The metric space of proteins—comparative study of clustering algorithms

Ori Sasson[1], Nathan Linial[1] and Michal Linial[2]

[1]*School of Computer Science and Engineering and* [2]*Department of Biological Chemistry, Institute of Life Sciences, Hebrew University, Jerusalem 91904, Israel*

ABSTRACT

Motivation: A large fraction of biological research concentrates on individual proteins and on small families of proteins. One of the current major challenges in bioinformatics is to extend our knowledge to very large sets of proteins. Several major projects have tackled this problem. Such undertakings usually start with a process that clusters all known proteins or large subsets of this space. Some work in this area is carried out automatically, while other attempts incorporate expert advice and annotation.

Results: We propose a novel technique that automatically clusters protein sequences. We consider all proteins in SWISSPROT, and carry out an all-against-all BLAST similarity test among them. With this similarity measure in hand we proceed to perform a continuous bottom-up clustering process by applying alternative rules for merging clusters. The outcome of this clustering process is a classification of the input proteins into a hierarchy of clusters of varying degrees of granularity. Here we compare the clusters that result from alternative merging rules, and validate the results against InterPro.

Our preliminary results show that clusters that are consistent with several rather than a single merging rule tend to comply with InterPro annotation. This is an affirmation of the view that the protein space consists of families that differ markedly in their evolutionary conservation.

Availability: The outcome of these investigations can be viewed in an interactive Web site at http://www.protonet.cs.huji.ac.il.

Supplementary information: Biological examples for comparing the performance of the different algorithms used for classification are presented in http://www.protonet.cs.huji.ac.il/examples.html.

Contact: ori@cs.huji.ac.il

Keywords: protein families; protein classification; sequence alignment; clustering.

INTRODUCTION

Recent years have seen an explosive growth in the amount of biological data gathered by the scientific community. Specifically, the number of publicly available protein sequences is growing rapidly, primarily as a result of many large-scale sequencing projects, including that of the human genome.

The large volumes of data collected make it necessary to automatically classify and sort such data on a very large or even global scale. However, currently used methods, mostly those based on automated procedures, have had limited success in attempts to infer proteins' function (Bork and Koonin, 1998). A multitude of techniques exists for sequence comparison (Needleman and Wunsch, 1970; Smith and Waterman, 1981; Lipman and Pearson, 1985; Althschul *et al.*, 1990). Nevertheless, we are still far from being able to determine in general whether two proteins share the same function. It is well known that sequence similarity or structural similarity imply a high likelihood for similar biological function. Furthermore, as more protein sequences are determined from DNA sequences and gene discovery prediction methods, we encounter an increasing number of poorly annotated protein sequences. Likewise, more sequences are labelled as 'hypothetical proteins'.

Several techniques were proposed for detecting weak homologies among proteins. Such methods may incorporate structural similarity scores as in FSSP (Holm and Sander, 1997), or phylogenetic information as in COG (Tatusov *et al.*, 2000). These classifications as well as those based on standard sequence comparison algorithms do not appear to provide a systematic and rigorous way to predict proteins' functionality. Several papers attempt to evaluate the potential of achieving two major goals: (i) Making functional predictions at the whole genome level and (ii) Assigning function to non-annotated proteins by extrapolating the function of their relatives (Fleischmann *et al.*, 1999; Marcotte *et al.*, 1999; Bilu and Linial, 2001; Di Gennaro *et al.*, 2001).

The shortcomings of plain sequence analysis algorithms give rise to attempts to classify proteins through clustering. Clustering can be based on the simple observation that homology is by definition a transitive relation. By clustering sequences into groups based on similarity, one may expect to discover relations that direct sequence comparisons

fail to uncover. The rationale here is that if a sequence A is similar to a sequence B, and B is similar to C, sequences A and C might exhibit function similarity, even if they do not necessarily have high sequence similarity. In other words, clustering may reveal unexpected relationships among protein sequences that sequence comparisons are unable to discern. A major reason for that is sequence similarity is not transitive, whereas homology (and biological function) is transitive. As is well known and as indicated below, transitivity has its perils and must carried out with great care.

