Towards a complete map of the protein space based on a unified sequence and structure analysis of all known proteins

Golan Yona* and Michael Levitt
Department of Structural Biology, Fairchild Bldg. D-109, Stanford University CA 94305, USA.
* Corresponding author, email: golan@gimmel.stanford.edu
Tel +1 650-725-0754, Fax +1 650-723-8464

Abstract
In search for global principles that may explain the organization of the space of all possible proteins, we study all known protein sequences and structures. In this paper we present a global map of the protein space based on our analysis. Our protein space contains all protein sequences in a non-redundant (NR) database, which includes all major sequence databases. Using the PSI-BLAST procedure we defined 4670 clusters of related sequences in this space. These clusters are centered on a sequence of known structure. All 4670 clusters were then compared using either a structure metric (when 3D structures are known) or a novel sequence profile metric. These scores were used to define a unified and consistent metric between all clusters. Two schemes were employed to organize these clusters in a meta-organization. The first uses a graph theory method and cluster the clusters in an hierarchical organization. This organization extends our ability to predict the structure and function of many proteins beyond what is possible with existing tools for sequence analysis. The second uses a variation on a multidimensional scaling technique to embed the clusters in a low dimensional real space. This last approach resulted in a projection of the protein space onto a 2D plane that provides us with a bird's eye view of the protein space. Based on this map we suggest a list of possible target sequences with unknown structure that are likely to adopt new, unknown folds.

Keywords: protein sequence classification, protein structure classification, protein map.

Introduction
Today sequence analysis techniques are used daily by thousands of biologists to elucidate relationships between novel genes and existing proteins. These techniques have been carefully optimized for detecting remote homologies and were thoroughly studied. Almost all these techniques, which have been thoroughly tested, are based on pairwise comparisons of the query sequence with the sequences in one or more of the sequence databases.

Copyright © 2000, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.

Over a decade ago, biologists began to realize that the amount of biological data accumulated in the databases can no longer be analyzed solely by means of pairwise sequence comparison. Thus powerful tools to analyze and organize this data became a necessity. New methods for utilizing the information in these databases have been applied, such as multiple alignment techniques, profile analysis, and pattern recognition methods. These techniques enhanced the ability to detect relationships between distantly related proteins. An extensive use of these tools, combined with expert knowledge helped to derive databases of protein families and domains (e.g. PROSITE (Hofmann et al. 1999), Blocks (Henikoff et al. 1999), PRINTS (Attwood et al. 1999), Pfam (Bateman et al. 1999), ProDom (Corpet et al. 1999), Cogs (Tatusov et al. 1997), ProtoMap (Yona et al. 1999)). For a review of the different methodologies that were used to cluster protein sequences into families see (Yona et al. 1999; Yona 1999).

Today, database searches can be applied with much success and can help to identify the biological function of new protein sequences, as well as to reveal many distant and interesting relationships between protein families. In fact, to date more sequences have been putatively characterized by database searches than by any other single technology. Yet, these tools still fail to assign a function for a substantial fraction of the protein space. Consequently, between 40% and 50% of the sequences in the sequence databases have an unidentified function (Pennisi 1997; Doolittle 1998).

The direct and most obvious reason is our limited ability to trace the common origin of related, homologous, proteins. Detecting homology can help to assign a putative function to a new protein sequence, since it implies evolutionary relationship. Homologous sequences usually have similar fold and close or related biological function. However, in many cases sequences have diverged to such an extent that sequence similarity is undetectable by current means of sequence comparison. Structural similarities are even more elusive since they are not well defined. Moreover, structure comparison algorithms are mostly heuristic and can fail to detect all structural similarities. In some cases functional similar-
ity does not even imply similar fold (e.g. the Cellulases family (Wilson & Irwin 1999)).

A major drawback of current schemes of sequence analysis is that they are restricted to local considerations (pairwise comparisons). In other words, they do not provide us with a global view of the protein space. We believe that a global view (a "map") of this space can compensate for much of the missing information. Such a map is expected to add insights about the nature of new proteins and can help to explore high order organization in the space, to identify clusters of related proteins, to gain a better understanding of the geometry of this space and to derive general principles that will explain the way this complex system was created.

Another drawback of current large-scale studies of proteins is that the analysis is limited to either protein sequences or structures. The essential difference between the representation of a protein as a sequence of amino acids, and its representation as a 3D structure, dictated different methodologies, different similarity/distance measures and different comparison algorithms. Consequently, sequence based techniques were traditionally applied to a subspace of protein sequences, while structure based techniques were applied only to the subspace of known structures (see SCOP (Hubbard et al. 1999), CATH ( Orengo et al. 1997) and FSSP (Holm & Sander 1997a)).

Clearly, classification of protein structures based on structural similarities is extremely important. Structure is often conserved more than sequence, and detecting structural similarity may help infer function, allowing more accurate predictions of functional roles. However, structural information is available only for several thousand proteins, while the number of known sequences is over 300,000. Given the difficulty of determining the 3D structure of a protein, sequence based studies play a major role in genome analysis. Therefore, to accurately map the protein space it seem logical to combine sequence based techniques with structure-based techniques. Several sequence-structure studies were carried out in the last few years (e.g. [Elphsson & Somlaihammer 1999, Han & Baker 1996, Rigoutsos et al. 1999]). However, to the best of our knowledge, none of these studies tried to combine both structure-based metrics and sequence-based metrics to map the protein space.

