PROTEINS: Structure, Function, and Genetics 37:360–378 (1999)

ProtoMap: Automatic Classification of Protein Sequences, a Hierarchy of Protein Families, and Local Maps of the Protein Space

Golan Yona,[1]* Nathan Linial,[1] and Michal Linial[2]
[1]*Institute of Computer Science, Hebrew University, Jerusalem, Israel*
[2]*Department of Biological Chemistry, Institute of Life Sciences, Hebrew University, Jerusalem, Israel*

ABSTRACT **We investigate the space of all protein sequences in search of clusters of related proteins. Our aim is to automatically detect these sets, and thus obtain a classification of all protein sequences. Our analysis, which uses standard measures of sequence similarity as applied to an all-vs.-all comparison of SWISSPROT, gives a very conservative initial classification based on the highest scoring pairs. The many classes in this classification correspond to protein subfamilies. Subsequently we merge the subclasses using the weaker pairs in a two-phase clustering algorithm. The algorithm makes use of transitivity to identify homologous proteins; however, transitivity is applied *restrictively* in an attempt to prevent unrelated proteins from clustering together. This process is repeated at varying levels of statistical significance. Consequently, a hierarchical organization of all proteins is obtained.**

The resulting classification splits the protein space into well-defined groups of proteins, which are closely correlated with natural biological families and superfamilies. Different indices of validity were applied to assess the quality of our classification and compare it with the protein families in the PROSITE and Pfam databases. Our classification agrees with these domain-based classifications for between 64.8% and 88.5% of the proteins. It also finds many new clusters of protein sequences which were not classified by these databases. The hierarchical organization suggested by our analysis reveals finer subfamilies in families of known proteins as well as many novel relations between protein families. Proteins 1999;37:360–378. © 1999 Wiley-Liss, Inc.

Key words: clustering; protein families; protein classification; sequence alignment; homologous proteins

INTRODUCTION

In recent years we have witnessed a massive flow of new biological data. Large-scale, world-wide sequencing projects reveal new sequences, and many sequences that are added to the databases are unannotated and await analysis. Currently, 15 complete genomes (yeast, *C. elegans, Escherichia coli,* other eubacteria, and several archaea) are known. Between 35% and 50% of their proteins have not been assigned a function yet.[1,2] In the absence of structural data, analysis necessarily starts with the sequence. The most effective analyses compare the sequence under study with all known sequences, in search for close relatives that may have been assigned a function. In this way, properties of a new protein sequence are extrapolated from those of its neighbors.

Since the early 1970s, algorithms were developed for comparing protein sequences efficiently and reliably.[3–7] But even with the best alignment of two protein sequences, the basic question remains: Do they share the same biological function or not? It is generally accepted that two sequences with over 30% identity along much of the sequences are likely to have the same three-dimensional structure or fold.[8–11] Proteins of the same fold often have similar biological functions. Nevertheless, one encounters many cases of high similarity both in fold and function that is not reflected in sequence similarity.[11–13] Such cases are missed by current search methods that just compare sequences.

Detecting homology may often help in determining the function of new proteins. By definition, homologous proteins have evolved from the same ancestor protein. The degree of sequence conservation varies among protein families. However, homologous proteins almost always have the same fold.[14] 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 leads to many pitfalls. Although the common evolutionary origin of two proteins is almost never directly observed, we can deduce homology, with a high statistical confidence, given that the sequence similarity is significant. This is particularly useful in the so-called twilight zone,[15] where sequences are identical with, say, 10–25%. Transitivity can be used to detect related proteins, beyond the power of a direct search.

Although transitivity is an attractive concept, it has its perils: Similarity is not transitive, and similarity does not

Grant sponsor: Israel Academy of Sciences; Grant sponsor: Horovitz Foundation of Yissum, The Hebrew University; Grant sponsor: Program in Mathematics and Molecular Biology.
*Correspondence to: Golan Yona, Department of Structural Biology, Fairchild Building D-109, Stanford University, CA 94305. E-mail: golan@gimmel.stanford.edu
Received 9 March 1999; Accepted 28 June 1999

necessarily imply homology.† Therefore, similarity should be used carefully in attempting to deduce homology. 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 inadequacy of single-linkage clustering for the purpose of identifying protein families within the sequence space.

Expert biologists can distinguish significant from insignificant similarities. However, the sheer size of current databases rules out an exhaustive manual examination of all potential homologies. Our goal here is to develop an automatic method for classification of protein sequences based on sequence similarity, through the detection of groups of homologous proteins (clusters) and high-level structures (groups of related clusters that are connected by weak but consistent sequence similarities) within the sequence space. Such organization would reveal relationships among protein families and yield deeper insights into the nature of newly discovered sequences.

Related Works—Large-Scale Analyses of Protein Sequences

Transitivity of homology has been used before, and the power of transitivity in inferring homology among distantly related proteins has been demonstrated in refs. 13 and 16–23. Some of these works have also addressed the perils of transitivity.[17,18,21]

To properly evaluate the present study, it is important to place it in the context of other large-scale analyses of protein sequences. This has been an active research field since the early 1990s. Several different approaches have been tested. In general, these studies are divided into two categories: those focused on finding significant motifs, patterns, and domains within protein sequences and those that apply to complete proteins.

Motif- and domain-based analyses

Most of these studies yielded databases of protein motifs and domains. Such databases have become an important tool in the analysis of newly discovered protein sequences. Among these are ProDom,[18] Pfam,[23] PROSITE,[24] PRINTS,[25] Blocks,[26] Domo,[27] and SMART.[28] The manually defined patterns in PROSITE have served as an excellent seed for several such studies.

There are several aspects in which these studies differ from each other. Some are based on manual or semimanual procedures (e.g., PROSITE, PRINTS), others are generated semiautomatically (Pfam), and the rest are generated fully automatically (e.g., ProDom, Blocks, Domo). Some focus on short motifs (PROSITE, PRINTS, Blocks), whereas others seek whole domains and try to infer domain boundaries (Pfam, ProDom, Domo). Most databases also give the domain/motif structure of proteins. Two databases make use of transitivity to enhance sensitivity. ProDom applies the transitive closure of high-scoring segments pairs obtained by BLAST (when the common segments overlap above a minimum overlap parameter). In the Pfam database, the construction of new families starts from an HMM model derived from multiple alignment of related proteins, which is then improved iteratively by searching for further related sequences in the database. These sequences are iteratively incorporated into the model, until the process converges. After each iteration the alignment is checked manually to avoid misalignments.

Protein-based analyses

Most studies in this category draw directly on pairwise comparison.[16,17,19,20,29–31] All these works cluster the input database, using transitive closure of similarity scores (i.e., single linkage clustering). Among these works, three[17,29,30] have addressed the problem of multidomain proteins. Harris et al.[17] allow groups to merge only if they share k overlapping regions. However, they concluded that $k = 1$ is the best choice for highest accuracy. Thus, their clustering procedure essentially remains a single-linkage clustering (in multidomain proteins, regions are classified to multiple classes). In the second study,[29] clusters are created starting from triangles formed by three homologous proteins from different species. Triangles which share an edge are merged (this requirement reduces the probability that unrelated clusters merge). An additional (manual) step is carried to split clusters that are incorrectly merged owing to multidomain proteins. The third study[30] classifies proteins into families based on global similarities and into homology domains based on local similarities. In this study sequences are classified into families and superfamilies based on similar overall architecture (same domains in the same order if the sequences belong to the same family; more flexibility is allowed if the sequences belong to different families within the same superfamily). Homology domains are defined using multiple alignments of homologous segments (identified based on local similarities). Both classifications depend on semiautomatic procedures and careful manual inspection.

Several other studies of complete proteins employed alternative representations of protein sequences, e.g., their dipeptide composition[32,33] or combination of compositional properties and other physical/chemical properties.[34] These representations induced measures of similarity/dissimilarity between complete protein sequences, which were used to classify the sequences into a fixed number of clusters,[32,33] or to search for close relatives.[34]

Our approach

The important role of motifs and domains in defining a protein's function is unquestionable: Detecting a known motif within a new protein sequence can help reveal its function and lead to the correct assignment of the new sequence to an existing protein family. Indeed, domain-based studies have added much to our knowledge. The

†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 ref. 14).

domain-based databases usually offer much biologically valuable information about domains and the domain structure of proteins, through multiple alignments and schematic representations of proteins. Therefore, function prediction procedures should include a domain-based protocol, in which the sequence is scanned first for known domains.

However, in many cases, characterizing a new protein only by its domain content is insufficient. This happens, for example, when no known domains are apparent in the new protein. In some instances, only a few related sequences are available, too few to define a reliable prototype signature or a profile of the common domains. Therefore, a proper analysis of a new protein sequence should incorporate comparisons against domain-based databases, as well as sequence databases. In this view, an analysis which identifies groups of related proteins in databases of protein sequences is invaluable. It may amplify the outcomes of a database search. When close hits are already grouped together based on mutual similarity, this may highlight a similarity with a group which could otherwise be missed by a simple manual scanning. Moreover, if several groups are found related to the query sequence, this may indicate the existence of several distinct functional/structural domains. If all groups share the same region of similarity, this adds insights about the relations between the different groups. In some cases, this may suggest that the groups belong to the same family or superfamily. Hence, protein-based analyses are important as a complementary tool for sequence analysis.

Our study attempts to define a classification of whole protein sequences. It draws on pairwise similarities and seeks strongly connected sets of proteins. It applies a moderate version of transitive closure, in an attempt to eliminate chance similarities and avoid indirect multiple-domain–based connections.

In our analysis, the protein space is represented as a weighted graph whose vertices are the sequences. The weight of an edge between two sequences corresponds to their degree of similarity. Clusters of related proteins correspond to strongly connected sets of vertices in this graph. To detect these sets, we begin with a very strict, high-resolution classification that employs only connections of very high statistical significance. The resulting clusters are then merged to form bigger and more diverse clusters. The algorithm operates hierarchically: Each step adds new weaker connections to the previously considered connections. A statistical test is applied in order to identify and eliminate problematic connections as well as possibly false connections between unrelated proteins. The result of our study is thus a hierarchical organization of all known protein sequences. The classification uses only standard similarity scores and does not depend on further biological information.

The method described here is applied to the set of all SWISSPROT[35] sequences and yields an exhaustive classification. This leads to the definition of a new pseudo-metric on the space of all protein sequences. In most cases, this measure turns out to be more sensitive than the existing measures from which it was derived. Such measures are important for a global self-organization of all protein sequences, as discussed in refs. 36 and 37.

METHODS

This section contains a description of our computational procedure. The procedure was carried out on the SWISS-PROT database[35] release 33, with a total of 52,205 proteins and 18,531,385 amino acids.

