

# A Unified Sequence-Structure Classification of Protein Sequences: Combining Sequence and Structure in a Map of the Protein Space.

Golan Yona\* and Michael Levitt

Department of Structural Biology, Fairchild Bld D-109,  
Stanford University CA 94305, USA

\* Corresponding author, email: [golan@gimmel.stanford.edu](mailto:golan@gimmel.stanford.edu)

## Abstract

We analyze all known protein sequences in search for a global map of protein space that is consistent in terms of both sequence and structure. Our goal is to define clusters of homologous protein domains, beyond those detected by sequence-based methods alone, and then to build a three-dimensional (3D) model for each of the sequences that are homologous to sequences of known 3D structure. This analysis uses both sequence and structure based metrics in the analysis of all protein sequences in a non-redundant (NR) database, comprising all major sequence databases.

The analysis starts from the sequences of the SCOP database domains, which have known three-dimensional structures. These sequences are clustered first into families based on sequence similarity alone, without incorporating any information from the SCOP classification. Each sequence-based family is represented by a profile, and this profile is used to search the NR database, using PSI-BLAST. Since PSI-BLAST can lead to false similarities, several different indices of validity are used to control the procedure. Each of the detected sequences is marked and a profile is built for the whole cluster of similar sequences. A 3D model is then built for each sequence in the cluster using an alignment made using the profile as well as the known structures of the SCOP representatives in the cluster. Clusters based on SCOP domains are called type-I clusters. In all we find 1421 type-I clusters with total of 168,431 sequences (44.5% of our NR database).

After all members of type-I clusters have been marked, we analyze the remaining sequences. The PSI-BLAST procedure is applied repeatedly, each time with a different query, to search what is left over from the previous

run. This give type-II clusters, which may overlap.

Type-I and type-II clusters are then grouped using higher level measures of similarity. Those pairs of clusters that contain the same common protein (significant overlap in membership), are marked first. The pairs of clusters are then compared using either a structure metric (when 3D structures are known) or a novel sequence profile metric, and clustered into superfamilies and "fold" families.

This analysis avoids the limitation of classifications that are based just on sequence comparison, and allows us to construct a 3D model for a substantial portion of the sequences in the NR database.

## 1 Introduction

Since the early days of genomic research molecular biologists have tried to make sense out of the accumulated information on protein sequences and structures by classifying proteins into families. The terms "family", "superfamily" and "fold" family were used during the 70's [Dayhoff 1976, Levitt & Chothia 1976], and preliminary tools were developed for the analysis of protein sequences and structures and identification of biological relatedness. During the past two decades there were many attempts to classify proteins based on either sequence analysis or structure analysis. However, the essential difference between the representation of a protein as a sequence of amino acids, and its representation as a 3D structure, dictates different methodologies, different similarity/distance measures and different comparison algorithms. Consequently, sequence based techniques were traditionally applied to a space (or subspace) of protein **sequences**, while structure based analyses were applied only to the space of known **structures**.

As sequence analysis techniques improved and became more sophisticated, better models were created for protein families and domains (e.g. profiles, Hidden Markov Models), allowing subtle relationships to be detected. Several large-scale analyses were carried

out, most of which focused on finding common motifs, domains and patterns of biological significance in protein sequences. This led to the compilation of useful databases, which can be used to search for significant patterns in new sequences. Among these are PROSITE [Hofmann et al. 1999], Blocks [Henikoff et al. 1999], PRINTS [Attwood et al. 1999], ProDom [Corpet et al. 1999], Pfam [Bateman et al. 1999], and Domo [Gracy & Argos 1998]. Other studies were applied to complete proteins. Most draw directly on pairwise comparison and cluster the input database using single linkage clustering [Gonnet et al. 1992, Harris et al. 92, Watanabe & Otsuka 1995, Koonin et al. 1996, Barker et al. 1996, Tatusov et al. 1997, Krause & Vingron 1998, Yona et al. 1999] ProtoMap [Yona et al. 1999] apply somewhat more elaborate considerations. However, in many cases sequences have diverged to such an extent that their common origin is untraceable by all these methods.