Protein classification algorithms are roughly divided to those based on motif and domain analyses and to those that rely on whole protein (reviewed by Kriventseva *et al.*, 2001a). When whole proteins are taken as the elementary units, several difficulties arise due to the fact that many proteins consist of multiple structural/functional domains. Consequently, transitivity may result in classifying unrelated proteins that share some highly conserved domains.

The advantages of clustering proteins using transitivity have been observed previously (Yona *et al.*, 2000; Bolten *et al.*, 2001). This concept has been studied and implemented in various systems, such as ProtoMap (Yona *et al.*, 2000), SYSTERS (Krause *et al.*, 2000), ClusTr (Kriventseva *et al.*, 2001b), Picasso (Heger and Holm, 2001) and MetaFam (Silverstein *et al.*, 2001).

Our clustering methods depend on standard similarity measure, namely gapped BLAST. We rely on the notion of *restricted transitivity* (see Yona *et al.*, 1999) in order to perform a continuous process of clustering. As this process progresses, we discover larger clusters that rely on progressively weaker similarity relations. In our previous work (e.g. Yona *et al.*, 2000) we used predetermined thresholds to construct the hierarchy, which resulted in discrete, somewhat arbitrary, stages. Here we allow the procedure of clustering to progress continuously so that the resulting clusters have different levels of granularity. The consistency of the results is validated through comparison with InterPro annotation. InterPro is an integrated documentation source that combines information from several independent domain-based systems including SMART, ProSite, Pfam, Prints and ProDom. According to the coverage of InterPro with proteins in our clustering we provide a validated flexible view on the resulting clusters.

Another significant aspect of our work is a method that allows for incremental updating of the current classifications using current clusters as anchors for new sequences. This facilitates the incorporation of large amounts of new protein sequences with minimal need for recalculation (to be described elsewhere).

METHODS

This section details the computational aspects of our work, including the required pre-computation and the clustering algorithms.

Pre-computation

We begin our clustering process with a comprehensive all-against-all sequence comparison. This is done using standard gapped BLAST based on BLOSUM62 with filtration of low complexity sequences. BLAST associates a numerical value (the E-Score) with each pair of proteins. Pairs with very low (or no) similarity, receive a very large (or infinite) E-Score. When this score exceeds a predetermined threshold value, the threshold value is used as the score. In this present work this threshold value is set at 10, well above any expected direct biological significance. Comparison of two proteins with E-Score 10 or above rarely shows any significant similarity.

Note that this cut-off value is significantly higher than that used in previous works (e.g., Yona *et al.*, 2000). As we explain below, this choice allows us to detect weak but biologically relevant relations.

It was already established (see e.g. Portugaly and Linial, 2000) that similarity among clusters at such low levels of confidence does encode a good deal of significant biological information. In previous systems such information was either absent or present in a very noisy form. In the latter case its interpretation required an expert view. A major advance of the present study is that the levels of noise are reduced and the important biological information becomes much more apparent.

Clustering methodology

Our clustering method is an adaptation of the widely accepted hierarchical clustering paradigm. The schematic clustering algorithm is as follows:

```
procedure clustering()
 {
  for each protein p {
    create_cluster(p);
  }
  t = 0;
  while ( not done )
  {
   find clusters x,y such that
    merge_score(x,y) is minimal;
   merge_clusters(x,y,t);
   t++;
   done = finished();
  }
}
```

This procedure makes use of several subfunctions. The procedure `create_cluster(p)` takes a protein and

creates a singleton cluster that includes this protein alone. The function `merge_score(x,y)` associates a numeric value with the merging of clusters x and y. The issue of merging scores is addressed in the next subsection.

The function `merge_clusters(x,y,t)` takes two clusters and creates a new cluster that is the union of the input clusters x,y. This step takes place in time t.

The function `finished()` implements the termination rule for the clustering. A variety of termination rules can be implemented. In this work we focus on termination rules derived from the number of non-singleton clusters generated.