In the study described in this paper we try to bridge the gap between the sequence space and the structure space. Our study attempts to establish a consistent, reliable and unified framework for sequence and structure analysis, and to map the whole space of protein sequences and structures. By these means we wish to extend our ability to predict the biological function of proteins based on their location in this map.

**Methods**

**Detecting clusters of related sequences**

Our computational procedure starts by defining clusters of homologous sequences within the SCOP database. The domains in this database provide a natural definition of the basic building blocks of protein structures. These sequences are clustered first into families using sequence similarity alone, without incorporating any information from the SCOP classification. These clusters are then used as seed clusters in the search for homologs in the space of all known sequences. The related sequences are marked and the clusters are extended accordingly.

**Databases** The SCOP database, release 1.39, serves as the starting point for our analysis. This release contains 15,198 sequences which are classified into 7 classes, 440 folds, 640 superfamilies, 938 families and 2716 protein domains. This defines the space of protein structures.

Our protein sequence space consists of all known protein sequences. We combine all major databases of protein sequences to create non-redundant (NR) database. Specifically, we include SWISSPROT, TREMBL, TREMBL new, PIR, GenPept, GenPept new, SCOP, PDB, NRL3D, and the complete genomes of yeast, C. elegans, and 20 bacteria available from the NCBI ftp website. All databases were up to date on June 15 1999. Our non-redundant database was created out of all these databases by removing exact duplicates and all entries shorter than 20 amino acids. This database contains a total of 378,407 sequences and 119,881,584 amino acids.

**Type-I clusters** The sequences of the SCOP database are clustered first based on all-against-all sequence similarities, using the ProtoMap clustering algorithm (Yona et al. 1999). Pairwise similarities were calculated using Capped-BLAST (Altschul et al. 1997) with the BLOSUM 62 scoring matrix (Henikoff & Henikoff 1992).

The ProtoMap clustering algorithm uses a graph representation of the sequence space and a hierarchical two-phase clustering algorithm to automatically cluster the protein sequences (for more details on the algorithm see (Yona et al. 1999)). At the end of the clustering process we end with 1421 clusters. For each cluster with more than one sequence we select a seed sequence. This is the sequence whose average distance from all other members in the cluster is the minimal (or equivalently, similarity is maximal).

Each sequence-based family is represented by a profile, and this profile is used to search the NR database, using PSI-BLAST. For all seed clusters with 2 or more members, the search starts with the seed sequence and the profile that was generated from all the sequences in the cluster. Otherwise, the search starts with a single sequence.

Although PSI-BLAST is a powerful tool, it can lead
to false positives by diverging from the original query sequence, and in doing so, create a profile representing unrelated sequences. To combat such drift, several different indices of validity are used to control the procedure. These indices are based on statistics of clusters such as growth rate and self-score, and were used to detect suspicious clusters and "correct" them. In all cases where the profile has diverged, we run the procedure again, stopping the iterations before divergence occurs. For more detail see (Yona & Levitt 2000).

Clusters based on SCOP domains are called type-I clusters. In all we find 1421 type-I clusters with total of 168,431 sequences and 34,259,323 amino acids. This is 44.5% of all sequences in the NR database, and 28.5% of the amino acids in our database.

Type-II clusters Once all members of type-I clusters have been marked, we analyze the remaining sequences by applying the PSI-BLAST procedure repeatedly. Each round, a random query is selected from the set of undetected sequences, a search is performed with that query, and the set of detected sequences is updated. The resulting clusters may overlap as the search is performed against the whole NR database (including the sequences in type-I clusters). The same validation criteria discussed above are used to control the PSI-BLAST process, and the clusters are refined accordingly. The resulting clusters are called type-II clusters.

At the time of writing, 3883 randomly selected query sequences had been used as search seeds for type-II clusters, giving 3249 clusters (there are 634 in singletons). Overall, type-II clusters contain 180,633 sequences and 37,791,626 amino acids. We are still in the process of generating all the type-II clusters.

The largest 20 type-I clusters are given in Table 1. The complete statistics about type-I and type-II clusters is available on the BioSpace website http://biospace.stanford.edu.

### Combining the sequence-based metric with the structure-based metric

Sequence based metrics often miss weak similarities between sequences that have diverged greatly. In many cases proteins may have the same fold and close biological function without significant sequence similarity (Murzin 1993; Pearson 1997; Brenner et al. 1998). Using structural information where available to deduce relationship between proteins can greatly increase the accuracy of predictions and functional analysis. Therefore, we incorporate structural metric in our analysis. However, structural information is currently available for only a small fraction of the protein space. Therefore, to achieve maximum sensitivity and benefit from both structures and sequences, we need to combine these two metrics.

Currently, our protein space consists of 1421 type-I clusters and 3883 type-II clusters. It is reasonable to assume that some of type-II clusters are related to the structure-based type-I clusters, since very weak sequence similarities might have been missed by PSI-BLAST. If we had a 3D structure for a representative sequence in each type-II cluster, such relationships could have been detected. However, in the absence of structural information for these clusters we must rely on sequence information. Although we have already used the most powerful tools available for data base searches, we have not fully utilized the sequence information in these clusters. Specifically, we believe that by creating a profile (e.g. PSI-BLAST profile) for each cluster and comparing these profiles we can enhance our ability to detect remote homologies. These profiles can provide a new, potentially powerful measure of similarity between clusters. For this task we have developed a novel procedure for profile-profile comparison. Detail follow.