Structure based analysis was focused on the classification of protein structures. The proteins of known 3D structure were first grouped into families based on sequence similarity, and then into structural classes and folds based on overall similar shape and architecture. Examples of such structure based classifications are SCOP [Hubbard et al. 1999], CATH [Orongo et al. 1997] and FSSP [Holm & Sander 1997a]. Since structure comparison algorithms are mostly heuristic, and are usually very computationally intensive, expert knowledge was usually required to obtain reliable classifications.

Because structure is often conserved more than sequence, classification of protein structures based on structural similarities is extremely important. For example, structural relationships help understand function, allowing more accurate predictions of functional roles. However, for most known proteins only the sequence is available (currently, only several thousand structures have been determined, 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

Several sequence-structure studies were carried out in the last few years. Some of these studied the correspondence of sequence-based classifications and structure-based classifications (e.g. [Elofsson & Sonnhammer 1999]), while others tried to associate structural properties with sequence patterns [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 this paper we address the problem of bridging 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. Our goal is to map the protein space through a scheme that combines sequence based metrics with structure-based metrics and considers domains as well as entire proteins.

Our analysis starts from a manual classification of protein structures which is considered to be a state-of-the-art in structure classification, namely, the SCOP database. The domains in this database provide a natural definition of the basic building blocks of protein structures. Each of which is a well-defined part of a protein structure which can be assigned a structural or a functional role. However, our main use of this database is not the classification it provides, but the actual definition of these structural units. The sequences of these domains were clustered based on their sequence similarity independent of the SCOP classification. These clusters were then used to identify as many as possible homologs in the NR database. Once these structure-based clusters are defined, the yet unanalyzed areas of the protein space are subjected to sequence analysis. This analysis provides sequence-based clusters much in the same way as the structure-based clusters. The acquired information about the sequence similarity between and within clusters of both types, as well as their structural similarity enabled us to develop a framework for unification of these two metrics

## 2 Methods

Our computational procedure starts by defining clusters of homologous sequences within the SCOP database. These clusters are then used as seed clusters in search for homologs in the space of all known sequences. The related sequences are marked and the clusters are extended accordingly. At the second stage we analyze the sequences which are not covered by these clusters, and cluster them as well. In what follows we describe this procedure in detail

### 2.1 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 form this non-redundant (NR) database by combining all major databases of protein sequences. Specifically, we include the SWISSPROT database release 37 and updates until June 10, 1999 (79,626 entries), the TrEMBL database release 9 minus data integrated into SWISSPROT as of June 10 1999 (200,821 entries), the new preliminary TrEMBL database entries created since release 9 of TrEMBL until June 10, 1999 (42,840 entries), the PIR database

(parts 1,2,3,4) release 60.00, March 31, 1999 (108,716 entries), the GenPept database release 111, April 5, 1999 (387,685 entries) plus new entries added to GenPept until June 15, 1999 (33,485 entries), the SCOP database release 1.39 (15,198 entries), all PDB entries until April 15, 1999 (16,293 entries), the NRL-3D database release March 31, 1999 (14,791 entries), and the complete genomes of yeast, *C. elegans*, and 20 bacteria (*A. fulgidus*, *A. aeolicus*, *B. burgdorferi*, *B. subtilis*, *C. trachomatis*, *C. pneumoniae*, *E. coli*, *H. influenzae*, *H. pylori*, *M. genitalium*, *M. jannaschii*, *M. pneumoniae*, *M. thermoautotrophicum*, *M. tuberculosis*, *P. horikoshii*, *Rhizobium* sp. NGR234, *R. prowazekii*, *Synechocystis* PCC6803, *T. pallidum* and *T. maritima*).

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. It should be noted that due to mutations, errors and different sequence reports the same protein may be stored as a different sequence entry in the source databases, and consequently, will appear as multiple entries in our composite database<sup>1</sup>. However, as will be shown below, this residual redundancy will not affect our results.

## 2.2 Seed clusters - the SCOP database

We begin by eliminating all redundant entries in the SCOP database and discarding all artificial sequences (sequences that are concatenation of discontinuous subsequences). The 6195 remaining sequences are compared pairwise in an all-against-all manner, using Gapped-BLAST [Altschul et al. 1997] with the BLOSUM 62 scoring matrix [Henikoff & Henikoff 1992]. Based on these similarities, the sequences are clustered using the ProtoMap clustering algorithm [Yona et al. 1999]. This algorithm uses a graph representation of the sequence space and a hierarchical two-phase clustering algorithm to automatically cluster the protein sequences. Global properties of the graph are taken into account, and the clustering process is closely monitored to prevent the transitive chaining that is observed for multi-domain proteins, as well as to eliminate false clusters due to chance similarities.