The Graph

We represent the space of protein sequences as a directed graph, whose vertices are the protein sequences. An edge between two vertices is weighted to reflect the dissimilarity between the corresponding pair of sequences, i.e., a high similarity translates to a small weight. To compute the weight of the directed edge from A to B, one compares sequence A against all sequences in the SWISS-PROT database including sequence B and obtains a distribution of scores. The weight is taken as the expectation value (e-value) of the similarity score found for A and B, based on this distribution.[38] This is an estimate for the number of occurrences that the appropriate score could have been obtained by chance, i.e., when compared with random sequences drawn from the same background distribution (usually defined as the distribution of amino acids overall the database). A low expectation value reflects a significant, strong connection, whereas a high expectation value reflects an insignificant, weak connection. Not all edges are retained in the graph as edges of statistically insignificant similarity scores are discarded (details below). In other words, in the final graph, an edge between sequences A and B indicates that the corresponding proteins are likely to be related.‡

This graph is constructed using the common algorithms for protein sequence comparison: Smith-Waterman dynamic programming method (SW),[4] FASTA,[5] and BLAST.[6] The SW algorithm was run with the BLOSUM62 matrix[39] and gap penalties of $-12,-2$ using either the Bioccelerator hardware[40] or the **ssearch** program which is part of the W. Pearson's FASTA 2 package. FASTA was run using the **fasta** program with the BLOSUM50 matrix[39] and gap penalties $-12,-2$ (the default setting). Both **ssearch** and **fasta** calculate expectation values based on empirically derived distribution of scores[41] (the Bioccelerator applies the same procedure for assessing the significance of results as in **ssearch**). The BLAST algorithm was also run with the BLOSUM62 matrix using the **blastp** 1.4.9 program available from the NCBI ftp site.* The program reports similarity scores along with the probability (P value) that

‡Within the scope of this work we used the expectation values as are, and weights reflect statistical significance rather than distance (intuitively, the term distance can be used instead of expectation value, to indicate that two proteins are either close or far, but practically, no metric is defined).

*We are aware of the new version of BLAST which also accounts for gaps and gives a very good approximation to the SW algorithm, while being much faster.[7] However, the old version of BLAST occasionally detects similarities that are missed by SW (e.g., for the Glucagon precursor family, and the H^+-transporting ATP synthase[42]). We have not yet tested whether the new version of BLAST preserves this merit.

the scores could have occurred by chance. Blastp probabilities are transformed to expectation values by the formula e-value = log $1/1-P$ value (see manual). All 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, especially when combined with different scoring matrices.[42] Therefore we chose to incorporate all three methods into our graph to achieve maximum sensitivity (indeed, many similarities were reported exclusively by only one of the three methods—in some cases as many as tens of hits per sequence, which were not detected by the other methods).

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 three methods. In order to place the expectation values for all three methods on comparable scales, a numerical normalization is determined and applied. Then, only statistically significant similarities are maintained in these lists. Finally, the weight of an edge is defined as the minimum expectation value associated to it by any of the three methods, so as to capture the strongest relationship (recall that edge weights represent expectation values, so small values indicate high similarity).

Placing all Methods on a Common Numerical Scale

It is relatively easy to compare scores that a particular method assigns to different comparisons. However, how does one compare 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 against one another for two methods at a time. These scores show a remarkably strong linear relation on a log–log scale (Fig. 1); therefore, by introducing a (usually small) correction factor, per each protein and per method, the three methods get scaled to a single reference line. Because SW e-values tend to be more reliable (see next section), they were chosen as a baseline. FASTA e-values and BLAST e-values were correlated with the SW e-values, and were corrected accordingly.

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.[43] This resulted in decreasing the expectation values (increasing significance) reported by FASTA. The differences between SW and BLAST may be due to approximations in estimating the parameters λ and K.[44] In general, our procedure resulted in increasing the expectation values (decreasing significance) of hits reported by BLAST.

Neighbors' Lists

It is, of course, very difficult to set a clear dividing line between true homology and chance similarity. An expectation value below 10^{-5} indicates that a false match would occur once in 100,000 searches and can be safely considered significant. On the other hand, an expectation value above 10 reflects mostly pure chance similarities. However, the mid-range is more difficult to characterize, and homologous proteins can have expectation values around 1. An overly strict threshold will miss important similarities within this "twilight zone," whereas an excessively liberal criterion will create many false connections. The exact threshold for each pairwise comparison method was set to best discriminate among related and unrelated proteins. Our choice is based on the overall distribution of expectation values over the entire protein space, as given by each of the three methods (Fig. 2).

The distribution shown may be thought of as the average distribution of expectation values for a "typical" protein sequence as a query. The distribution drawn on a log–log scale is nearly linear at low expectation values (where pairs of related sequences dominate), but starts to rise rapidly at a certain value. The steep slope at high

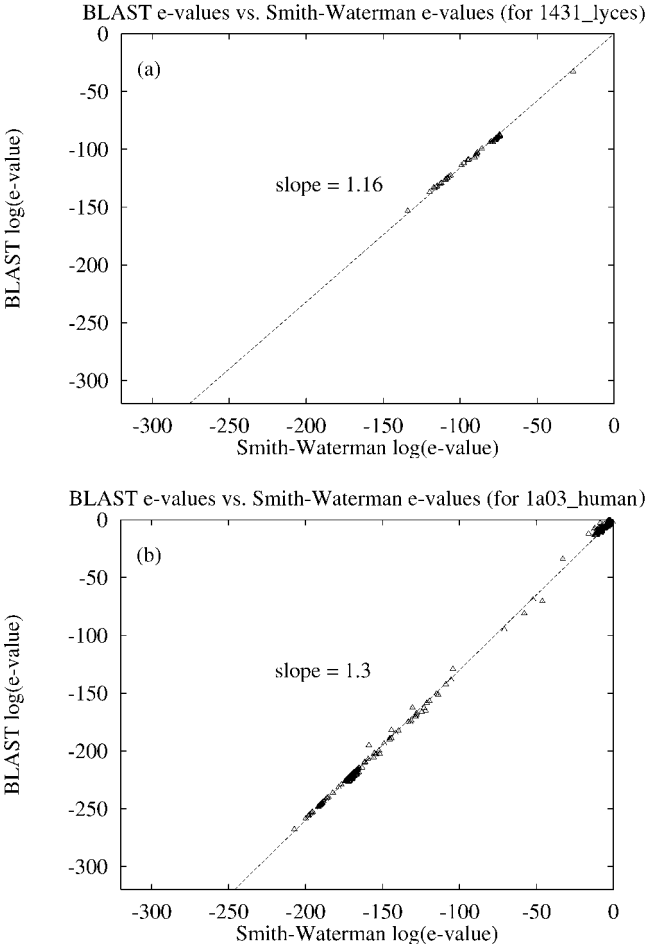

Fig. 1. Correlation of BLAST e-values and Smith-Waterman e-values. (**a**): BLAST e-values of neighboring sequences of 1431_lyces (P42651) vs. the SW e-values of the same neighbors. The graph is plotted in log–log scale. Note the strong linear correlation between the scores assigned by the two methods, where the slope is 1.16, i.e., $evalue_{BLAST} = (evalue_{SW})^{1.16}$. (**b**): BLAST e-values of neighboring sequences of 1a03_human (P04439) vs. the SW e-values of the same neighbors. The slope is 1.3 in this case.

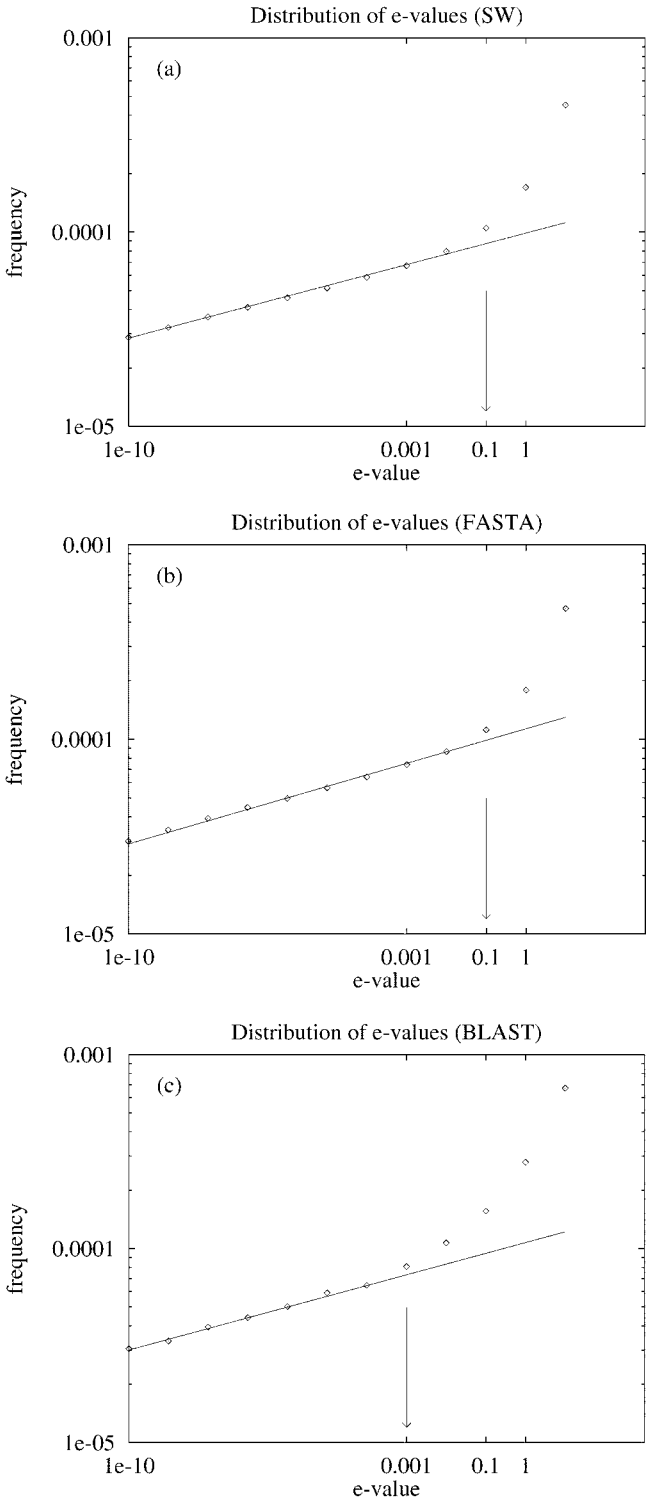

Fig. 2. Overall distribution of e-values according to the three main algorithms for sequence comparison. (**a**): SW. (**b**): FASTA. (**c**): BLAST. The distributions are based on the neighbors lists of *all* protein sequences in the SWISSPROT database and are plotted in a log–log scale. Frequency is the relative number of pairwise similarity scores with e-value that is equal to the value shown on the x-axis. 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}.

expectation values indicates a rapid growth in the number of sequences that are unrelated to the query sequence. Although the distribution may differ from one sequence to another, there is not enough data to deduce a reliable threshold for each sequence. Only when the distributions are averaged is the derived threshold reliable.

In this view, we set the threshold at the value at which the slope rapidly changes. The thresholds for SW, FASTA, and BLAST are set at 0.1, 0.1, and 10^{-3} respectively. An edge from vertex A to vertex B is maintained only if a significant score is obtained by any of the three methods used to compare the corresponding sequences—namely, if either SW or FASTA yields an expectation value of ≤ 0.1 or BLAST's expectation value is $\leq 10^{-3}$.

Although the self-normalized statistical estimates of FASTA and SW[41] are quite reliable (see also ref. 11), the statistical estimates of BLAST may be effected by the amino acid composition of the query sequence, and an unusual composition (e.g., low-complexity segments within the sequence) may bias the results of a search.[38] Therefore, we also generated the results using BLAST following a filtering of the query sequence, to exclude low-complexity segments (filtering was carried out with the SEG program[45]). On the other hand, many relations of biological significance can be missed if only sequences that pass the filter are to be considered. Therefore, this was handled with the more stringent BLAST threshold of 10^{-6}.