It should be noted that the unification of these two measures is not trivial. Since they are based on different protein features and different considerations no simple relation exists between the two metrics, and one should be careful not to arbitrarily bias the combined metric in favor of one metric over the other. Fortunately, both measures use the same statistical framework, what enables us to derive a unified measure of similarity between clusters. In both cases the distribution of scores of random matches follow the extreme value distribution (Gumbel 1958) as is shown below, and therefore the statistical estimates that are obtained from these

<table>
<thead>
<tr>
<th>Cluster number</th>
<th>Size</th>
<th>No. of amino acids</th>
<th>Family</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>17,596</td>
<td>1,857,330</td>
<td>Immunoglobulin (variable domain)</td>
</tr>
<tr>
<td>1296</td>
<td>15,862</td>
<td>1,823,363</td>
<td>Envelope Glycoprotein</td>
</tr>
<tr>
<td>891</td>
<td>5,514</td>
<td>1,862,581</td>
<td>Tropomyosin, Myosin, Kinesin (coiled coil segments)</td>
</tr>
<tr>
<td>11</td>
<td>4,945</td>
<td>1,251,316</td>
<td>Protein kinase</td>
</tr>
<tr>
<td>51</td>
<td>4,123</td>
<td>1,245,273</td>
<td>Ribulose bisphosphate carboxylase (large chain) domain II</td>
</tr>
<tr>
<td>127</td>
<td>4,026</td>
<td>523,363</td>
<td>Ribulose bisphosphate carboxylase (large chain) domain lb</td>
</tr>
<tr>
<td>146</td>
<td>3,640</td>
<td>396,697</td>
<td>Ribulose bisphosphate carboxylase (large chain) domain lb</td>
</tr>
<tr>
<td>3</td>
<td>3,583</td>
<td>374,549</td>
<td>Polypeptide, Protease</td>
</tr>
<tr>
<td>13</td>
<td>2,726</td>
<td>273,938</td>
<td>Reverse transcriptase</td>
</tr>
<tr>
<td>95</td>
<td>2,658</td>
<td>610,565</td>
<td>Homeobox</td>
</tr>
<tr>
<td>88</td>
<td>2,407</td>
<td>137,968</td>
<td>MHC class II</td>
</tr>
<tr>
<td>177</td>
<td>2,194</td>
<td>175,560</td>
<td>MHC class I</td>
</tr>
<tr>
<td>85</td>
<td>2,037</td>
<td>307,682</td>
<td>Gap protein</td>
</tr>
<tr>
<td>134</td>
<td>1,867</td>
<td>200,024</td>
<td>ADP-ribosylation factor, Ras-like proteins, GT-Binding proteins</td>
</tr>
<tr>
<td>453</td>
<td>1,854</td>
<td>286,378</td>
<td>Gag protein</td>
</tr>
<tr>
<td>542</td>
<td>1,780</td>
<td>184,894</td>
<td>Hydroxysteroid dehydrogenase</td>
</tr>
<tr>
<td>42</td>
<td>1,545</td>
<td>319,017</td>
<td>EF hand (Calmodulin, Troponin C, Parvalbumin)</td>
</tr>
<tr>
<td>7</td>
<td>1,514</td>
<td>191,313</td>
<td>Ras-like proteins</td>
</tr>
<tr>
<td>31</td>
<td>1,486</td>
<td>218,200</td>
<td>Ras-like proteins</td>
</tr>
<tr>
<td>212</td>
<td>1,460</td>
<td>81,843</td>
<td>Transcription factors</td>
</tr>
</tbody>
</table>

**Table 1: Twenty Largest Type-I clusters.** The preliminary family description states the feature common to most or many of the member proteins.
distributions can be used to derive a unified metric.

**Structural comparisons** Our procedure starts by calculating structural similarities between all seed proteins of type-I clusters (see section “Type-I clusters”). These matches were calculated using **strucutal** (Gerstein & Levitt 1998). The output of this program gives the RMS of the match between the two structures as well as the statistical significance of the match, i.e., the probability that the observed structural similarity could have been obtained by chance. The statistical estimate is based on the distribution of random matches, that was shown to follow the extreme value distribution (Levitt & Gerstein 1998).

![Image 1: Significant structural similarity between proteins from different SCOP classes.](image)

Out of the 1,008,910 type-I cluster pairs, 5185 pairs share significant structural similarity with e-value better than 1e-2. An additional 3841 pairs were found with structural similarity e-value better than 1e-1 (all the structural alignments used here will be made available on the BioSpace website). To evaluate the sensitivity and selectivity of this algorithm in detecting weak relationships between protein families, we have checked these similarities with respect to the SCOP classification. Clearly, proteins that belong to the same superfAMILY or fold are expected to share substructures. In some cases even proteins with different folds within the same class may show structural similarity. Even belonging to different classes does not necessarily entail that the proteins cannot be structurally similar as most SCOP classes share secondary structure elements. Out of the seven classes in SCOP (all alpha, all beta, alpha/beta, alpha and beta, multi-domain proteins (alpha and beta), membrane and cell surface proteins, and small proteins) only two classes (all-alpha and all-beta) do not suppose to share any secondary structure elements and therefore are not expected to be structurally similar.

A structural pairwise similarity that was reported as significant is considered a true relationship if both structures belong to the same SCOP class, an error if one protein is all-alpha and the second is all-beta, and suspicious otherwise. In what follows we adhere to these definitions. Out of all 9026 pairwise similarities detected with e-value < 1, 7842 are true relationships, 1163 are suspicious and only 21 are errors. Out of the 7842 true relationships, 3502 are between different families and superfamilies within the same fold (this is 41% of all pairs of clusters with the same fold), and additional 4340 pairs are of clusters with different folds within the same class.