Merging rules

When merging two clusters, we seek the most beneficial merge step. In a metric-space setting this typically entails merging two clusters to minimize the diameter of the new cluster. In the context of clustering proteins, this is based on the *E*-Score of pairs of proteins in a cluster. To capture the typical inter-protein distance within a cluster, we consider averages of *E*-score. Previously (in Yona *et al.*, 1999), we considered only one kind of average; here, we look at several kinds. In particular, the score of a cluster is the appropriate mean of pair-wise distances

- *Arithmetic mean*
- *Square (l_2) mean*
- *Geometric mean*
- *Harmonic mean*

The definition of the various means follows. For numbers $x_1, x_2, \ldots, x_n$, the arithmetic mean is defined as:

$$\text{ArithMean}(x_1, x_2, \ldots, x_n) = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

the geometric mean is:

$$\text{GeoMean}(x_1, x_2, \ldots, x_n) = \sqrt[n]{x_1, x_2, \ldots, x_n}$$

and the harmonic mean is:

$$\text{HarMean}(x_1, x_2, \ldots, x_n) = \frac{n}{\frac{1}{x_1} + \frac{1}{x_2} + \cdots + \frac{1}{x_n}}$$

A simple set of inequalities relates all these averages. The harmonic mean never exceeds the geometric mean which is less than or equal to the arithmetic mean which is in turn less than or equal to the square root of the arithmetic mean of squares.

This comes into play in the clustering process by the way weak similarities are considered. Compared with other averaging schemes, the arithmetic mean of squares places more weight on weak similarity scores (large *E*-values). This tendency weakens as we move to the arithmetic mean, geometric mean, and finally harmonic mean. Similarly, the harmonic mean puts much more emphasis on strong similarities (small *E*-values) than does the geometric mean (and so on for the rest of the means).
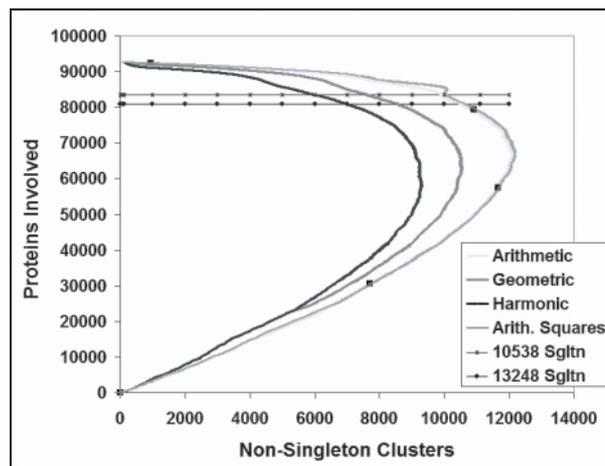

Fig. 1. The number of proteins participating in the clustering process versus the number of non-singleton clusters for each of the four merging algorithms used. Note that the point at which non-singleton clusters number starts dropping shows the following order: Harmonic, Geometric, Arithmetic, and Arithmetic square. The two horizontal lines represent two levels in the hierarchy with 10 538 and 13 248 singletons for the top and bottom line, respectively.

RESULTS

Pre-computation results

The underlying database used in this work is the SWIS-SPROT database release 39, containing 94 153 proteins. Using gapped BLAST with BLOSUM62, we have identified 6 871 715 relations among these proteins (i.e., sequence similarity *E*-Score of 10 or less).

The number of relations is quite large, due to the choice of a high cut-off value.

Comparison of merging rules and termination rules

In order to compare the various merging rules we investigate the number of proteins involved in the clustering process, in terms of the number of non-singleton clusters. This progression is shown in Figure 1. This figure reflects the aforementioned inequality tying the 4 averages used. The harmonic mean, being the smallest, creates the smallest number of clusters, and starts 'imploding' first.

Figure 1 can also assist us in selecting termination rules. In hierarchical classification all proteins are organized into nested sets of clusters. The merging criteria create an (artificial) rooted tree for each of the methods. A certain number of proteins are deemed to be singletons, as they do not exhibit significant similarity to any other protein. 2007 of the proteins have no relation whatsoever (not even a BLAST *E*-value of 10).