A major difference between BLAST and SW/FASTA is that BLAST does not include gaps in the alignments. BLAST detects similarities based on one or more high-scoring segment pairs (ungapped local alignments). Significance is assessed by applying Poisson or sum statistics.[38] Consequently, since gaps are ignored, BLAST tends to overestimate the statistical significance of fragmented alignments. We counter this behavior of BLAST by the above asymmetry in selecting the thresholds (Fig. 2). Although this property may help BLAST reveal significant similarities that the other methods miss (e.g., Pearson 1995[42]), we have to beware of highly fragmented alignments that cannot be considered biologically meaningful. Therefore, we ignore those BLAST scores that come from a large number of HSPs (high scoring pairs), when the MSP (maximal segment pair) is insignificant.

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 of ≤ 1.

This procedure is designed to screen most of the chance similarities in the neighbors list of each protein sequence. Unfortunately, chance similarities may occasionally pass our criteria. A major goal of the algorithm that is described next is to detect such similarities and eliminate them.

Exploring the Connectivity

We now turn to explore this graph. We seek clusters of related sequences which hopefully have a characteristic biological function. There are two major obstacles which should be considered: 1) Multidomain proteins can create undesired connection among unrelated groups; 2) Overes-

timates of the statistical significance of similarity scores may bias our decisions; chance similarities become more abundant as significance levels decrease. Therefore, transitivity should be applied restrictively. If transitivity is to be viewed as a force that attracts sequences, then it should be countered by some "repulsive force" to keep unrelated clusters apart and prevent collapse of the protein space.

Our approach

Our clustering procedure starts by eliminating all edges of significance below a certain, very high, significance threshold (i.e., very low expectation value). This operation splits the graph to many small components. In biological terms, we split the set of all proteins into numerous small groups of closely related proteins, which correspond to highly conserved subfamilies. To proceed from this basic highly restrictive classification, we lower the significance threshold (increase the expectation value) in a stepwise manner and gradually take into account similarities of lower statistical significance. In so doing, several clusters of a given threshold may merge. The process is closely monitored, and a merge is allowed only when strong statistical evidence is found for a true connection among the proteins in the resulting set. Detailed description of these two main steps follows.

Note that the graph is *directed,* 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, in a preliminary step, this graph is transformed into an undirected graph, by replacing the directed edges from A to B (with weight ω_1) and from B to A (with weight ω_2), with one undirected edge whose weight is defined as the maximum of ω_1 and ω_2. If there is only one directed edge, then in the new graph it is discarded.

Basic classification. If all edges of significance below a certain threshold are eliminated, the transitive closure of the similarity relation among proteins splits the space of all protein sequences into connected components or *clusters.* The transitive closure is equivalent to a single-linkage clustering and the resulting clusters 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 initially set the threshold at the very stringent significance level of 10^{-100}, and all edges with e-value $>10^{-100}$ are discarded. The remaining edges reflect similarities which correspond to highly conserved and relatively long regions (e.g., over 95% identity along at least 150 amino acids). Thus, neither chance similarities nor transitive chaining based on distinct common domains in multidomain proteins occur *at this level.* The resulting connected components can be safely expected to correlate with known highly conserved biological *subfamilies.*

The clustering algorithm. Our procedure is recursive. That is, given the classification at threshold T, we give a method for deriving the classification at the next more permissive level, that is $10^5 \times T$.

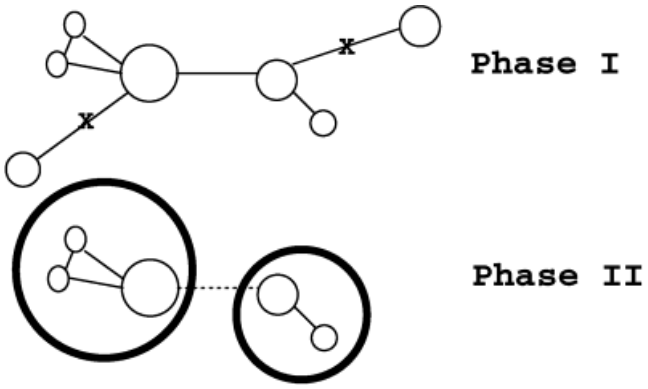

Fig. 3. The clustering algorithm. **Phase I:** Identify pairs of clusters that are considered as candidates for merging. Decisions are made based on the geometric mean of the pairwise scores of the connections between the two clusters. If this mean exceeds a specific threshold then the cluster is accepted as a candidate and enters a pool of candidates. Otherwise it is rejected (denoted in the figure 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 from the pool and merged provided that the quality of the connection exceeds the threshold (see text). Otherwise they stay apart (denoted by dashed line).

The algorithm runs in two phases. First we identify and mark groups ("pools") of clusters that are considered as candidates for merging (see Fig. 3). A local test is performed in which each candidate cluster is tested with respect to the cluster which "dragged" it to the pool, to check their degree of similarity.

To quantify the similarity of two clusters P and Q, we calculate the geometric mean of all pairwise scores between sequences in P and Q. Unrelated pairs are assigned the default (insignificant) e-value of 1. The geometric mean considers the distribution of all pairwise connections between the two clusters, so that random or unusual connections have little effect. The geometric averaging may equivalently be considered as (arithmetically) averaging the logarithms of the similarity scores. When the geometric mean of the e-values is below $\sqrt{T}$ (more significant) our interpretation is that P and Q are indeed related and that their connection reflects a genuine similarity. This threshold was chosen to obtain a sublinear decrease (since pairwise similarities are more frequent as confidence level decreases). It gave better results than other schemes and prevented a quick collapse of the whole space into a few huge clusters. The level of confidence in the reliability of the connection clearly decreases as T increases. We define the *Quality* of the P to Q connection as minus the log of the geometric mean. This quantity ranges between 0 and 100 and is higher for more significant connections.

In 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 thus not suspected of representing chance similarities). At each step the two closest clusters are chosen on the basis of the

quality of their connection and merged if their similarity (as quantified above, and based on all pairwise similarities between sequences within the new formed clusters) is more significant than the threshold. The process stops when the similarity of the next closest clusters do not pass the threshold (Fig. 3).

Any rejected merge is marked for further biological analysis. We refer to these rejected merges as *possibly related clusters,* and we comment on them in the next sections.

This analysis is performed at different thresholds, or *confidence levels,* to obtain a hierarchical organization. The analysis starts at the 10^{-100} threshold. Subsequent runs are carried out at levels 10^{-95}, 10^{-90}, 10^{-85}, 10^{-80}. . . . The process terminates at the threshold of $10^{-0} = 1$. Above the threshold of 1 almost all similarities are in fact chance similarities (see previous section).[§]

RESULTS

In the following sections we give brief general information about clusters and the results of the overall assessment of our classification in comparison with several well-known domain-based databases. We also offer a glimpse of our classification, through a few specific examples. Needless to say, the overwhelming body of information provided by our classification cannot be properly surveyed in a single report. 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 "map" of the protein space.

General Information

Table I shows the distribution of cluster sizes at various confidence levels. At each level, the set of all proteins splits into clusters, which merge to form larger and coarser clusters as the confidence level decreases. In particular, the number of isolated proteins (clusters of size 1) diminishes as well. At the lowest significance level ($10^{-0} = 1$) we have 10,602 clusters, of which 4435 contain at least two members. One thousand six clusters have size 10 and above.

The number of clusters (of size bigger than 1) at each level of confidence ranges between 4,228 and 5,543. In the attempt to evaluate these results, we ask: How many clusters should there be in the "ultimate" classification of all proteins? Are the numbers that we see here are close to this figure? A lower bound for this number is provided by the number of different folds, since we expect members of the same cluster to have similar folds. The current estimates place the total number of folds (known and unknown) between several hundreds and few thousands.[16,46–48] However, the same fold is usually adopted by few different superfamilies which share little or no significant sequence similarity. Therefore, a sequence-based classification would probably place these superfamilies in different (although possibly related) clusters. Moreover, superfamilies may consist of several families, sometimes with only a few percent sequence identity. Consequently, these families may be classified into different clusters, depending on the sensitivity of the method. On the average, each fold is adopted by two to three protein families.[49] Thus the total number of clusters is expected to exceed the number of folds, but it is likely to be less than an order of magnitude bigger. A number of clusters in the 4,000–5,000 range seems consistent with this estimation.[#]

Most of these clusters are biologically meaningful. Table II shows only the 50 largest clusters at the lowest confidence level (10^{-0}). The description attached in Table II to each cluster is based mainly on SWISSPROT annotations. [We do not have a fully automatic method for annotating all the clusters. A proper biological interpretation requires a substantial degree of biological sophistication and insight. However, a simple census of proteins based on SWISSPROT definition usually gives a good indication of the cluster's nature.] The table should be viewed only as a sample. For further information, the reader is referred to our web site. Henceforth, unless otherwise stated, cluster numbers refer to significance level 10^{-0}.

It should be noted that since our analysis concerns complete proteins, and it is not limited only to those subsequences which are identified as functionally or structurally important motifs and domains, not all the emerging clusters are correlated with a specific domain (e.g., clusters 23, 33). However, many of the clusters we encounter are characterized by a domain that is common to many or all member proteins, e.g., cluster 6 (homeobox domain) and cluster 9 (zinc finger). Some clusters exclusively consist of unknown proteins or hypothetical proteins (e.g., clusters 563, 606).

Performance Evaluation

It is very hard to evaluate the validity of classifications that emerge from a large-scale study of protein sequences, in that no generally accepted standards have been set yet in this field. Thus, new classifications are traditionally compared with what is considered a state of the art characterization of protein sequences, namely, the PROSITE dictionary of signature patterns, motifs, and domains.[24] For domain-based studies, a comparison with the manually derived PROSITE dictionary is inevitable and is essential for testing the biological significance of the

[§]The expectation values that we used are taken from the output of fasta, ssearch, and blastp, and are defined in the context of a single search against the database. Because we repeated this procedure 52,205 times (the number of proteins in the SWISSPROT 33 database), one may suggest correcting the expectation values and dividing them by 52,205. However, because hits were already screened to exclude most chance similarities (see "Neighbors' lists") and thresholds were defined based on the original e-values, and in order to fit the output of search programs such as **fasta, ssearch,** and **blastp,** the original e-values were kept.

[#]In protein-based analyses, further aggregation may be observed due to coupling of different structural units in the same class. However, the estimated number of clusters is not expected to change much: Out of 2984 ProtoMap clusters with at least three proteins, 1,128 clusters include at least one PROSITE domain and are associated with 1.6 PROSITE domain types on the average (the other 1,856 clusters do not contain PROSITE domains). Therefore, the estimated numbers of protein-based clusters and domain/fold based clusters are expected to agree up to a factor of 2.