Most of the 1163 suspicious connections are due to similarity along a relatively small substructure (few secondary structure elements). Such similarities are often observed between proteins from different SCOP classes and are not necessarily false similarities. Even most of the 21 similarities that are defined as errors are actually true similarities, and the error actually results from the SCOP definition of domains. For example the most significant error (with e-value = 3.1e-4) was reported between a short helical protein and an all-beta protein that contains three helices (see Fig. 1).

The SCOP classification is an excellent resource, which is based on extensive expert knowledge. Nonetheless, one should keep in mind that this is a man-made classification, and the definitions of domains, folds and classes do not necessarily conform with “nature’s definitions”. Therefore, any assessment using SCOP may be biased due to errors in that reference classification. The definition of domain boundaries is especially tricky since no consistent rules or principles are used. In such cases, the histogram of sequence matches we obtain, may help to detect errors and refine domain boundaries accordingly, as exemplified for d2gpc2 in Fig. 2.

The most significant pair of clusters with different folds within the same class has an e-value of 4.5e-9. Both proteins (dleur_pdb, cluster 1015 and dgoa_pdb, cluster 309) are beta-propeller barrel proteins and their alignment (Fig. 3) suggests that another level of hierarchy between the fold level and the class
level of SCOP should be introduced. We hope that our map can suggest a new level of granularity that is currently missing from SCOP.

Comparing profiles. A few procedures to compare profiles are reported in the literature. Gotoh (Gotoh 1993) proposed an iterative alignment method to align two groups of biological sequences, that incorporates profile-based operations. However, his method is based on optimizing a weighted sum-of-pairs score, and essentially compares pairs of sequences, with overall computation time that is proportional to the product of the numbers of sequences comprising the two groups. Pietrokovski (Pietrokovski 1996) compared profiles that were generated from multiple alignments of protein families in the Blocks database, but his method does not allow gaps in the alignment. Lyngso et al. (Lyngso et al. 1999) used the co-emission probability of two profile hidden markov models to measure their similarity. Despite the mathematical elegance of their approach, the metrics that are proposed are overly sensitive to the differences between the concentration of the probability distributions these models induce on sequences, and to the size of the training data. In other words, their proposed metrics emphasize the differences rather than the similarity of the two models, and it is not clear that the method can detect subtle similarities between protein families. Variations on these metrics may produce a more sensitive measure.

Our algorithm for profile-profile comparison is based on the classical dynamic programming algorithm. The novel ingredient in our procedure is the definition of profile similarity scores. To estimate the statistical similarity between two (empirical) probability distributions we utilize the Jensen-Shannon measure $D^{JS}$ (Lin 1991). We define the score of a match between $p$ and $q$ as a combination of its statistical similarity $D$ and its significance $S$

\[
\text{Score}(p,q) = \frac{1}{2}(1 - D)(1 + S)
\]

\[
= \frac{1}{2}(1 - D^{JS}(p||q))(1 + D^{JS}(r||p_0))
\]

where $p_0$ is the background distribution (defined as the amino acid distribution in our NR database), and $r = 0.5p + 0.5q$ can be considered as the most likely common source of both $p$ and $q$. This scoring scheme assigns a score close to one for two similar distributions (i.e. $D = D^{JS}(p||q) \to 0$), while assigning a score close to zero for two dissimilar distributions whose most likely common distribution resembles the background distribution. For more details, see appendix A.

Finally, a mass conserving transformation was applied to these scores, so as to make them suitable for detecting local sequence similarities. This transformation maps the distribution of JS similarity scores onto the distribution of BLOSUM62 matrix so as to preserve the “mass” along the BLOSUM62 distribution. The major benefit of this transformation is that it allows us to use the same gap penalties that were obtained from exhaustive optimization of parameters for sequence comparison (Henikoff & Henikoff 1993; Pearson 1995). The procedure is described in more detail in the appendix.

![Distribution of profile similarity scores](image)

**Figure 4:** Distribution of profile similarity scores (for cluster number 9). The distribution follows an extreme value distribution. The distributions for other clusters resemble this distribution.

The significance of profile similarity scores is estimated based on the distribution of those scores for unrelated, effectively random profiles. For each cluster the distribution of scores with all other clusters is derived and an extreme value distribution is fitted to that distribution after excluding high-scoring profiles.
(see Fig. 4). Based on the theoretical fit we estimate the statistical significance (e-value) of raw similarity scores.

![Distribution of type-I profile similarity scores by e-value](image1)

**Figure 5:** Distribution of profile similarity scores by e-value. Top: distribution of profile similarity scores for type-I clusters. At e-value ≤ 1 we detect 350 true relationships, 34 suspicious connections and only 5 errors. Bottom: distribution of profile similarity score for all clusters (logscale). The distribution suggests that the e-value threshold should be set to 1, and verifies that our statistical estimates are indeed reliable.

In Fig. 5 we show the distribution of profile similarity scores at different e-values, after normalization for the search space size. There are 242,366 cluster pairs with profile similarity score e-value < 50. Of these, 159,319 are pairs of type-II clusters, 74,204 are mixed pairs involving type-I and type-II clusters, and 8,843 are pairs of type-I clusters.