The algorithm starts from a very conservative classification, based on transitive closure of highly significant similarities (expectation value below  $1e^{-100}$ , as measured by BLAST). Subsequently, these clusters are merged to form bigger and more diverse clusters. The procedure operates hierarchically: at each step it adds

<sup>1</sup>The TrEMBL + SWISSPROT databases already account for all sequence entries in all major databases, after eliminating redundancy and false entries and the unification of sequence variations (see release notes of these databases). In the present study we included these variations as different entries, since parsing these variations from the SWISSPROT + TrEMBL entries is cumbersome

new, weaker connections to the previously considered connections. At each round of this process statistical information is gained on the connections between current clusters and used to derive the distribution of pairwise similarities between each pair of clusters. This information is then used to merge only certain clusters (given that their connection is statistically significant), thus forming the next round of larger, coarser clusters. For more details on the algorithm see [Yona et al. 1999, Yona 1999].

We stop the clustering process when the BLAST evalue exceeds  $1e^{-5}$ . This gives 1421 clusters, 762 of which contain at least 2 sequences. The 10 largest clusters are given in table 1. Comparison with SCOP shows that almost all clusters correspond to SCOP families. 18 clusters contain members from different families within the same superfamily. Only 2 clusters contain members from different classes. In both cases there is significant sequence similarity suggesting a biological relatedness.

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). The seed sequence is then searched against all the other sequences in its own cluster using PSI-BLAST [Altschul et al. 1997]. This is an iterative version of BLAST, with a position-specific scoring matrix that is generated from significant alignments found in round  $i$  and used in round  $i+1$ . This algorithm is able to detect weak similarities that are missed in database searches with pairwise sequence searches. At this stage, the use of PSI-BLAST is intended to create a profile which represents all sequences of known structure in the cluster. The full power of PSI-BLAST is exploited in the next stage.

Cluster number	Size	Family
1	288	Immunoglobulin (variable domain)
2	238	Phage T4 lysozyme
3	181	Immunoglobulin (constant domain)
4	113	Globin
5	113	Trypsin(ogen), Thrombin
6	105	Lysozyme
7	82	Calmodulin, Troponin C, Parvalbumin
8	60	Carbonic anhydrase
9	50	Concanavalin A, Lectin
10	43	Mitochondrial cytochrome c, Cytochrome c2

Table 1: Largest SCOP clusters at confidence level  $10^{-5}$

## 2.3 Forming the extended clusters

Each of the 1421 seed clusters of SCOP that we found in the previous stage is compared with all the sequences in our NR database. We use PSI-BLAST to search the database. This choice was made based on evaluations

which proved that this tool is more sensitive in searching for remote homologs [Park et al. 1998]. For all seed clusters with at least 2 members the search starts with the profile that was generated from all the sequences in the cluster. Otherwise, the search starts with a single sequence

The iterative PSI-BLAST process is repeated until it converges, or until the predefined maximum number of 10 iterations is reached. Out of the 1421 clusters, only 64 did not converge. 46 of these did not converge because of noise and divergence, and were "corrected" (see next section). The other 18 did not converge because of fluctuations (15 cases), or high initial threshold (3 cases) and were "approved".

### 2.3.1 Controlling the PSI-BLAST searches

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 that represents unrelated sequences. To avoid this problem, we used a high significance threshold as well as several indices of validity. First we set the threshold at  $1e^{-10}$  (PSI-BLAST evalue). Only hits that scored above this threshold were included in the profile. When a similar analysis was performed with a threshold of  $1e^{-5}$ , it resulted in 263 clusters that did not converge after 10 iterations (as opposed to 64 with the  $1e^{-10}$  threshold). It should be noted that although we set a high significance threshold for inclusion in a profile, many weak but genuine sequence similarities pass the threshold after a few iterations and are incorporated in the profile. Moreover, the threshold for reporting significant hits was set at  $1e^{-2}$ . Thus the results of this procedure are generally as sensitive as simple gapped-blast searches, and in most cases (950 out of 1421) are more sensitive, detecting many more hits than in a simple BLAST search.