TABLE I. Distribution of Clusters by Their Size at Each Confidence Level

Confidence level	Cluster size							Total number of clusters
	>100	51–100	21–50	11–20	6–10	2–5	1	
10^{-100}	8	18	90	234	528	3727	29870	34,475
10^{-95}	8	19	100	240	537	3806	29086	33,796
10^{-90}	8	20	111	256	545	3871	28224	33,035
10^{-85}	8	23	119	262	563	4004	27189	32,168
10^{-80}	8	25	133	264	594	4071	26140	31,235
10^{-75}	9	31	132	275	623	4131	25051	30,252
10^{-70}	10	34	138	293	653	4136	23943	29,207
10^{-65}	11	32	156	309	660	4180	22911	28,259
10^{-60}	13	34	171	319	677	4170	21772	27,156
10^{-55}	15	40	178	334	676	4194	20646	26,083
10^{-50}	15	51	184	350	676	4188	19463	24,927
10^{-45}	17	53	197	362	696	4181	18282	23,788
10^{-40}	21	54	203	383	714	4109	17129	22,613
10^{-35}	23	53	213	393	760	4101	15801	21,344
10^{-30}	26	53	232	415	774	4014	14428	19,942
10^{-25}	29	57	252	421	788	3897	13191	18,635
10^{-20}	32	64	263	436	779	3775	11839	17,188
10^{-15}	35	64	270	464	808	3645	10620	15,906
10^{-10}	38	76	293	457	802	3231	9112	14,009
10^{-5}	51	92	315	431	684	2655	7169	11,397
10^{-0}	51	94	315	456	703	2816	6167	10,602

results. However, when the analysis extends beyond regions which are known or suspected as domains, no standard benchmark exists to assess the quality of the results. One may then resort to comparisons against domain-based databases. Obviously, this may bias the assessment, a fact that should be kept in mind when evaluating the results.

To estimate the quality of our classification, we compared it with two well-established domain-based databases: PROSITE and Pfam.

The evaluation methodology

Given a reference classification A and a new classification B of the same set X, we evaluate the quality of the classification B in terms of the reference classification A, as reflected by their mutual agreement.¶ We consider four such indices of quality.

Gracy and Argos[27] have proposed a procedure for such a performance evaluation: Each class $a \in A$ is associated with the group $b \in B$ which maximizes the quantity $tp - fp - fn$. Here the number of true-positives (tp) is given by $|a \cap b| - 1$, that of false-positives (fp) is given by $|b \setminus a|$, and that of false-negatives (fn) is given by $|a \setminus b|$. Quality is defined by the percentage of the true positives $100 \cdot \mathrm{tp}/\mathrm{tp}+\mathrm{fp}+\mathrm{fn}$. In the same way, the percentage of false-positives and false-negatives are calculated. We call this index Q_{single}.

What if B is a further refinement of classification A (i.e., each group in A splits perfectly into several groups in B)? The Q_{single} parameter will be very small, since for each group in A only one group in B will be counted for. In order to counter this, we introduce the following modification. Now each group $a \in A$ is associated with all those groups b from B, for which $tp > fp$. Specifically, we say that a group $b \in B$ is a *relative* of a group $a \in A$ if more than 50% of b's members are also members of a (see Fig. 4). For each group $a \in A$ we identify all its relative groups b in B. The union of all the relatives of a is denoted by b_a. A protein is misclassified by classification B if it is a member of a missed by b_a (false-negative), or is a member of b_a, but not a member of a itself (false-positive). The intersection of b_a and a defines the group of correctly classified proteins.

We define the quality Q_{set} of the classification for the group a by the percentages of the true positives in the union $a \cup b_a$. Namely,

$$Q_{set}^{(a)} = 100 \cdot \frac{|a \cap b_a|}{|a \cup b_a|}$$

which accounts for both false-positives and false-negatives errors. This procedure is repeated for every group $a \in A$, and the total percentage of true-positives is given by the average over all groups $a \in A$.

As observed above, Q_{set} gives more favorable evaluations when B is a refinement of the partition A. This however, may lead to another problem, because dividing each group $a \in A$ into many small clusters (and in the extreme, to singletons clusters) is not desirable. Therefore, we also define another quality index, which accounts for the number of clusters which are relatives of a, and penalizes for excess in this number. This is done by first subtracting the number of relatives of a from the number of true-

¶The two classifications can be either "hard," i.e., each protein is classified to exactly one group, or "soft," in which each protein can be classified to more than one group. In our case, Pfam and PROSITE are soft, whereas (the current version of) ProtoMap is a hard classification.

TABLE II. Largest Clusters at the Lowest Confidence Level 10^{-0}†

Cluster number	Size	Order of transitivity	Family
1	718	2	Protein kinases
2	593	2	Globins
3	514	2	G-protein–coupled receptors
4	330	2	Immunoglobulin V region
5	326	2	Immunoglobulins and major histocompatibility complex
6	318	2	Homeobox
7	315	2	Ribulose bisphosphate carboxylase large chain
8	284	2	ABC transporters
9	260	1	Zinc-finger C2H2 type
10	256	2	Calcium-binding proteins
11	252	2	Serine proteases, trypsin family
12	229	2	GTP-binding proteins—ras/ras-like family
13	221	2	Myosin heavy chain, tropomyosin, kinesins
14	208	3	Collagens, structural proteins
15	206	2	Cytochrome P-450
16	198	2	GTP-binding elongation factors
17	196	2	Tubulins
18	190	1	Cytochrome b/b6
19	187	2	ATP synthases
20	172	2	Heat-shock proteins
21	171	2	Alcohol dehydrogenases (short-chain)
22	171	2	Snake toxins
23	152	2	NADH-ubiquinone oxidoreductase
24	142	2	Bacterial regulatory components of signal transduction
25	141	3	DNA-binding proteins of HMG
26	140	1	Nuclear hormones receptors
27	139	1	Actins
28	139	1	Intermediate filaments
29	138	2	GTP-binding, ADP-ribosylation factors family
30	136	1	Neurotransmitter-gated ion-channels
31	133	2	Zinc-containing alcohol dehydrogenases
32	133	2	Cellular receptors, EGF-family
33	130	3	Amylases
34	130	1	Hemagglutinin
35	129	2	RNA-directed DNA polymerase
36	125	1	Chaperones, chaperonins
37	122	2	Phospholipase A2
38	120	2	Insulins
39	115	1	Cytochrome c
40	115	3	Ketoacyl synthase
41	114	2	Growth hormones (somatotropin, prolactin, and related hormones)
42	113	1	Glyceraldehyde 3-phosphate dehydrogenase
43	113	3	Nuclear proteins, hn-RNP and sn-RNP, RNA-processing proteins
44	110	1	Viral nucleoprotein
45	109	1	Cytochrome c oxidase subunit II
46	108	3	Kazal serine protease inhibitors, secreted SPARC proteins
47	102	3	2Fe-2S ferredoxins, flavohemoproteins
48	102	2	Viral genome polyproteins
49	102	1	Developmental regulators—WNT family
50	101	1	Cation transport ATPases

†Clusters are ordered in decreasing order of size. The order of transitivity within each cluster is defined as follows: select the protein with the maximum number of neighbors and define it as the cluster's seed. The seed's order of transitivity is 0. Its neighbors are of order 1. Additional proteins that are neighbors of 1st order proteins, are of order 2, etc. The family description states the feature common to most of the member proteins.

positives, then calculating the percentages of the total number of proteins in the union $a \cup b_a$, i.e.,

$$Q^{(a)}_{set-relatives} = 100 \cdot \frac{|a \cap b_a - \text{number of relatives of } a + 1|}{|a \cup b_a|}$$

This way, a class with N elements which has a single identical relative cluster in classification B has quality of 1 (or 100%). On the other hand, if the relatives are N singletons in classification B, the corresponding quality is close to zero.

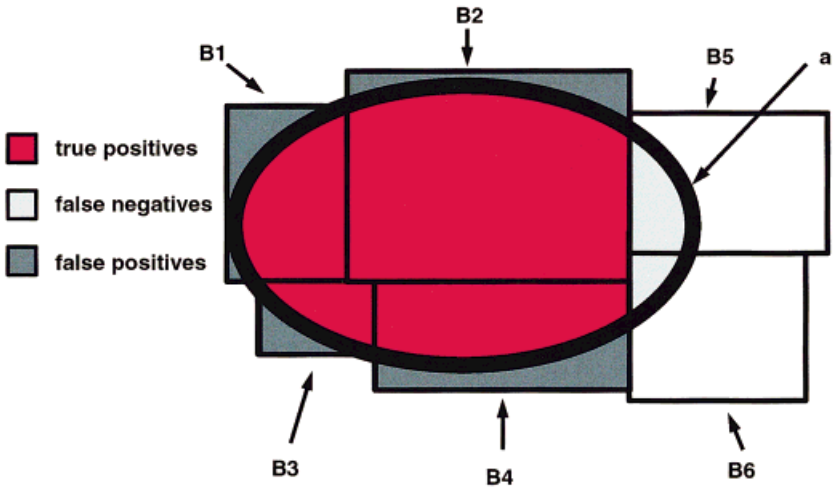

Fig. 4. Association of groups in classification B with groups in the reference classification A. Groups B1–B4 are relatives of the red ellipse (group of A), while groups B5 and B6 are not, as the overlap is too small. The set b_a is defined as b_a = B1 ∪ B2 ∪ B3 ∪ B4, and the quality Q_{set} is given by $Q^{(a)}_{set} = 100 \cdot |A \cap b_a| / |a \cup b_a|$. For comparison, Q_{single} is given by $Q^{(a)}_{single} = 100 \cdot |a \cap B2| / |a \cup B2|$.

The fourth index of quality is given by the quantity $tp/(tp + fn)$, which is the fraction of the reference family a which is in the set b_a. This measure does not take into account the false positives; the reference classifications may have not detected all homologies, and therefore not all seemingly false-positives are indeed such (see "Critique of the evaluation methodology" below). This index is based on the (overly generous) assumption that all false-positives are in fact potential related sequences. Obviously, this is not true, and we should be extra cautious in using it in automatic assignments of protein sequences. This estimate still provides a useful upper bound on the quality of the classification, because typically some of the false-positives are indeed related sequences. We denote this quantity by $Q_{upper\text{-}bound}$.

The consistency test proposed by Krause and Vingron[31] can be used for rough self-validation, but it is less useful for assessing the quality of a new classification with respect to a reference classification.

The reference databases

We compared our classification with two domain-based classifications: PROSITE and Pfam. Our classification contains 52,205 proteins of the SWISSPROT database, release 33, classified to 10,602 clusters, of which 1,006 have size 10 and above (see Table I). The PROSITE database, release 13 (released with the SWISSPROT 33 database), contains 24,156 proteins, characterized by 1,151 different signature patterns. PROSITE often associates the same family with two or more signature patterns. Therefore, sequences with different PROSITE patterns documented as the same PROSITE family are considered to belong to the same family. For example, all proteins that have either the ACTINS_1, the ACTINS_2, or the ACTINS_ACT_LIKE signature belong to the actins family. The exceptions are those patterns that never appear together in the same protein although they are documented the same (e.g., ANTENNA_COMP_ALPHA and ANTENNA_COMP_BETA). Patterns that are always associated with other patterns (e.g., INTEGRIN_BETA with EGF, POU_1 and POU_2 with HOMEOBOX) are ignored. Overall, within these terms, the 1,151 signatures characterize 874 protein families and domains, of which 600 are of size 10 and above. The Pfam database (release 1.0, associated with SWISSPROT 33) contains 15,604 proteins, classified to 175 families, of which 172 are of size 10 and above.

Recently, our clustering procedure was applied to a newer release of SWISSPROT (release 35 with updates up to May 6th 1998, with a total of 72,623 proteins). Because this was not one of the main releases, it complicated the assessment procedure, especially in comparisons with the PROSITE and Pfam classifications. Therefore, we applied the assessment procedure to ProtoMap, PROSITE, and Pfam releases associated with the SWISSPROT 33 database. The latest releases of PROSITE and Pfam are associated with SWISSPROT 35 database. This database includes 69,113 sequences, 43,053 (62%) of which are included in the 1,390 families of the Pfam database release 3.3, and 35,340 (51%) are included in the PROSITE 14 database. The results of the analysis of both SWISSPROT releases are available at the ProtoMap web site. In Appendix A we give the mutual correspondence/correlation of both releases.