To evaluate the performance of our profile-profile comparison algorithm we have tested it in the same way as for structural similarities. Of the 8,843 type-I pairs that are detected by profile-profile comparison at e-value < 50, 3712 are true relationships, 4734 are suspicious and only 397 are errors. Of the 3712 true relationships, 1410 are different superfamilies within the same fold (16.6% of all such pairs), 1194 are different families within the same superfamily (31% of all such pairs) and 790 are different subfamilies within the same family (50.5% of all such pairs). None of these similarities could be detected using the simple BLAST search

(see section “Type-I clusters”). Only 673 pairs were detected by PSI-BLAST (the clusters shared one or more members), but 110 of these had less than 100 amino acids in common. 1096 pairs were detected by structural comparison, of which 365 had more significant profile-profile similarity scores.

Setting the threshold more conservatively at e-value < 10, detects 1372 true relationships, 718 suspicious connections and 80 errors. 704 of the true relationships were detected by structural, and 572 by PSI-BLAST. One would expect to set the e-value threshold at = 1 (see Fig. 5). This gives 350 true relationships, 34 suspicious connections and only 5 errors. Of the true relationships, 240 were detected by structural, and 244 by PSI-BLAST.

These numbers are promising in view of recent evaluations (Park et al. 1998) that point to 30-35% success with iterative methods such as PSI-BLAST and SATTM98 (Hughley & Krogh 1998). Our success rate is even higher, considering the fact that our test set is more difficult than the one that was used in (Park et al. 1998), as the sequence similarity between any of our seed sequences is not significant according to BLAST and the percent identity is well below 40% (the threshold that was used in (Park et al. 1998)). Therefore, we believe that that this new procedure can increase the sensitivity of our map.

![Significant profile similarity between proteins from different SCOP classes](image2)

**Figure 6:** Significant profile similarity between proteins from different SCOP classes. Left: cytochrome c551.5, dnew (SCOP class 1.107.1.21). Right: the N-terminal domain of transcription factor IIA (TFIIA) dlytfd (SCOP class 2.42.1.1.1). Though dnew is defined in SCOP 1.39 as an α/α protein it contains two beta strands, which are matched, based on profile similarity, with dlytfd.

As for structural similarities, many of the suspicious connections between clusters are due to a common substructure or few common secondary structure elements. Some of the error connections are also due to a short common motif as demonstrated in Fig. 6. To evaluate the true error rate, we are currently checking the correlation of profile-profile matches with matches along secondary structure elements.

**Results - constructing the unified map**

The final step in our analysis is the construction of the map by integrating all the information we have accumu-
Table 2: Largest super-clusters that contain both type-I and type-II clusters.

<table>
<thead>
<tr>
<th>Super-cluster number</th>
<th>Num. of type-I clusters</th>
<th>Num. of type-II clusters</th>
<th>SCOP families</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3</td>
<td>105</td>
<td>heterogeneous</td>
</tr>
<tr>
<td>2</td>
<td>42</td>
<td>1</td>
<td>fold 3.1 beta/alpha (TIM)-barrel</td>
</tr>
<tr>
<td>7</td>
<td>23</td>
<td>1</td>
<td>class 3 (Flavodoxin-like, alpha/beta-Hydrolases, Formyltransferase)</td>
</tr>
<tr>
<td>8</td>
<td>14</td>
<td>9</td>
<td>fold 2.1 (mostly the Immunoglobulin superfamily)</td>
</tr>
<tr>
<td>11</td>
<td>17</td>
<td>1</td>
<td>class 3 (mostly NAD(P)-binding Rossmann-fold domains)</td>
</tr>
<tr>
<td>13</td>
<td>13</td>
<td>1</td>
<td>superfamily 1.3.1 Cytochrome c</td>
</tr>
<tr>
<td>17</td>
<td>11</td>
<td>1</td>
<td>class 1 (mostly &quot;Winged hecs&quot;, DNA-binding domain)</td>
</tr>
<tr>
<td>19</td>
<td>9</td>
<td>1</td>
<td>family 7.3.9.1 EGF-like module</td>
</tr>
<tr>
<td>22</td>
<td>9</td>
<td>1</td>
<td>class 7 (Laminin, Metaflavonoid, Flavoidin)</td>
</tr>
<tr>
<td>25</td>
<td>6</td>
<td>3</td>
<td>superfamily 3.53.1 S-adenosyl-L-methionine-dependent methyltransferases</td>
</tr>
<tr>
<td>26</td>
<td>1</td>
<td>8</td>
<td>family 4.108.1.1 gpl20 core</td>
</tr>
<tr>
<td>30</td>
<td>7</td>
<td>1</td>
<td>superfamily 3.29.1 P-loop containing nucleotide triphosphate hydrolases</td>
</tr>
<tr>
<td>31</td>
<td>7</td>
<td>1</td>
<td>superfamily 2.35.1 Trypsin-like serine proteases</td>
</tr>
<tr>
<td>36</td>
<td>6</td>
<td>1</td>
<td>superfamily 3.28.1 Thiamin-binding</td>
</tr>
<tr>
<td>39</td>
<td>3</td>
<td>4</td>
<td>superfamily 1.22.1 Histon-like</td>
</tr>
</tbody>
</table>

The graph-based approach - detection of super-clusters

The pairwise cluster similarity scores (based either on the structural similarities or the profile-profile similarities) are used to cluster the clusters into super-clusters, using the ProtoMap clustering algorithm (Yona et al. 1999). Based on our preliminary data, the 4,670 clusters (excluding the 634 type-II singletons) form 2,789 superclusters of which 774 contain more than one cluster. Many connections are observed between superclusters but they are currently rejected by the ProtoMap clustering algorithm because of insufficient statistical evidence for a relationship. Fine tuning of the algorithm may lead to further aggregation of clusters.