To validate the results of our clustering we have to determine where to 'cut' the tree, or determine a proposed

Table 1. Number of clusters generated for each of the four merging methods when stopping at two levels of the hierarchy. The levels are determined by the number of singletons (i.e., single proteins not clustered yet)

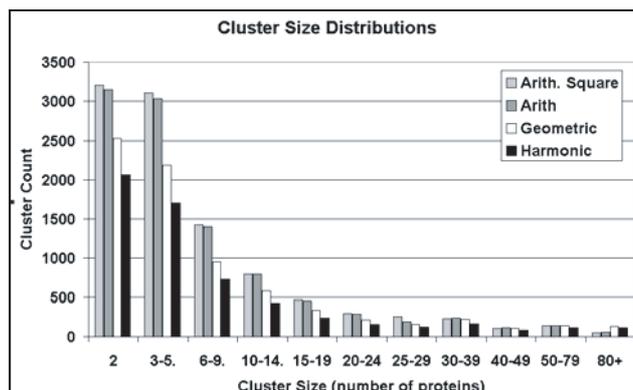| Method	#Clusters (10 538 singletons)	#Clusters (13 248 singletons)
Harmonic	5925	6991
Geometric	7553	8544
Arithmetic	9858	10437
Arith Square	10030	10626

Fig. 2. Distribution of cluster sizes for each of the merging rules, analysed at a level of 10 538 singletons. Note that the Geometric and Harmonic merging methods have a significantly larger number of clusters with 80 proteins or more.

level. It is known that the number of proteins which should correspond to singleton clusters may comprise a substantial fraction of the entire protein sequences. This is especially true in the case of protein deduced from genomes with no other phylogenetic relatives.

Based on the common perception of the database at hand (SWISSPROT in this case), we expect 10–20% of the proteins to remain as 'singletons' in the final clustering. To be more accurate, if we use $1E-5$ as a threshold for cutting off sequence similarities, then 10 538 of the 94 153 proteins should be 'singletons', i.e., populate a cluster of their own. This setting corresponds to a horizontal line (at the 83 615 level) in the graph shown in Figure 1. If we use a more conservative threshold of $1E-10$, the number of singletons is 13 248. The number of non-singleton clusters in each of the methods is shown in Table 1.

We study the distribution of cluster sizes using the first termination point (10 538 singletons). The results indicate that the distribution of cluster sizes is very different among the 4 tested merging methods and that the geometric method produces a larger fraction of bigger clusters (size >80 proteins each, Figure 2).

Some biological examples of generated clusters

In this section we look at some specific protein families and try to track the behaviour of the clustering algorithms. To this end, we need to use some cluster numbering system. The system we use is quite straightforward. Each of the proteins is numbered with an identifier (ID) from 1 to 94 153. A similar numbering system is used for each of the clustering methods. The serial number for each singleton cluster corresponds to the ID of the protein it represents, and the non-singleton clusters are assigned running numbers starting with 94 154.

In order to evaluate the quality of the resulting clustering, we tested a few tens of established protein families based on Protein Profiles (http://www.ebi.ac.uk/proteinprofiles) and on other knowledge-based sites. In the following section we perform a systematic evaluation based on comparison with Inter-Pro. Because of a lack of space we will describe only few selected cases. Other examples are presented in http://www.protonet.cs.huji.ac.il/examples.html.

Our first example looks at histone and histone fold (Sullivan *et al.*, 2002). Eukaryotic DNA associates with histones to form nucleosomes. Each nucleosome consists of a compact core containing histones of the core histone proteins: H2A, H2B, H3 and H4 that are wrapped with DNA. Histones H2A/H2B and H3/H4 dimerize through their histone fold motif while histones H1 and H5, which are structurally distinct and do not resemble the core histones, bind as monomers in linker DNA.

Histones are quite unique in that constraints on nucleosome structure in all archea and eukaryotes have preserved them with minimal changes during evolution. There are small differences in sequences throughout evolution, which probably relate to some early event of histone gene duplication. Histone H2A, H2B, H3 and H4 sequences form distinct classes (and have 4 distinct InterPro annotations).

Interestingly, the histone structural fold is shared with other DNA binding proteins such as TAFs that are part of the transcription TFIID complex. The evolutionary relationships between the members of the core histones and other histon fold-containing proteins are largely ambiguous but it was suggested that H2A and H4 histone folds had diverged prior to the appearance of eukaryotes.