Evaluation results

The results of the evaluation procedure for these reference databases are given in Table III. The evaluation is based on all families with at least ten members (the same analysis with all families with at least five members gave the same results up to within 1.5%).

Gracy and Argos[27] used a similar procedure (using the Q_{single} parameter) with respect to a reference classification that was a combination of PROSITE and PIR.[50] The assessment resulted in 96.6% true-positives (1.8% false-positives) for PROSITE, 93.2% true-positives for DOMO (0.3% false-positives), and 65.1% true-positives for ProDom (0.9% false-positives). All three are domain-based classifications. The high percentage of true-positives in PROSITE with respect to this reference classification indicates that the combined database is not much different

TABLE III. Performance Evaluation at Selected Confidence Levels†

ProtoMap confidence level	Reference database	Q_{single} (%)			Q_{set} (%)			$Q_{set\text{-}relatives}$ true-positives (%)	$Q_{upper\text{-}bound}$ (%)
		True-positives	False-positives	False-negatives	True-positives	False-positives	False-negatives		
10^{-100}	PROSITE	13.3	0.9	85.8	95.2	3.7	1.1	43.4	98.9
	Pfam	14.8	1.7	83.5	96.4	2.2	1.4	53	98.5
10^{-50}	PROSITE	34.3	2.5	63.2	92.1	5.3	2.6	61.8	97.3
	Pfam	35.0	2.5	62.5	92.0	3.0	5.0	68.9	94.7
10^{-10}	PROSITE	65.9	6.4	27.7	85.7	8.3	6.0	78.5	93.8
	Pfam	60.5	5.4	34.1	84.6	4.9	10.5	79.6	89.1
10^{-0}	PROSITE	68.4	9.5	22.1	77.8	11.1	11.1	75	88.5
	Pfam	64.8	7.3	27.9	76.7	6.9	16.4	75	83.1

†At the level of 10^{-100} most ProtoMap clusters are small (what explains the low Q_{single} value), but highly specific (what explains the high Q_{set} value). As the e-value threshold is increased, and more permissive similarities are taken into account, ProtoMap clusters merge to form bigger clusters (Q_{single} increases), which are more diverged (Q_{set} decreases). $Q_{set\text{-}relatives}$ is a compromise between these two measures. At the level of 10^{-100} its value is low, in spite of the high value of Q_{set}, because each PROSITE and Pfam family is matched with several small clusters of ProtoMap. At more permissive levels the value of $Q_{set\text{-}relatives}$ increases, because most families are matched with a single cluster (see Table IV). It reaches its maximum at 10^{-10}.

TABLE IV. Distribution of the Number of Relative Clusters†

	Number of relative clusters in ProtoMap								
	0	1	2	3	4	5	6–10	>10	
Number of PROSITE families	73	525	131	62	24	18	30	11	874 (total)
Number of Pfam families	15	76	33	18	10	1	15	7	175 (total)

†The majority of PROSITE and Pfam families are associated with a single cluster in ProtoMap. Out of the 73 PROSITE families which have no relative cluster in ProtoMap (see "The evaluation methodology" for a definition of "relative cluster") 62 are mapped exhaustively to a single cluster in ProtoMap. Most of these families are subfamilies of larger families, each of which is mapped to a single cluster, e.g., RAN (see "Critique of the evaluation methodology" for details). Some are classified to clusters that still need to be refined (see Discussion).

from the PROSITE database. The first three columns in Table III assess the performance of ProtoMap in the same way as Gracy and Argos[27] with respect to PROSITE. With 68.4% true-positives, our work compares favorably with ProDom, although we find many more false-positives. However, as we note in the next section, not all these false-positives should be counted as such.

Because ours is a hard classification, we find the second measure Q_{set} more appropriate than Q_{single}. This means that no single cluster is associated with one domain family. Consider, for example, the PROSITE's AA_TRNA_LIGASE _I family. Seven different clusters form the cover of this family corresponding to the subfamilies: leucyl/isoleucyl/valyl/methionyl-trna synthetase (cluster 152), glutamyl/prolyl-trna synthetase (cluster 429), tryptophanyl-trna synthetase (cluster 758), arginyl-trna synthetase (cluster 988), tyrosyl-trna synthetase (cluster 904), cysteinyl-trna synthetase (cluster 1216), and a singleton (cluster 6647) arginyl-trna synthetase (a very short fragment).

Although most of PROSITE families have only one relative in our classification (see Table IV), many families are associated with more than one cluster, where different clusters may correspond to different subfamilies (belonging to the same family is still detectable through connections between clusters, as discussed in section "possibly related clusters"). Consequently, with the Q_{set} index, the quality of performance reaches 77.8%.

Critique of the evaluation methodology

The above procedure may result in an over-strict measure. False-positives may be overcounted, because often supposedly false-positives with respect to the reference database are actually true-positives. For example, short fragments which surely belong to a specific family may be considered false-positive simply because they are too short to completely include the domain which is common to all other members in the family. Similarly, hypothetical proteins with a significant sequence similarity with a family are not necessarily false-positives. Even proteins which are documented as members of a family may be counted as false-positives simply because they do not have the exact family signature pattern, but rather a slightly modified one. For example, cluster 27 has 139 proteins of which 132 have the actin and actin-like signature. Five of the other seven proteins are documented as actins (and indeed show a remarkable similarity with other actins), and two are hypothetical proteins (again, with a strong similarity to actins). However, these seven proteins do not have the actin and actin-like PROSITE signature and therefore are counted, unjustifiedly perhaps, as false-positives. Similarly, cluster 7 has 46 proteins which do not have the rubisco_large PROSITE signature, and are thus counted as false-positives. They are all, however, annotated in SWISSPROT as ribulose bisphosphate carboxylase large

chain (some are fragments). Similarly, cluster 15 has 13 proteins which do not have the cytochrome P-450 signature; ten of these are variants of cytochrome P-450, two proteins are thromboxane-a synthase, and one is trans-cinnamate 4-monooxygenase. All of these are documented to be part of the cytochrome P-450 family, and indeed, show a strong similarity with cytochromes P-450.

Such "false-positives" are very common in our clusters, but should they count as false-positives? Obviously, sometimes they are, but we have no automatic way to discern those that are indeed biologically meaningful. At present we can only say with confidence that some false-positives should not count as such. Consequently, the true value of performance quality lies somewhere between Q_{set} = 77.8% and $Q_{upper\text{-}bound}$ = 88.5% of true-positives.

Here is another factor which limits the agreement: Some families have subfamilies, the members of which share a well-defined domain which other members in the family do not have. In the evaluation procedure these will be considered as false-positives, and if the latter are a majority, then no cluster will be matched with the subfamily. A case in point is the ran family, a subfamily of the small G-proteins. The ran proteins are classified to the same cluster as the ras/ras-like/rab proteins (cluster 12). However, because the ras/ras-like/rab proteins are not characterized by the same signature pattern (nor by any other signature pattern in release 13.0 of PROSITE), they count as false-positives. The functional relationship of the proteins in this cluster points to the problem of assessing a new classification by means of another, human-made classification.

Families for which our analysis performed worst, with less than 50% true-positives, are dominated by short/local domains (e.g., PH domain, EGF, ER_TARGET, C1Q, KRINGLE, C2 domain, SH2, SH3) or domains that are paired with other, more abundant domains (e.g., opsin paired with G-protein receptor). This is to be expected, because our analysis is not a domain-based.

The results of the evaluation procedure with Pfam as the reference database lead to similar conclusions: good performance for protein families, but short motifs are not detected well. The quality of the classification increased as the Pfam coverage (the total portion of the sequences which was included in the multiple alignment used to define the domain or the family in the Pfam database) increased: for coverage >0.3 (134 families) the quality raised to Q_{single} = 76.9% (6.1% false-positives), Q_{set} = 84.9% (6.2% false-positives), and $Q_{upper\text{-}bound}$ = 90.9%; for coverage >0.5 (109 families) the quality increased to Q_{single} = 80.6% (5% false-positives), Q_{set} = 88.3% (5% false-positives) and $Q_{upper\text{-}bound}$ = 93.3%.

New clusters

The above evaluation procedure is oblivious to the many new clusters in our classification that have no counterpart cluster in PROSITE, nor in Pfam. Our definition of a counterpart is very strict. A cluster has no counterpart family in the reference database if NONE of its members are associated with ANY reference family. Of the 1,006 clusters with over 10 members (total of 33,682 proteins), 308 clusters (6,989 proteins, which comprise 20.8%) have no counterpart family in PROSITE, 734 clusters (15,586 proteins 46.3%) do not have a counterpart in Pfam A, and 281 clusters are missing from both.

TABLE V. Largest Clusters With No Corresponding Family in PROSITE 13 or Pfam 1.0

Cluster number	Size	Family
23	152	NADH-Ubiquinone reductase (chains 2, 4, 5)
34	130	Hemagglutinin (virus)
44	110	Nucleoprotein (virus)
60	92	Phycocyanin (algae)
67	86	Histone H1
79	77	Envelope protein (virus)
85	74	Chlorophyll A-B binding protein (plants)
102	69	NADH-Ubiquinone oxidoreductase (chain 6)
110	64	60S ribosomal protein (P0, P1, P2)
114	61	Envelope protein (virus)
120	58	E6 (viral protein)
129	56	Neurotoxins (insect)
132	55	NADH-Ubiquinone oxidoreductase (chain 3)
133	55	RNA polymerase B
135	54	Probable L1 protein (virus)
137	53	E1, helicase (virus)
138	53	E2, transactivator (virus)
139	53	Probable L2 protein (virus)
142	52	TAT protein, transactivator (virus)
145	51	E7, transforming (virus)

The largest 20 unannotated clusters (both by PROSITE and Pfam) are listed in Table V. They are documented based on their SWISSPROT definition. The purity of these clusters (in terms of definition consensus) is very high. Still proteins in these clusters are not characterized in PROSITE 13 or in Pfam 1.0. It should be noted that both databases have been extended since then (as did ProtoMap). Yet, many proteins in the updated SWISSPROT database are not classified by the latest releases of these databases (see "The reference databases" above).

ProtoMap as a Tool for Analysis

Aside of the direct use of ProtoMap as an automatic classification of protein sequences in the SWISSPROT database, ProtoMap offers additional information, which is available interactively in the web site.

Tracing the formation of clusters

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

By moving from one level to a more restrictive one, we obtain a subdivision of clusters into smaller subsets. These subsets suggest a natural division of the corresponding family, as illustrated in the following example for the transport system permease proteins.