SCOP sequences (or type-I clusters) are scattered among 611 superclusters. This number is bigger than the number of folds in SCOP (440) and is slightly less than the number of superfamilies in SCOP (640). Only 22 superclusters contain SCOP entries from different classes (heterogenous superclusters), with a total of 96 clusters within these superclusters (less than 7% of all type-I clusters). Within superclusters that contain at least one type-I cluster we find 312 type-II clusters, of which 201 are in homogeneous superclusters. Recall that these similarities were not detected by PSI-BLAST, and hence we believe that this map will extend predictions of structural and functional similarities beyond that obtainable with existing methods. The largest superclusters that contain type-I clusters as well as type-II clusters are listed in Table 2.

The projection map

To embed the protein space into an Euclidean space we used the classical approach to dimensionality reduction, namely the multi-dimensional scaling method (see (Duda & Hart 1973)). This method represents the data points as points in some lower-dimensional space in such a way that the distances between the points in the lower-dimensional space correspond to the dissimilarities between points in the original space. A stress function which compares proximity values with Euclidean distances of the corresponding points (for example, a sum-of-squared-errors function) is used to measure the quality of the embedding, and a gradient descent procedure is applied to improve the embedding till a local minimum, i.e. local optimum of the error function, is reached.

A major drawback of multi-dimensional scaling (MDS) techniques is that the deterministic algorithms (gradient descent) usually applied to solve this problem, tend to get trapped in a local minimum which can be far from optimal. A random projection usually serves
as the initial embedding, what increases the chances of getting trapped in a suboptimal configuration. It is possible to use stochastic techniques, like simulated annealing or deterministic annealing (Klöck & Buhmann 1997), to reduce the probability of being trapped in a local minimum, but for a considerable cost in computation time.

The results that we are presenting here are based on a modification of the MDS algorithm. The distances are weighted so as to preserve small distances while allowing larger distortion for large distances. To reduce the dependency on the initial seed, the starting configuration is created by mapping close clusters to adjacent points within a ball whose radius is the distance between the clusters.

In Fig. 7 we show the results of this method when applied to the 2789 super-clusters described in the previous section. The distances between super-clusters are based on the overall connections between them (those are rejected merges in ProtMap terminology (Yona et al. 1999), and their quality is defined as the geometric mean of all pairwise connections between clusters). In this 2D projection of the protein space only 970 super-clusters are shown, since the rest made no connections and could not be positioned in this map at the time of writing.

![Figure 7: A 2D projection of the protein space. Note that this is a preliminary map, and we are still in the process of optimizing it.](image)

Having plotting the map we had to assess its validity. Tests with respect to SCOP classification show that super-clusters that belong to the same fold are generally mapped to the same vicinity in this map (see Fig. 8), but we also observe counter examples. As a reassuring test, note that all-alpha proteins occupy one part of the space, while all-beta proteins occupy another part, and the overlap between the two parts is small (see Fig. 9).

![Figure 8: Distribution of super-clusters that correspond to the Knottin fold. Families with this fold include small inhibitors, toxins and lectins.](image)

Having plotting the map we had to assess its validity. Tests with respect to SCOP classification show that super-clusters that belong to the same fold are generally mapped to the same vicinity in this map (see Fig. 8), but we also observe counter examples. As a reassuring test, note that all-alpha proteins occupy one part of the space, while all-beta proteins occupy another part, and the overlap between the two parts is small (see Fig. 9).

![Figure 9: Distribution of super-clusters corresponding to SCOP classes 1 and 2. Top: class 1 (all-alpha). Bottom: class 2 (all-beta).](image)

![Another way to obtain a better configuration is to start with a non-random seed, based on some a-priori information we may have about the space. For example, classical linear embedding techniques, such as principal component analysis (PCA), can provide a reasonable initial seed. However, these techniques require the data to come from a real normed space while the space of protein sequences is represented by proximities. To overcome this, we made use of what we call the canonical representation of protein sequences (Yona, unpublished results, (Dubnov et al. 2000)). Given a protein family or a group of related proteins, the new representation is based on the statistical significance of the match of each sequence with all the other sequences in the group. This assigns a position in a high dimensional real space for each protein sequence, where every coordinate is associated with one member and its value is correlated with its similarity to the particular protein. The representation leads to the definition of a new distance measure among samples. Monotonicity tests show that this metric mostly preserves the original structure of the space. Therefore, we believe that this is a faithful representation of the protein data in an Euclidean space. Another approach for embedding is distance geometry (Havel et al. 1983). Both those directions are currently being tested.](image)


Discussion

One of the most important problems in genomics is the clustering of related proteins in the protein space. Usually proteins are grouped using either sequence similarity or structure similarity (when structural information is available). Here we have tried to combine both measures of similarity in order to build a unified map of the protein space.

So far we have clustered the protein space into type-I clusters (clusters with structural representatives) and type-II clusters (clusters with no structural representative). Higher level measures of similarity were then used to group these clusters into superclusters. Specifically, we made use of efficient and sensitive structural comparison algorithm, and a novel method for profile-profile comparison. Our method is also designed to be compatible with standard scoring schemes of sequence-sequence comparison. This new method outperforms the currently available means of sequence comparison and detects subtle similarities between clusters. The significance of scores is assessed using the same statistical framework as for structural similarities. Taken together, these tools enable us to derive a single sequence/structure metric. Using a graph-based method we cluster type-I clusters and type-II clusters into superclusters. Finally, by applying a multidimensional scaling procedure we create a global map of the protein space.