The clustering profile of histone H2A members in all 4 clustering method is summarized in Figure 3. This figure clearly shows the superiority of the geometric hierarchy, in the sense that it clusters the evolutionary remote histone core proteins. The purity of its clusters drops only at the top of the hierarchy with less than 1000 clusters or so. In the geometric hierarchy, all H2A were collected in a 75-protein cluster (cluster #159 927, at ∼10 000 non-singleton clusters). Later in the process (cluster #179 689) the number of proteins is doubled to 152 proteins, and the
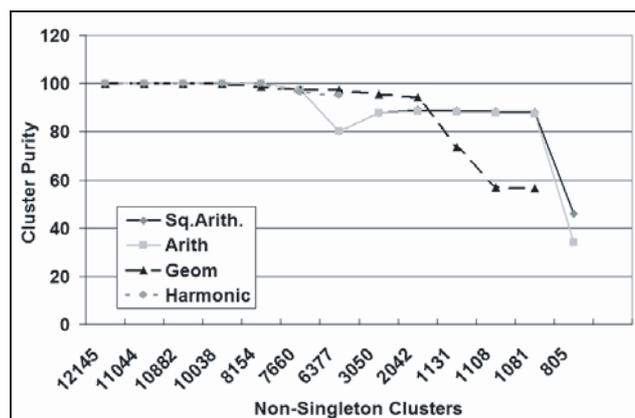
Fig. 3. Cluster purity as a function of the number of non-singleton clusters, for clusters containing H2A proteins, for all four merging methods. The line stops before reaching a cluster of 1000 members and above. Each point in this graph corresponds to a cluster.

Fig. 4. Breakdown for cluster 183 028 of the geometric merging method according to the protein associated with different histone subfamilies. The cluster includes 290 proteins.

H4 proteins join the cluster. It is worth mentioning that at this point the clustering process makes use of weak similarities between protein sequences. Many very low scoring pairs of proteins (with E-Score ranging from 1 to 10) participate in the clustering process at this low level.

Going further in the clustering process provides even more intriguing results. In cluster #183 028 the number of proteins has grown to 290, where 273 are within the histone fold including all 4 elements of the core. The chart in Figure 4 shows the InterPro classification breakdown for the 290-member cluster. This figure highlights the fact that our geometric clustering identifies the relation between H2A and H4 histones without adding significant amounts of 'noise' (in the form of non-histone proteins). The fact that the H4 elements join the H2A elements in the same cluster appears to imply that the clustering is quite powerful. The relation between H2A and H4 is well known from a biological function perspective, but is difficult to identify with automated procedures. The order by which the different core histone families merged into a combined cluster follows the evolutionary pathway predicted (Thatcher and Gorovsky, 1994). Applying PSI-BLAST iterative search (until convergence) failed to detect this weak but evolutionary relevant map of the histone core proteins. Interestingly, proteins marked as 'others' (Figure 4) consist of DNA binding transcriptional activators having a histone fold.

The other clustering methods do not fare as well. In the Arithmetic method, already at cluster #177 019 (105 proteins) the coverage of histone H2A is only 70% with non-related proteins such as cdc25 are included. A similar phenomenon occurs with the arithmetic-squares method.
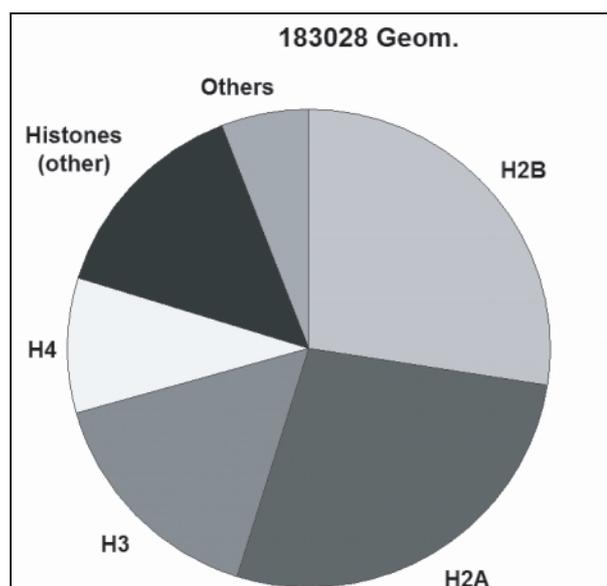
The harmonic merging rule did generate a pure clus-

ter covering the entire H2A family with 85 proteins (#174 786) but did not progress further to accumulate other related histones. Instead, further clustering leads to generating a huge cluster ($>31 000$ proteins).