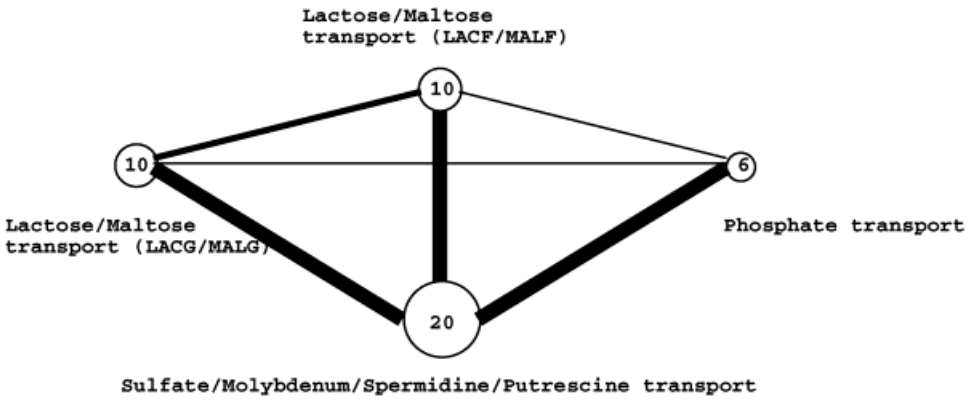

Fig. 5. Tracing the formation of cluster 170 (at level 10^{-0})—the transport system permease proteins. As we move on to level 10^{-5} and further to level 10^{-10} the cluster splits into several subclusters. Each circle stands for a cluster at threshold = 10^{-10}. Radii of the circles are proportionate to the sizes of the clusters (numbers indicate sizes of clusters). The drawn edges appear upon changing the threshold from 10^{-10} to the more permissive 10^{-5}. Edge widths are proportionate to the number of connections between the corresponding clusters.

Cluster 170 at level 10^{-0} with 46 proteins consists of transport system permease proteins. These proteins participate in multicomponent transport systems in bacteria. Specifically, they are the integral inner-membrane proteins which translocate the substrate across the membrane.

The cluster decomposes into four subclusters at level 10^{-10}, which form a clique (Fig. 5). These smaller subclusters correspond to the lactose/maltose transport system lacG/malG, the lactose/maltose transport system lacF/malF, the phosphate transport system, and other transport systems (of sulfate, molybdenum, spermidine, and putrescine). The subgroups of lacG/malG and lacF/malF form already at level 10^{-25}. Some proteins that combine features from F and G subtypes are denoted in SWISS-PROT as malGF proteins. However, based on this subclassification and the fact that the malG group and the malF group form at such high levels of significance, these proteins may be classified either to malG or to malF.

Hierarchical organization within protein families and superfamilies

The hierarchical organization also suggests classification within known families. This classification is suggested by scanning the hierarchy over all levels, as illustrated for the small G-protein/Ras superfamily (Fig. 6).

The ras gene is a member of a family that has been found in tumor virus genomes and that is responsible for the viruses' carcinogenic effect. In most cases this viral oncogene is closely related to a cellular counterpart, called a 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, all of which share the guanine nucleotide-binding site. They are referred to as the small G-protein superfamily.[51] This family of proteins has several subfamilies: ras, rab, ran, rho, ral, and smaller subfamilies. Like ras, these proteins participate in regulatory processes, such as vesicle trafficking (rab) and cytoskeleton organization (rho).

In Figure 6 we depict the relations within this family, based on our hierarchical organization. A total of 229 proteins, all from the small G-protein superfamily, are presented. All were clustered into cluster 12 at the lowest level of significance 10^{-0}. Small clusters, which correspond to subfamilies, are formed at higher confidence levels, and fuse to larger clusters when the threshold is lowered. The four main branches coincide with (I) rab subtypes; (II) ras, ral, and rap; (III) rac, rho, and cdc (cell division control proteins); (IV) ran. Interestingly, the linkage of rac to cdc and rho seems stronger than that between ran and rho or rab and rho. This proposed subdivision suggests a common root for all the subtypes, but splits them in a way that resembles the evolutionary tree of the small G-protein superfamily.[52]

As we proceed to include weaker similarities, we identify other families which are related to the small G-protein family. According to our map, the clusters which are detected as related clusters include cluster 29 (138 proteins), cluster 646 (14 proteins) cluster 461 (20 proteins), and cluster 1400 (7 proteins). All these clusters are *possibly related clusters* (rejected merges) of cluster 12 (see next section for a discussion of possibly related clusters). Cluster 29 consists of ADP-ribosylation factors family (ARF) that are involved in vesicle budding and of guanine nucleotide-binding proteins from the sar subfamily, whose members participate in a different type of vesicle budding. Another family which is classified to this cluster is G_α proteins of heterotrimeric G proteins. The connection between this cluster and cluster 12 is based on similarity of the ARFs and the rab subfamily, as shown in Figure 6. Cluster 646 consists of the GTP-binding protein ERA and of thiophene/furan oxidation proteins (both being groups of GTP-binding proteins). This cluster and cluster 12 are related through the similarity of the thiophene/furan oxidation proteins and the ras subfamily. Cluster 461 (GTP-binding proteins of the OBG family) and cluster 1400 (hypothetical small G-proteins) are not directly related to cluster 12. However, these clusters are related to clusters 29 and 646, as well as to each other (see Fig. 6).

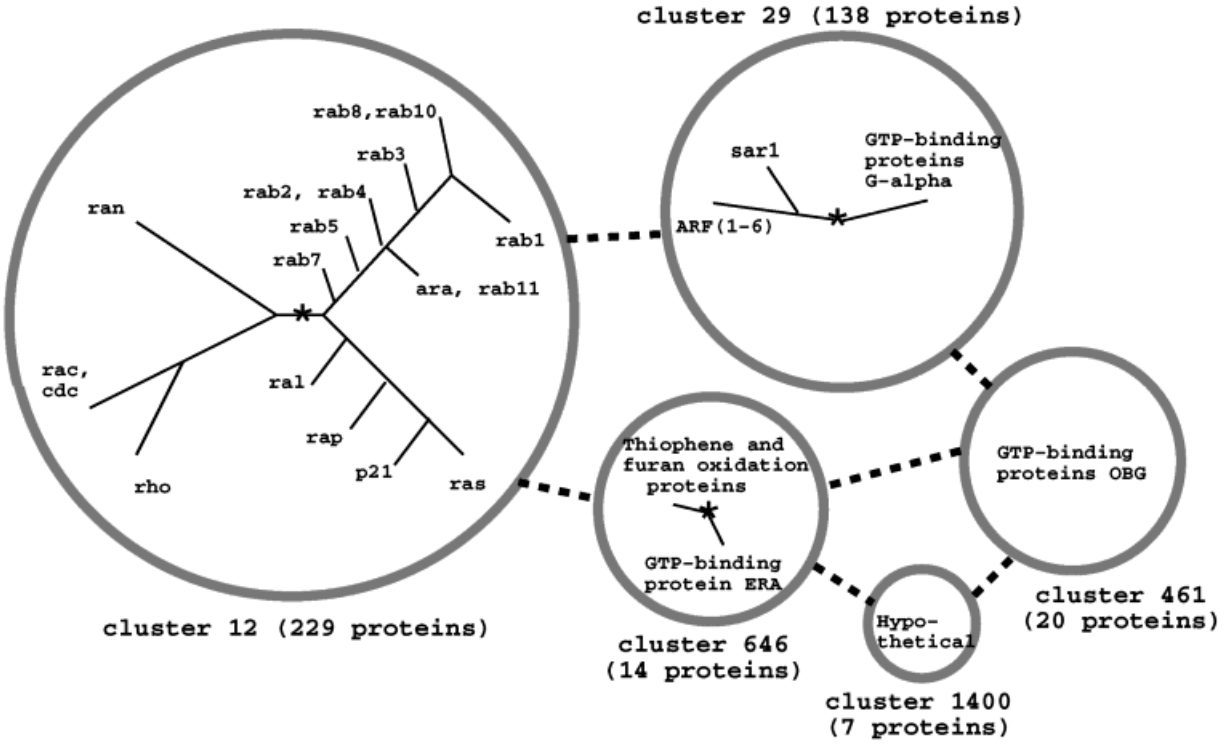

Fig. 6. The small G-protein family. This family is composed of several subfamilies. A total of 229 proteins, combined in cluster 12 (level 10^{-0}), were grouped together into isolated sets at different levels of confidence, to form a natural subclassification of the family. This hierarchical organization is much enriched by combining possibly related clusters (see text). The related clusters are connected by dashed lines.

Possibly related clusters and local maps

Recall that our clustering algorithm differs from a single linkage algorithm. First, the algorithm identifies groups of possibly related clusters using "local considerations" (strong connections between *pairs* of clusters). Then, a "global test" is applied to identify nuclei of strong relationships within these groups of clusters, and clusters are merged accordingly. During this process, the algorithm automatically rejects many possible connections among clusters. This happens whenever the quality associated with a connection falls below a certain threshold (see Fig. 3). Many of these rejected connections are nevertheless meaningful and reflect genuine though distant homologies. We refer to the rejected mergers as *possibly related clusters*.

For almost all clusters, much insight can be gained by observing their possibly related clusters. On average, there are three possibly related clusters per cluster (at the level of 10^{-0}). Even though some of these connections are justifiably rejected, in particular at the lowest level of confidence we consider (10^{-0}), many others do reflect structural and functional similarities, 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 pairwise alignments of protein pairs, one from the cluster under study and one from a possibly related cluster (the alignments are shown in the web site).

Based on the connections with possibly related clusters we can plot local maps (at this stage, mostly schematic) for the neighborhoods of protein families. These schematic maps can expose interesting relationships between protein families. Here we present a map for the immunoglobulins superfamily. Two of the big clusters in Table II belong to this superfamily. These are cluster 4 (immunoglobulin V region) and cluster 5 (immunoglobulins C region).

Table VI shows those clusters which are possibly related to cluster 4, ordered by their quality value. These clusters include proteins which are involved in aspects of recognition at the immune system via the variable regions.

Likewise, Table VII shows the clusters which are possibly related to cluster 5. These clusters, unlike the clusters related to cluster 4, consist mostly of proteins that adopted the immunoglobulin fold of the Ig constant region. Clusters which we suspect to be unrelated appear in italics (one can validate the significance of possibly related clusters using the quality of their relation and the alignments, and insignificant connections can be easily traced and ignored by a manual examination of the alignments in the web site).

The two sets of clusters are mostly disjoint, with the exception of cluster 1796. Members of this cluster contain both regions, whence the cluster is related to both clusters 4 and 5. The different parts of the proteins account for the appropriate relationships.

There is also a direct connection between cluster 4 and cluster 5. The connection is based on 226 pairwise similarities between cluster 4 and 5. However, all the similarities are due to a single protein in cluster 5, a T-cell receptor

TABLE VI. Clusters Possibly Related to Cluster 4 (Level: 1e-0)†

Cluster number	Size	Connection quality	Number of edges	Family
1,643	5	0.29	219	B-cell antigen receptor complex–associated protein
927	10	0.11	193	T-cell surface glycoprotein CD4
2,613	3	0.03	20	Polymeric-immunoglobulin receptor
5	326	0.01	226	Immunoglobulins and major histocompatibility complex
1,137	8	0.01	18	T-cell–specific surface glycoprotein CD28, cytotoxic T-lymphocyte protein
1,189	8	0.01	9	Myelin P0 protein precursor
1,796	5	0.01	9	Poliovirus receptor precursor

†Clusters are sorted by quality (i.e., minus the log of the geometric mean of similarity scores). Note that all clusters belong to the superfamily of immunoglobulins.