The work described here is part of a multi-stage analysis that aims to organize the protein space into domains as well as into protein families, to create a uniform framework for sequence and structure analysis, to build a 3D model for a substantial part of the protein space, and finally to plot a map of the protein space.

The implications of this study are many. First there is the large scale modeling of protein sequences. Since each of the type-I 1421 clusters is associated with a seed sequence that has a known 3D structure, we can build a 3D model for all other sequences in the cluster based on the similarity with the seed sequence. The PSI-BLAST alignment as well as the known structure of the SCOP representative have been used to create an all-atom 3D model for each of the 168,431 sequences (34,259,323 amino acids), using SegMod (Levitt 1992) and EnCad (Levitt et al. 1995). The alignments as well as the models are available on the web, at the BioSpace website http://biosa.bioinformatics.stanford.edu.

Second, it can help to select target sequences for structural genomics. It has been a major goal of structural genomics to select those targets for structure determination that will enable reliable modeling of as many sequences as possible. Recent efforts concentrate on identifying those sequences that are likely to have a new fold, since by solving the structure of these sequences we will be able to significantly extend our knowledge of the structural repertoire in the protein space. Our study can be useful in this aspect as well. The probability for having a new fold corresponds to distances in our map, and the target sequences are those that represent clusters that are far from known structures. Since this organization is hierarchical, clusters that are mapped to the same vicinity are ranked and only one member of each super-cluster is chosen. At lower priority other cluster centroids within the same super-cluster can be selected. A preliminary sample list is given in appendix B. The full list of targets will be available at the BioSpace website.

![Figure 10: Eigenvalues of the clusters covariance matrix, after embedding in Euclidean space. Only the first 50 eigenvalues are shown. The eigenvalues were calculated by applying the singular value decomposition method (SVD) to the sample covariance matrix. At least 10 eigenvalues appear to be significant. See footnote 3 on page 8 for details.](image-url)

Third, we intend to use this study as a base for multi-structure classification of proteins. A typical protein comprises domains and motifs. Each part has a specific role, but it is the combination that assigns a function to a protein sequence. By studying these combinations we search for new levels of hierarchy within the protein space that are unknown today.

Naturally, one would be concerned that large multi-domain proteins may impose a problem in our approach as they may connect unrelated families. It has not slipped our mind and our computational scheme was designed so as to address this issue. For example, the decision to define type-I clusters based on SCOP sequences was made following this line. Yet, this is a still ongoing project and we have not finished processing type-II clusters. Study of sequence-sequence correlation within each cluster can help to detect the existence of several domains, or to refine domain boundaries as is demonstrated in Fig. 2. Also, a better scheme for defining type-II clusters will be employed. In the current scheme, a PSI-BLAST run with a random representative of a protein family may not detect all family members, and several runs with several different queries may be needed to cover all family members. A wise selection of the seed sequence can minimize the number of such queries and limit the redundancy. Databases such as PIM and ProtoMap which provide pre-defined groups of related sequences can help to reveal these center sequences or can suggest an alternative for type-II clusters. A preliminary analysis of these families, in search for the constituting domains (using the techniques men-
tioned above) would reduce the problems that may arise in the presence of multi-domain proteins. These modifications will be integrated soon in our procedure and are expected to improve and refine our map.

Finally, it should be kept in mind that the non-linear projection of the original space onto a 2D plane as in Fig. 7 does not and cannot fully capture the complexity of the original protein space. There is no bound on the expected distortion after the embedding and some of the pairwise distances may have changed a great deal during the embedding process. Moreover, this map is created based on partial information since not all distances are defined. Therefore, one should be careful when trying to deduce conclusions from this map. This is especially true since our current findings suggest that the intrinsic dimension of this space is much higher than 2, i.e. a few tens of dimensions are needed to capture most of the variance that is observed in this space and faithfully embed it in an Euclidean space (see Fig. 10).

In spite of these reservations, we believe that this projection can provide us with a lot of insight on the geometry of the protein space from which better, more accurate maps can be generated. It can affect and direct the development of new tools for comparison of proteins, and may help to redefine the boundaries between protein “classes”. We have developed several new tools and tested a few approaches for this task. Yet, many new tools still need to be studied, new directions should be explored, means for validation need to be developed and new concepts of similarity, relationship and family or class need to be defined.

Acknowledgments

We thank Steven Brenner for help and advice regarding the SCOP database. Golan Yona is supported by a Burroughs-Wellcome Fellowship from the Program in Mathematics and Molecular Biology (PMMB). This work was supported by DOE award DE-FG03-95ER62135 to Michael Levitt.

Appendix A - Profile comparison

The profile comparison is performed using the classic dynamic programming algorithm, and the alignment is assigned a score that accounts for matches, insertions and deletions. Unlike sequence-sequence comparison, where a match means identity, a profile-profile comparison is more subtle. The core of our new procedure is the definition of profile similarity scores, and the parameters used to quantify this measure of similarity. This was done to obtain maximal sensitivity, as well as to be compatible with standard scoring schemes of sequence-sequence comparison as will be clarified below.

