins participate in cell regulation processes, such as vesicle trafficking (rab) and cytoskeleton organization (rho). In figure Fig. 4 we depict the relations within this family, based on the hierarchical organization obtained by our analysis. Total of 229 proteins, all from the small G-protein super-family, are

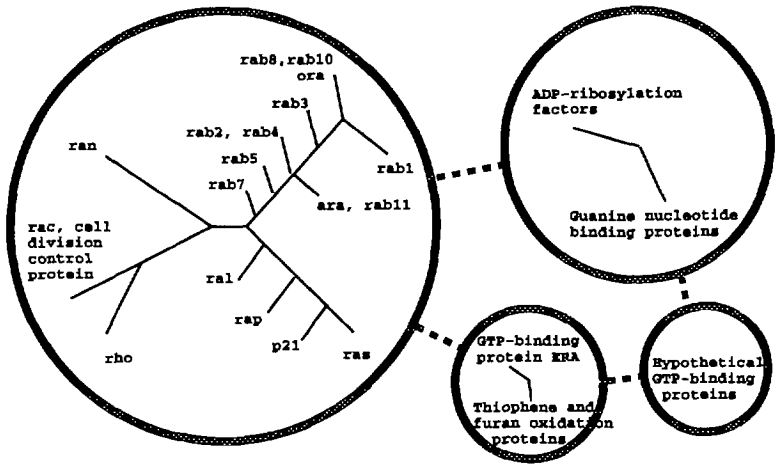

Figure 4: **The small G-protein family.** This family is composed of several sub-families. A total of 229 proteins, combined in cluster 12 (table 2), were grouped together into isolated sets at different levels of confidence, to form a natural sub-classification within the family. This hierarchical organization is much enriched by combining possibly related clusters. Clusters which are detected as related to this family are the cluster of the ADP-ribosylation factors family and guanine nucleotide-binding proteins and the cluster of the ERA proteins and thiophene and furan oxidation proteins.

presented. Small clusters, which correspond to subfamilies, are formed at the high levels of confidences, and fuse to larger clusters when the threshold is lowered. Clusters which are detected as related to this family are the ADP-ribosylation factors family and guanine nucleotide-binding proteins (total of 138 proteins) and GTP-binding protein ERA and thiophene and furan oxidation proteins (total of 14 proteins), all of which are GTP-binding proteins.

Discussion

This paper addresses the problem of identifying high order features within the sequence space. We aim at an exhaustive "charting" of all proteins and at sketching the map of the proteins space, based on pairwise similarities. This is a daunting task and many difficulties are encountered. One must begin from well established statistical measures, in order to identify significant similarities. Great caution and biological expertise are needed to eliminate connections which are unacceptable or misleading. The main culprits are chance similarities and multi-domain-based connections. The sheer volume of data makes it inevitable to seek automatic identification methods.

We tried to address these major obstacles, and obtain a hierarchical organization. This organization corresponds to a functional partitioning of all proteins. It reveals interesting relations between and within protein families, and provides a global view ("map") of the universe of all proteins.

Clustering proteins into functional groups would greatly benefit the understanding of protein function, structure and evolution. It would serve as a key tool for the analysis of new sequences, and the relationships among known proteins. When currently available means of comparison fail to give satisfactory information about the features of a protein, a global view may provide the missing clues.

How can a project such as this one be evaluated? Unfortunately, this area has no standard benchmarks on which to try one's algorithm. We suspect that other groups investigating this field would benefit from such a benchmark. Our results may be useful in this respect as well, in that we offer well-defined groups which can be used in testing and refining new algorithms and software tools.

For a comprehensive view of this project the reader is again encouraged to visit our site (http://www.protomap.cs.huji.ac.il).

Acknowledgments

We thank Hanah Margalit for many valuable discussions.

References

Altschul, S. F.; Carrol, R. J.; and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* 215: 403-410.

Altschul, S. F. (1991). Amino acid substitution matrices from an information theoretic perspective. *Journal of Molecular Biology* 219: 555-565.

Altschul, S. F.; Boguski, M. S.; Gish, W. G.; and Wooton, J. C. (1994). Issues in searching molecular sequence databases. *Nature Genetics* 6: 119-129.

Bairoch, A., and Boeckman, B. (1992). The SWISS-PROT protein sequence data bank. *Nucleic Acid Research* 20: 2019-2022.

Doolittle, R. F. (1992). Reconstructing history with amino acid sequences. *Protein Science* 1: 191-200.

Flores, T. P.; Orengo, C. A.; Moss, D.; and Thoronton, J. M. (1993). Comparison of conformational characteristics in structurally similar protein pairs. *Protein Science* 2: 1811-1826.

Harris, N. L.; Hunter, L.; and States, D.J. (1992). Mega-classification: Discovering motifs in massive datastreams. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, 837-842, AAAI press/The MIT Press, Menlo park/Cambridge.

Hilbert, M.; Bohm, G.; and Jaenicke, R. (1993). Structural relationships of homologous proteins as a fundamental principle in homology modeling. *Proteins* 17: 138-151.

Henikoff, S., and Henikoff, J. G. (1992). Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Science (USA)* 89: 10915-10919.

Karlin, S., and Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proceedings of the National Academy of Science (USA)* 87: 2264-2268.

Koonin, E. V.; Tatusov, R. L.; and Rudd, K. E. (1996). Protein sequence comparison at genome scale. *Methods Enzymol* 266: 295-321.

Linial, M.; Linial, N.; Tishby, N.; and Yona, G. (1997). Global self organization of all known protein sequences reveals inherent biological signatures. *Journal of Molecular Biology* 268: 539-556.

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

Murzin, A. G. (1993). OB(oligonucleotide/oligosaccharide binding)-fold: common structural and functional solution for non-homologous sequences. *EMBO Journal* 12(3): 861-867.

Needleman, S. B., and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* 48: 443-453.

Neuwald, A. F.; Liu, J. S.; Lipman, D. J.; and Lawrence, C. E. (1997). Extracting protein alignment models from the sequence database. *Nucleic Acid Research* 25(9): 1665-1677.

Nuoffer, C., and Balch, W. (1994). GTPase: multifunctional molecular switches regulating vesicular traffic. *Annual Review of Biochemistry* 63: 949-990.

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

Pearson, W. R. (1996). Effective protein sequence comparison. *Methods Enzymol* 266: pp 227-258.

Pearson, W. R. (1997). Identifying distantly related protein sequences. *Computer Applications in the Biosciences* 13,4: 325-332.

Pearson, W. R. (1998). Empirical statistical estimates for sequence similarity searches. *Journal of Molecular Biology* 276: 71-84.

Pennisi, E. (1997). Microbial genomes come tumbling in. *Science* 277: 1433.

Pisier, G. (1989). The volume of convex bodies and Banach space geometry. Cambridge University Press, Cambridge.

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

Smith, T. F., and Waterman, M. S. (1981). Comparison of Biosequences. *Advances in Applied Mathematics* 2: 482-489 .

Watanabe, H., and Otsuka, J. (1995). A comprehensive representation of extensive similarity linkage between large numbers of proteins. *Computer Applications in the Biosciences* 11(2): 159-166.

Wootton, J. C., and Federhen, S. (1993). Statistics of local complexity in amino acid sequences and sequence databases. *Computers in Chemistry* 17: 149-163.