Another family of proteins we look at is that of actin and actin-like proteins (Fyrberg *et al.*, 1994). The clustering generated for this family appears to be quite 'stable' for all methods. Each of the methods identifies the $\sim$250 actins present in SWISSPROT. Note that the best coverage occurs at very different levels in the hierarchy for each of the methods.

Specifically, in geometric merging, cluster #177 226 (at $\sim$5500 non-singleton clusters) contains 259 actins (out of 262 proteins). In harmonic merging only 236 proteins are identified (#135 194, at $\sim$9000 non-singletons). The arithmetic merging exhibits the best behaviour by detecting all 261 actin proteins (#179 099, at $\sim$5200 non-singletons). Further on, Arithmetic clustering only moderately adds non-actin proteins (e.g. #183 708, 261 actins out of 272 proteins, at $\sim$1700 non-singletons). Note that in this example, the geometric method quickly reaches a large non-pure cluster while all other 3 methods gradually accumulate non-actin proteins.

An additional interesting case is that of the cyclophilin family. Cyclophilins are high-affinity binding protein for immunosuppressive drug cyclosporin A in vertebrates and other organisms. Another group of proteins that shares similar enzymatic activity are proteins that bind the immunosuppressive drug FK506, named FKBPs.

From the information provided by InterPro, it seems that sequences from this family are not related to that of cyclophilin and thus assigned with a different InterPro label. Inspecting the clustering procedure revealed that both harmonic and geometric methods collected the entire set of cyclophilin (∼80 proteins). However, during clustering via the harmonic method, a group of ∼50 KEBP proteins were joined (#166 549, 142 proteins, 6300 non singletons) while in the case of the geometric method, other (non-related) proteins were combined (#178 458, 355 proteins, 5100 non singletons). This example suggests that while in most cases the geometric method outperforms the other methods we considered, this is not always true.

To conclude this section, we look at the Wnt family. The Wnt genes encode a large family of secreted growth factors that function in determining cell fate, proliferation, migration, polarity, and cell death. The family members are found in vertebrates and invertebrates. Following classification of Wnt proteins using different merging methods we found that all 4 methods were able to detect eventually all Wnt proteins (137 proteins). Surprisingly, all merging methods indicated a weak connection to phospholipase A2 (PLA2) family (∼150 proteins). The PLA2 are small enzymes that release fatty acids and are involved in numerous physiological processes. Most tested cases show that each of the merging methods attracts non-overlapping sets of proteins during the clustering process (as described above; for more examples, see http://www.protonet.cs.huji.ac.il/examples.html). The case of Wnt-PLA2 and similar instances in which all merging methods result in clustering similar sets of proteins are of special interest and invite deeper biological inspection. In summary, we believe that at junctures where the clustering result is questionable, agreement among several merging methods could corroborate the results.

Validation against InterPro

In order to evaluate the properties of our clustering systematically, we cross-validate our clusters against the InterPro classification. This validation method can provide meaningful results only for protein families which have an InterPro annotation (about 70% of all proteins in SWISSPROT database).

We focus on the geometric merging rule and study the correspondence between the generated clusters and InterPro classification. This correspondence is somewhat problematic since the underlying database of protein sequences is not identical. InterPro provides about 5000 entries that represent the cumulative information taken from Pfam, PRINTS, PROSITE, ProDom, SMART and TIGRFAM. In our analysis we consider clusters for which at least one protein has an InterPro annotation. For such clusters, the level of purity is measured as the fraction of proteins in the cluster which have the InterPro
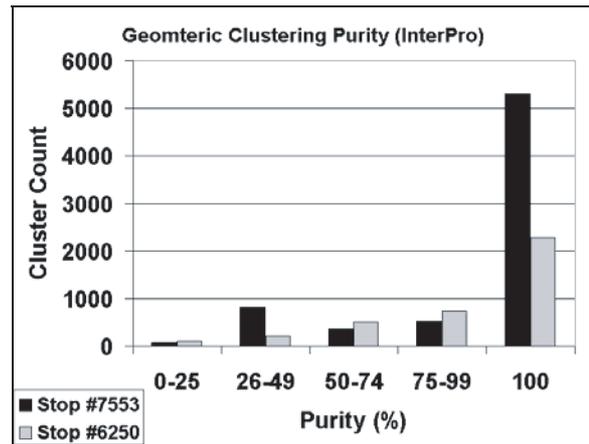