Several validity indices, which are based on simple statistics of clusters such as growth rate and self-score, were used to detect suspicious clusters and "correct" them. At this stage most of the analysis involves extensive manual examination of the results. This enabled us to extract rules to control the iterative process, and in many cases we could use these rules to control the process automatically.

### 2.3.2 The rank index

After each PSI-BLAST iteration, a new profile is generated based on all significant hits. In the next iteration, the database sequences are compared with this new profile, and scored accordingly. Since the new profile represents all significant hits (above a certain threshold), it may happen that after few iterations the query sequence is no longer the most significant hit. The profile

may have become more generalized to account for related sequences. If the seed sequence is not a "typical" sequence, then this may result in better score for other, "typical", sequences. We still want the seed sequence to be one of the most significant hits. Therefore, if at some point the seed sequence becomes less significant, we stop the iterative process. Specifically, we require that the seed sequence be one of the 10 most significant hits (where multiple hits with the same score count as one), and stop the process otherwise. Out of 1421 clusters, 135 scored the seed sequence below the first 10 hits at some iteration. Based on manual examination of these clusters, we observe in 105 clusters what we call "shift to center of family", meaning, the profile has diverged from the query sequence and its corresponding subfamily to better represent a family that includes this subfamily. The other 30 cases are due to low complexity sequences and coiled coil segments. In all cases where the profile has diverged, we run the procedure again, stopping the iterations before divergence occurs (see Fig 1).

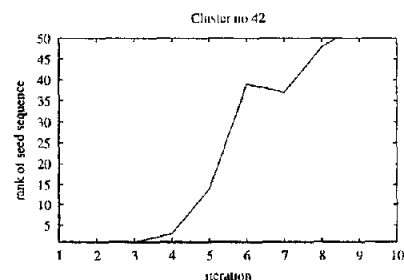


Figure 1. **The rank index.** At the fourth iteration of PSI-BLAST the profile of cluster 42 (mostly 7-alpha-hydroxysteroid dehydrogenases) diverges from the query sequence (Human estrogenic 17beta-hydroxysteroid dehydrogenase), to represent a larger family of reductases and dehydrogenases. Consequently, the seed sequence becomes less significant and is ranked 14 at the fifth iteration. Based on this graph, PSI-BLAST is repeated, this time stopping after 3 iterations.

### 2.3.3 The size index

The rank index helps to detect many cases where the PSI-BLAST process may have led to false hits. We also observe cases in which the cluster keeps growing and yet the query sequence is ranked in the top 10 hits. Such growth may happen for example in cases where sequence similarity is due to convergent evolution (such as coiled coil and transmembrane domains). Most of these cases can be easily traced by checking the growth rate and the final size of clusters. Those clusters that were not detected by the rank index test and in the last iteration are more than 4 times bigger than the size of the cluster at the first round are marked. 41 such clusters were detected, 31 of which were declared suspicious. These suspicious clusters were rerun, this time only for 2 iterations.

At the moment both the rank index and the size test use parameters which are defined ad hoc. These choices were made to facilitate the manual examination of suspicious clusters in a reasonable amount of time. However, these tests initiated a framework for the development of a validated version of PSI-BLAST

### 2.3.4 The SCOP classification test

Though it may happen that different sequences share the same fold [Murzin et al. 1993, Pearson 1997, Brenner et al. 1998], it is very uncommon that similar sequences adopt different folds. Therefore, as another validation test we checked if sequences with different SCOP classifications were placed in the same cluster. Each cluster has a seed cluster in SCOP, and therefore is assigned a specific fold. The seed clusters correspond to either families or superfamilies in the SCOP hierarchy (see section 2.2). Therefore, in each of the "clean, corrected" set of clusters we look for SCOP entries from other superfamilies, folds or classes. Such conflicting entries are detected only in 13 clusters. In 3 clusters the entries belong to different superfamilies within the same fold family. In 4 clusters, the entries belong to different fold families within the same class. The remaining 6 clusters indicate similarity between sequences from different classes. However, manual inspection of all 13 cases showed that in 9 cases the similarities are significant and genuine and the clusters were left as generated. In the remaining 4 cases, the questionable similarities had low significance value, and were not used in the profile building process.