TABLE VII. Clusters Possibly Related to Cluster 5 (Level: 1e-0)†

Cluster number	Size	Connection quality	Number of edges	Family
1,831	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-β precursor, and more
578	16	0.01	28	High/low-affinity immunoglobulin ϵ/γ FC receptor
596	*16*	*0.01*	*33*	*Recombination-activating proteins, zinc finger, c3hc4 type*
856	*11*	*0.01*	*11*	*Cornifin (small proline-rich protein)*
1,262	7	0.01	21	T-lymphocyte activation antigen CD80/CD86 precursor
1,636	5	0.01	8	Basigin precursor
1,796	5	0.01	7	Poliovirus receptor precursor

†Only clusters with two members or more are shown, and are sorted by quality. Almost all clusters belong to the superfamily of immunoglobulins. Clusters 596 and 856 are probably unrelated (in italic).

beta chain (P11364). This protein has one V region aside from a C region, and the similarity with the 226 proteins in cluster 4 is limited to the V region. No other protein in cluster 5 has a V region. Note that despite these similarities, clusters 4 and 5 did not merge, and the connection was automatically rejected by our clustering algorithm.

This information depicts the "geometry" of the protein space in the vicinity of the immunoglobulin superfamily as in Figure 7. In this map we include also indirectly related clusters, i.e., possibly related clusters of order 2 and above (related clusters of related clusters, etc). The map includes almost all protein families which belong or are related to the immunoglobulin superfamily defined by the SWISS-PROT database (see Table VIII for details), except for three clusters: cluster 373 (isolated cluster of periplasmic pilus chaperones), cluster 2172 (isolated cluster of THY-1 membrane glycoproteins), and cluster 2363 (B-lymphocyte antigen cd19).

DISCUSSION

This study addresses the problem of identifying high-order features and organization within the space of all protein sequences. Our aim is to exhaustively "chart" all proteins and to automatically classify them into families, based on pairwise similarities.

A complete charting of the protein space 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 exclude connections which are unacceptable or misleading. The main difficulties stem from chance similarities among sequences and multi-domain–based connections. Semi-automatic procedures were developed mainly for domain-based family identification (see Introduction). However, the sheer volume of data makes it necessary to develop automatic methods to complement such attempts.

The work we present here addresses these major problems. We start by creating, for each protein sequence, an exhaustive list of neighboring sequences. These lists take into account the scores of the three major methods for pairwise sequence comparisons. These three types of scores are jointly normalized and the lists are filtered. The link is maintained in the lists only when a significant relationship seems to exist. A two-phase clustering algorithm is then applied to identify groups of related sequences. There are two pitfalls to avoid here: 1) It is very easy to follow a very strict rule and generate many small clusters, within each of which the proteins are very closely related. This approach safely avoids the creation of false connections, but adds little to our understanding of the protein space, because it includes only fairly obvious and well-known connections. 2) A procedure that declares a connection without sufficient scrutiny does not miss interesting connections. Instead it generates so many false connections that it becomes impossible to recognize significant relations. In other words, such a permissive method quickly collapses the whole space into a small number of gigantic but biologically meaningless clusters. The problem is to find a golden path where nonobvious relationships are discovered without "littering" the classification with too many false connections.

TABLE VIII. Clusters Belonging to the Immunoglobulin Superfamily†

Cluster number	Size	Family
4	330	Immunoglobulin V region
5	326	Immunoglobulins and major histocompatibility complex (MHC)
104	66	Cell adhesion molecules (CAM, n-CAM, ng-CAM, v-CAM, contactin, fascilin II), myelin-associated glycoprotein, axonin-1 precursor, B-cell receptor CD22-β precursor
373*	25	Periplasmic pilus chaperones
578	16	High/low-affinity immunoglobulin ϵ/γ FC receptor
621	15	Interleukin-1 receptor, interleukin-1 binding protein, surface antigen
854	11	T-cell surface glycoprotein CD3 δ/ϵ/γ chain precursor
927	10	T-cell surface glycoprotein CD4
1,075	9	MHC class I NK cell receptor precursor
1,137	8	T-cell–specific surface glycoprotein CD28, cytotoxic T-lymphocyte protein 4
1,189	8	Myelin P0 protein
1,262	7	T-lymphocyte activation antigen CD80/CD86 precursor
1,301	7	Intercellular adhesion molecule precursor (ICAM)
1,468	6	Hemagglutinin precursor
1,636	5	Basigin precursor
1,637	5	Proable cell adhesion molecule involved in regulating T-cell activation
1,643	5	B-cell antigen receptor complex–associated protein
1,727	5	Interleukin-12 beta chain precursor
1,796	5	Poliovirus receptor, ox-2 membrane glycoprotein
1,831	5	T-cell receptor γ chain C region
1,938	4	T-cell surface antigen CD2 precursor
2,172*	4	THY-1 membrane glycoproteins
2,363*	3	B-lymphocyte antigen cd19
2,294	3	α-1b-glycoprotein
2,613	3	Olymeric-immunoglobulin receptor
4,622	1	β-2-microglobulin
4,763	1	T-cell surface glycoprotein CD4
5,186	1	Fasciclin III precursor
5,583	1	Lymphocyte activation gene
5,847	1	Immunoglobulin μ chain C region membrane-bound

†All clusters (other than those marked with an asterisk) appear in the local map of the immunoglobulins (Fig. 7).

Our algorithm can be described as a moderate version of the transitive closure algorithm. At each round of this process we gain statistical information on the relationships among current clusters. This information is then used to merge certain clusters, thus forming the next round of larger, coarser clusters. The algorithm starts from a very conservative classification, and is repeatedly applied, at varying levels of confidence, the input for each stage being the classification output at the previous stage. Finally, a hierarchical organization of all protein sequences is obtained, strongly correlated with a functional partitioning of all proteins. This data structure reveals interesting relations between and within protein families, and provides a global view ("map") of the space of all proteins.

The classification consists of several thousand clusters, the largest of which contain several hundred members each. It is interesting to assess the effect of slowing down the transitive closure algorithm. Indeed, a straightforward application of the transitive closure algorithm (a.k.a. single-linkage clustering) leads to an avalanche as discussed above (details not shown). Already at a confidence level 10^{-20}, most of the space is made up of a small number of very large clusters. This avalanche is caused by chance similarities and chains of domain-based connections that cause 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.

We should note that some connections found in our analysis are still questionable. Some domain-based connections did escape our filters and caused unrelated clusters to merge (e.g., cluster 122 which contains the pancreatic trypsin inhibitor [Kunitz] family as well as the amyloidogenic glycoprotein intracellular/extracellular domains). Also, high-scoring low-complexity segments that may be biologically meaningless can lead to false connections and to the formation of nonhomogeneous clusters (e.g., cluster 14 of collagens and other structural proteins). Although we did take into account the effect of these segments, not all of them were filtered out. Because our goal was to detect many remote homologies we used weak filters in this case and considered many similarities at very low levels of confidence. We are currently testing more stringent filtration criteria and improving our algorithm to handle domain-based connections better. At the moment, such connections can be easily traced manually, by observing the alignments (at the web site). Multiple alignments will be available as well in the next release.

It is very difficult or even impossible to properly classify all proteins. The space of proteins has many different facets, all of which should be considered in future, more thorough classification: 3D structure/fold, biological function, domain content, cellular location, tissue specificity, organism (source), metabolic pathways, etc. This work differs from previous large-scale analyses in several ways: 1) We do not attempt to identify protein domains or motifs. 2) No predefined groups or other classification are being employed in our analysis. Moreover, no multiple alignments of the proteins are needed. 3) We chart the space of *all* protein sequences in SWISSPROT, not just particular families. 4) We offer a global organization of all protein sequences. In the ideal scheme, a combined strategy should be developed which includes protein-based considerations as well as domain-based considerations and structural information in a more rigorous way. Such a scheme is currently being developed (Yona and Levitt, unpublished observations).

Our algorithm has turned out well-defined groups which are strongly correlated with protein families and subfami-

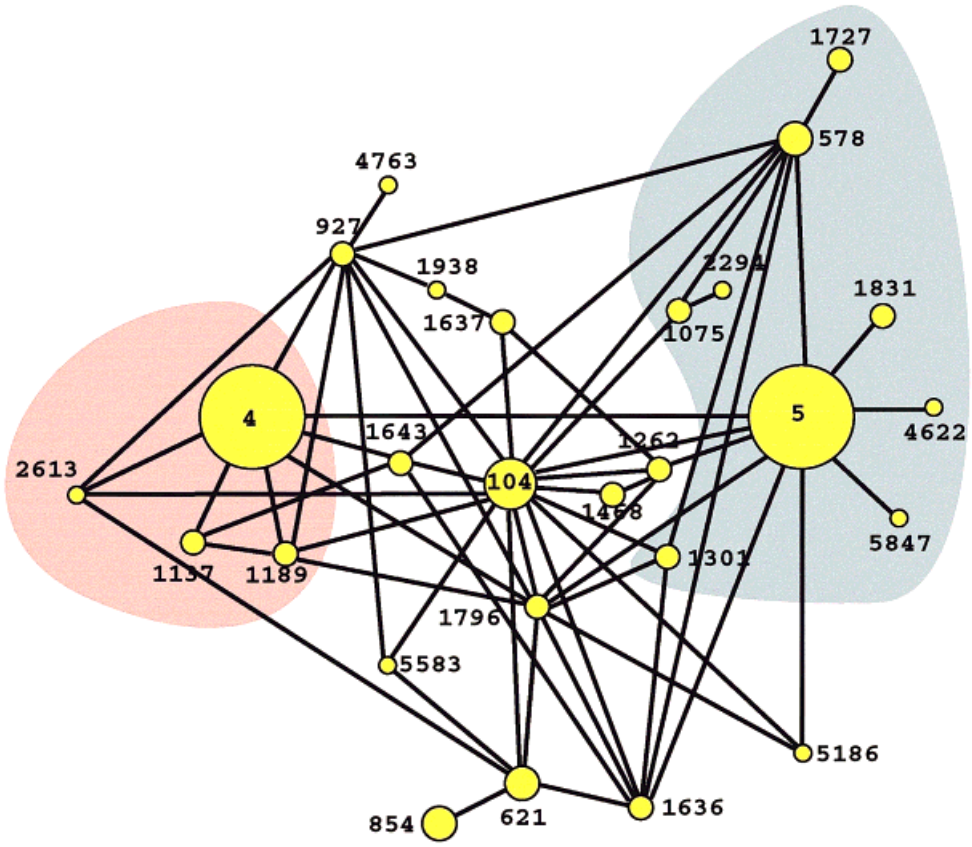

Fig. 7. The immunoglobulin superfamily. All clusters related to clusters 4 and 5 are shown (these are referred to as *directly related clusters*), as well as other clusters which are indirectly related clusters (clusters that are possibly related to directly related clusters). In this schematic map, each cluster is represented as a circle, whose radius is proportional to the cluster's size. Only clusters' numbers are given. See Table VIII for more information on these clusters. Clusters on the left (colored group) contain an Ig V-region. The group on the right has an Ig C-region. The clusters in between share both regions (except for cluster 1468, which consists exclusively of V-region). This ''local map'' of relationships is plotted to distinguish between three main groups (this map may change upon the accumulation of new data). The left one corresponds to the variable regions (V domains) of the immunoglobulins, the right one to the constant regions (C domains). In the middle appear many clusters that are a mixture of the two types. The many alternative connections between clusters whose proteins resemble V-domain to these of the C-domains indicate that adhesion molecules, fasciclin II and vascular CAM, are positioned between the classical V and C-regions. Indeed, studies of the evolutionary pathway between the structural classes of Ig and Ig-like domains have pointed to the important role of these non-Ig molecules, and they are considered as the I-set according to their intermediate nature.[53,54]

lies. When compared with the well-accepted databases PROSITE and Pfam, our classification performed well for most of the families, although many of them were domain based rather than protein families. Many new clusters in our classification did not have a match from either PROSITE nor Pfam. The hierarchical organization can indicate the existence of subfamilies within families, and the concept of possibly related clusters exposes distant relationships which reflect functional similarities. These relations offer a basis for sketching local maps near protein families. These connections can be of biological interest and should be taken into account in the study of protein families. For example, an overall view of Figure 6 or Figure 7 implies that tracing the hierarchical organization of a cluster and/or its rejected merges can provide new information about the corresponding family and reveal relationships among protein families.