Fig. 5. Representation of the degree of cluster purity, compared against InterPro family and domain annotations. The results shown are at two levels of the hierarchy of clustering process at 7553 and 6250 non-singleton clusters.

annotation most common in the cluster. Clearly proteins that are not assigned with InterPro entry may belong to the biologically relevant cluster but may represent protein fragments or proteins whose signatures do not fully agree with the classifications signatures used by InterPro. This method of measurement is therefore biased against our clustering since all proteins not classified in InterPro are automatically assumed to be false positives.

When analysing the hierarchy according to the termination rule described (Table 1, 7553 non-singleton clusters), the number of clusters with any InterPro classification is 3899. Out of these, 3090 exhibit 100% purity in terms of their InterPro classification. Clearly, we would like to present a coarser level that allows exposure of remote evolutionary connections as illustrated for some biological examples (see http://www.protonet.cs.huji.ac.il/examples.html).

By allowing the clustering process to continue until a level of only 6250 clusters, the clustering obtained becomes less pure (in terms of InterPro classification). However, we obtain larger clusters which still exhibit a relatively high level of purity (out of a total of 2640 clusters, 2201 are of purity 75% and above).

Figure 5 shows the distribution of cluster purity for these two different levels of hierarchy. Note that the cluster purity percentage is essentially the complement of the number of false positives (e.g., 80% cluster purity is the equivalent of 20% false positives).

An important aspect of the validation process is the number of false negatives, namely the protein annotated by a certain InterPro entries but are not included in the analysed cluster. Table 2 shows the distribution of false

Table 2. False negatives for clusters of size $\geq$ 50 proteins with high purity (>75%), compared against InterPro, at a level of 6250 non-singleton clusters

% False Negatives	# Clusters (in these clusters)	Total # Proteins
0	103	9854
1–10	96	14684
11–25	27	5104
26–50	31	3310
51–75	21	2078
75–100	10	825
Total	288	35855

negatives for the large clusters of high purity (75% and above) for the second level considered in Figure 5. False negatives for a cluster are measured against the most common InterPro entry in the cluster. The percentage of false negatives shown is the fraction of the respective InterPro entry not included in the cluster (e.g., for a cluster of size 100 with 0 false positives, whose entries belong to an InterPro entry with 125 proteins, the false negatives value is 25%).

DISCUSSION

This paper presents a technique for clustering a database of proteins and studies several clustering merging rules. Our results show that this clustering provides valuable information when analysing the function of a new protein, as well as understanding the function of existing proteins in the database.

The clustering results exceed those provided by other automated clustering methods not based on protein names or keywords. We attribute this to the fact our clustering continues further beyond other clustering systems, and does so using weak relations (obtained by low-similarity values in the BLAST output).

We are able to obtain a biological meaningful clustering while reducing the number of clusters significantly compared to Picasso (Heger and Holm, 2001). Picasso provides 10 000 clusters based on about 150 000 proteins using hierarchical classification with profile-based approach.

In our comparison of the various merging score formulas, the geometric scoring system appears to be the most useful. However, we conjecture that some of the other scoring systems might be useful in specific settings, or might be used in conjunction with the other merging methods.

From a general perspective, the problem of clustering proteins can be divided into two separate sub-problems. For certain parts of the proteins space, proteins exhibit high levels of similarity that stem from evolutionary conservation. In these parts of the protein space, it is quite easy to identify clusters, even without employing sophisticated clustering techniques. The remainder of the proteins poses a more challenging problem. More sophisticated tools are needed there, since weak relations are more difficult to analyse correctly. It is tricky at times to distinguish weak relations from random relations. The use of several merging rules provides a natural tool for tackling this difficulty.

Our clustering method lends itself to working with variable resolutions in different parts of the tree. The results described in this paper imply that it is possible to reduce the number of clusters without significantly damaging the quality of the clustering. We intend to investigate this issue further and to develop a set of automated rules, which will allow a varying scale of clustering in different parts of the clustering hierarchy. This would provide a useful tool for navigating in protein space.