Given two profiles $P = p_1p_2p_3...p_n$ and $Q = q_1q_2q_3...q_m$, where $n$ and $m$ are the lengths of the profiles (the number of positions or columns) and $p_i, q_j$ are probability distributions over the 20 letter alphabet of amino acids, the score of a “match” between two columns $p_i$ and $q_j$ is based on their statistical similarity.

A possible measure of statistical similarity between two (empirical) probability distributions $p_i(x)$ and $q_j(x)$ is the Kullback-Leibler (KL) divergence (Kullback 1959) which is defined as

$$D_{KL} [p_i || q_j] = \sum_{k} p_{ik} \log_2 \frac{p_{ik}}{q_{jk}}$$

This measure has a few disadvantages, being asymmetric and unbounded. A better measure of statistical similarity is the Jensen-Shannon (JS) divergence between probability distributions (Lin 1991). This measure is also known as “divergence to the mean”.

Given two (empirical) probability distributions $p$ and $q$, for every $0 \leq \lambda \leq 1$, their $\lambda$-JS divergence is defined as

$$D_{JS}^\lambda [p||q] = \lambda D_{KL} [p||r] + (1 - \lambda)D_{KL} [q||r]$$

where

$$r = \lambda p + (1 - \lambda)q$$

can be considered as the most likely common source of both $p$ and $q$, with $\lambda$ as a prior. Without apriori information, a natural choice is $\lambda = 1/2$. We denote the corresponding measure by $D_{JS}^\lambda$. This measure is symmetric and ranges between 0 and 1, where the score for identical distributions is 0. It is proportional to the minus logarithm of the probability that the two empirical distributions represent samples from the same (“common”) source (El-Yaniv et al. 1997).

While a statistical measure estimating the probability that two distributions represent the same source distribution seems appropriate for the comparison of profiles, a major ingredient is ignored: the apriori probability of the source distribution. This information can help to assess the significance of a “match”. Say that the two given empirical distributions resemble the overall distribution of amino acids in the database (i.e. the distribution of the common source is similar to the background distribution). Should they be considered significantly similar in that case? Obviously not, as the distribution of the common source is observed for random profiles. Therefore this “match” is not as significant as a “match” of two probability distributions which both resemble a unique distribution (i.e. a distribution which is distinct from the overall distribution of amino acids in the database). In other words, the similarity of two random distributions is not as significant as the similarity of two unique distributions.

To assess the significance $S$ of a match $D$ we measure the JS divergence of the (common) source from the base (background) distribution $P_0$ (defined as the amino acid distribution in our NR database). This score reflects the probability that this distribution could have been obtained by chance. Then, the score of a match between $p$ and $q$ is defined as

$$\text{Score}(p, q) = \frac{1}{2}(1 - D)(1 + S)$$

$$= \frac{1}{2}(1 - D_{JS}^\lambda [p||q])(1 + D_{JS}^\lambda [r||P_0])$$
Thus, the score for two similar distributions (i.e. \( D = D^{\text{SS}}[p][q] \to 0 \)) whose common source is far from the background distribution (i.e. \( S = D^{\text{SS}}[r][P_0] \to 1 \)), approaches 1, while the score for two dissimilar distributions whose most likely common distribution resembles the background distribution approaches zero. This scoring scheme also distinguishes two distributions that resemble the background distribution from two distributions that are far apart, but whose common source resembles the background distribution (assigning a higher score in the first case).

For local alignments, a general scoring scheme \( \text{Score}(a, b) \) should satisfy two requirements: (i) \( E(\text{Score}(a, b)) < 0 \) and (ii) \( a^* = \max\{s(a, b) \} > 0 \). The first requirement implies that the average score of a random match will be negative (otherwise, an extension of a random match would tend to increase its score, and this contradicts the idea of local similarity). The second condition implies that a match with a positive score is possible (otherwise a match would always consist of a single pair of residues). It is necessary that the profile similarity scores be adjusted so as to meet these requirements. A simple transformation would be to subtract a constant from all similarity scores. We applied a more elaborated transformation. The distribution of similarity scores for 1,000,000 profile probability distributions was derived, together with the distribution of similarity scores in the BLOSUM62 matrix. The distribution of profile similarity scores is then mapped onto the distribution of BLOSUM62 matrix so as to preserve the “mass” along the BLOSUM62 distribution. There are clear advantages of applying this kind of transformation. It allows us to use the same gap penalties that were obtained from exhaustive optimization of parameters for sequence comparison (Henikoff & Henikoff 1993; Pearson 1995). Moreover, it allows the augmentation of sequence-sequence comparison with profile-profile comparison and profile-sequence comparison on the same sequence, at the same time. This is useful, for example, when profiles are available only for part of the sequence. A detailed description of the profile comparison procedure and the transformation procedure will be described elsewhere.

Appendix B - target sequences for structural genomics

<table>
<thead>
<tr>
<th>Prot. ID</th>
<th>Length</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>65460</td>
<td>88</td>
<td>trembl: acid.</td>
</tr>
<tr>
<td>66801</td>
<td>89</td>
<td>trembl: polar flagellar assembly protein.</td>
</tr>
<tr>
<td>66348</td>
<td>91</td>
<td>swissprot: (p04967) early e3b protein precursor.</td>
</tr>
<tr>
<td>66448</td>
<td>91</td>
<td>genepnt: (e01128)</td>
</tr>
</tbody>
</table>

Table 3: Target sequences for structural genomics. Protein ID is as in our NR database. These sequences are the seeds of type-II clusters that are positioned far from clusters of known structures. The complete list, ranked by distance from clusters of known structures, is being compiled these days.
References