Another aspect of the possibly related clusters is that this list can be viewed as a "soft clustering," where the same protein can participate in several different clusters. Proteins that are composed of several domains, some of which are shared by proteins from different families, are naturally associated with more than one group. Therefore, although in this version of ProtoMap a multidomain protein is classified to a single cluster, its multitrait nature is revealed when we examine its relations with the other clusters as well.

The concept of soft clustering will be integrated in the next releases of ProtoMap and is already applied for the online classification of new protein sequences (submitted by the user), in the newest version of ProtoMap. Each new sequence is classified to the existing clusters based on its distribution of connections with the existing clusters. The sequence can be classified to more than one cluster, with different qualities, to help in predicting its nature.

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

TABLE IX. Correlation of ProtoMap Releases

A. 1,514 clusters mapped to a single new cluster				B. 378 clusters mapped to several new clusters				
"Stable"				"Stable"				
The same	Only new sequences added	Merged with singletons	Others	Perfect split	Split; only new sequences added	Split and merged with singletons	Others	Total stable
399	947	102	66	29	162	39	148	1,678 (89%)

ACKNOWLEDGMENTS

We thank Michael Levitt and Steven Brenner for critically reading this manuscript and for many helpful comments and Hanah Margalit for many valuable discussions. The results shown here are based on very extensive computations. We were generously helped by access to Amnon Barak's MOSIX parallel system. Compugen Ltd. donated a Bioccelerator which was crucial to the completion of this computational task. We also thank Alex Kremer, Avi Kavas, Yoav Etsion, and Daniel Avrahami for creating the ProtoMap web site. Golan Yona is supported by the Program in Mathematics and Molecular Biology (PMMB).

APPENDIX A

Since the first release of ProtoMap (releases 1.0–1.2), we have run our algorithms on a newer version of SWISSPROT. That is, SWISSPROT 35 with updates (dated May 6th 1998) with total of 72623 proteins (39% increase), the corresponding release numbered 2.0. As was mentioned in section 2, we preferred to run the evaluation procedure on the SWISSPROT 33, since the updated version is not one of the major releases, and therefore no corresponding PROSITE or Pfam databases were available. However, for reference, we checked the correlation of the old release and the new release. Specifically, we checked: How many clusters remained unchanged, how many clusters grew larger due to the addition of new protein sequences, how many clusters split, and how many clusters merged. This procedure is used to identify those clusters which seem "stable" vs. "unstable" ones. A cluster in the first release of ProtoMap is considered stable, if one of the following conditions hold:

- The cluster remains the same in the new release.
- Only new protein sequences are added to the cluster in the new release.
- Only new protein sequences and old singletons join the cluster.
- The cluster perfectly splits into several smaller subclusters.
- The cluster splits into several clusters, to which only new sequences are added.
- The cluster splits into several clusters, each of which augmented only by new sequences and old singletons.

Such clusters are well correlated with protein families. All other clusters are considered unstable. Unstable clusters are not necessarily "false" clusters. Some merges and splits of clusters are proper responses to the new information carried by the addition of new protein sequences. The new clusters may in fact be better correlated with protein families or subfamilies. Some are unfortunately improper.

The ratio between the two types of clusters can help in assessing the stability of the ProtoMap system. The results are summarized in Table IX, based on the statistics of 1,892 clusters with more than five members in ProtoMap 1.2. This analysis shows that 89% of the clusters are stable.

REFERENCES

1. Pennisi E. Microbial genomes come tumbling in. Science 1997;277:1433.
2. Doolittle RF. Microbial genomes opened up. Nature 1998;392:339–342.
3. Needleman SB, Wunsch CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 1970;48:443–453.
4. Smith TF, Waterman MS. Comparison of Biosequences. Adv Appl Math 1981;2:482–489.
5. Lipman DJ, Pearson WR. Rapid and sensitive protein similarity. Science 1985;227:1435–1441.
6. Altschul SF, Carrol RJ, Lipman DJ. Basic local alignment search tool. J Mol Biol 1990;215:403–410.
7. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
8. Sander C, Schneider R. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 1991;9:56–68.
9. Flores TP, Orengo CA, Moss D, Thoronton JM. Comparison of conformational characteristics in structurally similar protein pairs. Protein Sci 1993;2:1811–1826.
10. Hilbert M, Bohm G, Jaenicke R. Structural relationships of homologous proteins as a fundamental principle in homology modeling. Proteins 1993;17:138–151.
11. Brenner SE, Chothia C, Hubbard TJP. Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. Proc Natl Acad Sci USA 1998;95:6073–6078.
12. Murzin AG. OB(oligonucleotide/oligosaccharide binding)-fold: common structural and functional solution for non-homologous sequences. EMBO J 1993;12(3):861–867.
13. Pearson WR. Identifying distantly related protein sequences. Comp Appl Biosci 1997;13(4):325–332.
14. Pearson WR. Effective protein sequence comparison. Methods Enzymol 1996;266:227–258.
15. Doolittle RF. Reconstructing history with amino acid sequences. Protein Sci 1992;1:191–200.
16. Gonnet GH, Cohen MA, Benner SA. Exhaustive matching of the entire protein sequence database. Science 1992;256:1443–1445.
17. Harris NL, Hunter L, States DJ. Mega-classification: discovering motifs in massive datastreams. In: Proceedings of the 10th national conference on AI. Cambridge, MA: The MIT Press; 1992. p 837–842.
18. Corpet F, Gouzy J, Kahn D. Recent improvements of the ProDom database of protein domain families. Nucleic Acids Res 1999;27:263–267.

19. Watanabe H, Otsuka J. A comprehensive representation of extensive similarity linkage between large numbers of proteins. Comp Appl Biosci 1995;11(2):159–166.
20. Koonin EV, Tatusov RL, Rudd KE. Protein sequence comparison at genome scale. Methods Enzymol 1996;266:295–321.
21. Neuwald AF, Liu JS, Lipman DJ, Lawrence CE. Extracting protein alignment models from the sequence database. Nucleic Acids Res 1997;25:1665–1677.
22. Park J, Teichmann SA, Hubbard T, Chothia C. Intermediate sequences increase the detection of homology between sequences. J Mol Biol 1997;273:349–354.
23. Bateman A, Birney E, Durbin R, Eddy SR, Finn RD, Sonnhammer EL. Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. Nucleic Acids Res 1999;27:260–262.
24. Hofmann K, Bucher P, Falquet L, Bairoch A. The PROSITE database, its status in 1999. Nucleic Acids Res 1999;27:215–219.
25. Attwood TK, Flower DR, Lewis AP, Mabey JE, Morgan SR, Scordis P, Selley J, Wright W. PRINTS prepares for the new millennium. Nucleic Acids Res 1999;27:220–225.
26. Henikoff JG, Henikoff S, Pietrokovski S. New features of the Blocks Database servers. Nucleic Acids Res 1999;27:226–228.
27. Gracy J, Argos P. Automated protein sequence database classification. I. Integration of copositional similarity search, local similarity search and multiple sequence alignment. II. Delineation of domain boundaries from sequence similarity. Bioinformatics 1998; 14(2):164–187.
28. Ponting CP, Schultz J, Milpetz F, Bork P. SMART: identification and annotation of domains from signalling and extracellular protein sequences. Nucleic Acids Res 1999;27:229–232.
29. Tatusov RL, Eugene VK, David JL. A genomic perspective on protein families. Science 1997;278:631–637.
30. Barker WC, Pfeiffer F, George DG. Superfamily classification in PIR-international protein sequence database. Methods Enzymol 1996;266:59–71.
31. Krause A, Vingron M. A set-theoretic approach to database searching and clustering. Bioinformatics 1998;14(5):430–438.
32. van Heel M. A new family of powerful multivariate statistical sequence analysis techniques. J Mol Biol 1991;220:877–887.
33. Ferran EA, Pflugfelder B, Ferrara P. Self-organized neural maps of human protein sequences. Protein Sci 1994;3:507–521.
34. Hobohm U, Sander C. A sequence property approach to searching protein database. J Mol Biol 1995;251:390–399.
35. Bairoch A, Boeckman B. The SWISS-PROT protein sequence data bank. Nucleic Acids Res 1992;20:2019–2022.
36. Yona G. Methods for global organization of the protein sequence space. Ph.D. thesis, The Hebrew University, Jerusalem, Israel; 1999.
37. Linial M, Linial N, Tishby N, Yona G. Global self organization of all known protein sequences reveals inherent biological signatures. J Mol Biol 1997;268:539–556.
38. Altschul SF, Boguski MS, Gish WG, Wootton JC. Issues in searching molecular sequence databases. Nature Genet 1994;6: 119–129.
39. Henikoff S, Henikoff JG. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci USA 1992;89:10915–10919.
40. Compugen LTD. BIOCCELERATOR Manual. http://www.compugen.co.il
41. Pearson WR. Empirical statistical estimates for sequence similarity searches. J Mol Biol 1998;276:71–84.
42. Pearson WR. Comparison of methods for searching protein sequence databases. Protein Sci 1995;4:1145–1160.
43. Altschul SF. Amino acid substitution matrices from an information theoretic perspective. J Mol Biol 1991;219:555–565.
44. Karlin S, Altschul SF. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. Proc Natl Acad Sci USA 1990;87:2264–2268.
45. Wootton JC, Federhen S. Statistics of local complexity in amino acid sequences and sequence databases. Comp Chem 1993;17:149–163.
46. Chothia C. One thousand families for the molecular biologist. Nature 1992;357:543–544.
47. Green P, Lipman D, Hillier L, Waterston R, States D, Claverie JM. Ancient conserved regions in new gene sequences and the protein databases. Science 1993;259:1711–1716.
48. Wang Z. How many fold types of protein are there in nature? Proteins 1996;26:186–191.
49. Brenner SE, Chothia C, Hubbard TJP. Population statistics of protein structures: lessons from structural classifications. Curr Opin Struct Biol 1997;7:369–376.
50. George DG, Barker WC, Mewes HW, Pfeiffer F, Tsugita A. The PIR-International protein sequence database. Nucleic Acids Res 1996;24:17–20.
51. Nuoffer C, Balch W. GTPase: multifunctional molecular switches regulating vesicular traffic. Annu Rev Biochem 1994;63:949–990.
52. Downward J. The ras superfamily of small GTP-binding proteins. Trends Biochem Sci 1990;15:469–472.
53. Harpez Y, Chothia C. Many of the immunoglobulin superfamily domains in cell adhesion molecules and surface receptors belong to a new structural set which is close to that containing variable domains. J Mol Biol 1994;238:528–539.
54. Smith DK, Xue H. Sequence profiles of immunoglobulin and immunoglobulin-like domains. J Mol Biol 1997;274:530–545.