The clustering system described in this paper is available online at http://www.protonet.cs.huji.ac.il.

REFERENCES

Althschul,S.F., Carrol,R.J. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 410.

Bilu,Y. and Linial,M. (2001) On the predictive power of sequence similarity in yeast. *Proceedings of the Fifth Annual International Conference on Computational Biology*. pp. 39–48.

Bolten,E., Schliep,A., Schneckener,S., Schomburg,D. and Schrader,R. (2001) Clustering protein sequences-structure prediction by transitive homology. *Bioinformatics*, **17**, 935–941.

Bork,P. and Koonin,E.V. (1998) Predicting functions from protein sequences-where are the bottlenecks? *Nature Genet.*, **18**, 313–318.

Di Gennaro,J.A., Siew,N., Hoffman,B.T., Zhang,L., Skolnic,J., Neilson,L.I. and Fetrow,J.S. (2001) Enhanced functional annotation of protein sequences via the use of structural descriptors. *J. Struct. Biol.*, **134**, 232–245.

Fleischmann,W., Moller,S., Gateau,A. and Apweiler,R. (1999) A novel method for automatic functional annotation of proteins. *Bioinformatics*, **15**, 228–233.

Fyrberg,C., Ryan,L., McNally,L., Kenton,M. and Fyrberg,E. (1994) The actin protein superfamily. *Soc. Gen. Physiol. Ser.*, **49**, 173–178.

Heger,A. and Holm,L. (2001) Picasso: generating a covering set of protein family profiles. *Bioinformatics*, **17**, 272–279.

Holm,L. and Sander,C. (1997) Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Res.*, **25**, 231–234.

Krause,A., Stoye,J. and Vingron,M. (2000) The SYSTERS protein sequence cluster set. *Nucleic Acids Res.*, **28**, 270–272.

Kriventseva,E.V., Biswas,M. and Apweiler,R. (2001a) Clustering and analysis of protein families. *Curr. Opin. Struct. Biol.*, **11**, 334–339.

Kriventseva,E.V., Fleischmann,W., Zdobnov,E.M. and Apweiler,R. (2001b) CluSTr: a database of clusters of SWISS-PROT+TrEMBL proteins. *Nucleic Acids Res.*, **29**, 33–36.

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

Marcotte,E.M., Pellegrini,M., Thompson,M.J., Yeates,T.O. and Eisenberg,D. (1999) A combined algorithm for genome-wide prediction of protein function. *Nature*, **402**, 83–86.

Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search of similarities in the amino acid sequence of two proteins. *J. Mol. Biol.*, **48**, 443–453.

Portugaly,E. and Linial,M. (2000) Estimating the probability for a protein to have a new fold: A statistical computational model. *Proc. Natl Acad. Sci. USA*, **97**, 5161–5166.

Silverstein,K.A., Shoop,E., Johnson,J.E. and Retzel,E.F. (2001) MetaFam: a unified classification of protein families. I. Overview and statistics. *Bioinformatics*, **17**, 249–261.

Sullivan,S., Sink,D.W., Trout,K.L., Makalowska,I., Taylor,P.M., Baxevanis,A.D. and Landsman,D. (2002) The histone database. *Nucleic Acids Res.*, **30**, 341–342.

Smith,T.F. and Waterman,M.S. (1981) Comparison of biosequences. *Adv. Appl. Math.*, **2**, 428–489.

Tatusov,R.L., Galperin,M.Y., Natale,D.A. and Koonin,E.V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res.*, **28**, 33–36.

Thatcher,T.H. and Gorovsky,M.A. (1994) Phylogenetic analysis of the core histones H2A, H2B, H3 and H4. *Nucleic Acids Res.*, **22**, 174–179.

Yona,G., Linial,N. and Linial,M. (1999) ProtoMap: Automatic classification of protein sequences, a hierarchy of protein families, and local maps of the protein space. *Proteins*, **37**, 360–378.

Yona,G., Linial,N. and Linial,M. (2000) ProtoMap: Automatic classification of protein sequences and hierarchy of protein families. *Nucleic Acids Res.*, **28**, 49–55.