### 2.3.5 The low complexity test

It is well known that low-complexity segments may drastically affect the results of BLAST search [Altschul et al. 1994]. Statistical estimates of significance of similarity scores in database searches assume that the compositions of the compared sequences are similar to the overall composition of amino acids in the database [Karlin & Altschul 1990]. However, this is not true for sequences with low-complexity regions, and statistical estimates overestimate the significance of matches to proteins with unusual amino acid composition. This can be avoided by excluding these segments using filtering programs such as SEG [Wootton & Federhen 1993]. On the other hand, many relationships of biological significance can be missed if only sequences that pass the filter are to be considered.

Setting the threshold for PSI-BLAST at the value of  $1e^{-10}$  eliminated most of the similarities that are due to low complexity regions. However, to test our validation procedures, and in order to detect those false clusters which escaped our pruning criteria, we repeated the whole computational process after the NR database

was filtered to exclude low complexity segments, using the SEG program [Wootton & Federhen 1993]. Out of the 1421 clusters, 695 (49%) are not affected by filtering, and 1211 (85%) clusters have the same size to within 10 members

All clusters whose membership increased by more than 100 when the non-filtered database was searched, were manually checked. Out of 46 clusters, 29 had already been detected using our validation criteria and corrected. An additional 5 clusters were detected but approved without correction. Another 11 clusters suffered a substantial decrease in true positives as a result of the filtering process. Only in one case (cluster 851 that contains mostly GAG Polyproteins) we observed an excess of hits in the non-filtered database due to similarity with proline rich proteins. Yet, use of the high threshold prevented a collapse, and the process stopped after 2 iterations. The proline rich regions had low significance value (greater than  $1e^{-8}$ ).

Many biologically meaningful hits are not reported with the filtered database (as observed for the 16 clusters mentioned above), indicating that masked regions contain important information which can be missed by straight-forward filtering. Since our pruning criteria revealed most of the problem cases, we chose to define our clusters using the non-filtered NR database.

## 2.4 Mapping the second part of the space - the séquence-based clusters

After all members of type-I clusters were marked, we turn to the remaining sequences. We apply the PSI-BLAST procedure repeatedly; to cover each time what is left over from the previous search. Each time a random query is selected from the set of undetected sequences, a search is performed with that query, and the set of detected séquences is updated. The resulting clusters may overlap as the search is performed against the whole NR database (including members of type-I clusters). The same validation criteria as discussed above are applied to control the PSI-BLAST process, and the clusters are refined accordingly. The resulting clusters are called type-II clusters.

## 3 Results

### 3.1 General information

Overall, the 1421 type-I clusters contain 168,431 sequences (34,259,323 amino acids) at significance level (evalue) of better than  $1e^{-2}$ . This is 44.5% of all sequences in the NR database, and 28.5% of the amino acids in our database. 155,690 of the sequences have sequence identity of at least 20% with the seed sequence. 100,261 sequences share 20% identity with the seed sequence, along their whole sequence. The largest 20

type-I clusters are given in Table 2. At the time of writing, 3883 random queries 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, and the complete statistics about type-I and type-II clusters will be available soon on the **BioSpace** website <http://biospace.stanford.edu>.

Cluster number	Size	No of amino acids	Family
1	17,596	1,657,530	Immunoglobulin (variable domain)
1296	15,882	1,823,563	Envelope Glycoprotein
891	5,514	1,982,581	Tropomyosin, Myosin, Kinesin (coiled coil segments)
11	4,845	1,251,316	Protein kinase
51	4,123	1,245,273	Ribulose biphosphate carboxylase (large chain) domain II
127	4,026	523,363	Ribulose biphosphate carboxylase (large chain) domain Ia
146	3,640	386,697	Ribulose biphosphate carboxylase (large chain) domain Ib
3	3,583	374,549	Immunoglobulin (constant domain)
13	2,726	273,938	Pol polyprotein, Protease
95	2,658	610,565	Reverse transcriptase
88	2,407	137,968	Homeobox
177	2,194	175,560	MHC class II
85	2,037	307,682	MHC class I
134	1,887	200,024	Gag protein
453	1,854	268,378	ADP-ribosylation factor, Ras-like proteins, GTP-binding proteins
542	1,780	148,894	Gag protein
42	1,545	319,017	Hydroxysteroid dehydrogenase
7	1,514	191,313	EF hand (Calmodulin, Troponin C, Parvalbumin)
31	1,486	218,200	Ras-like proteins
212	1,460	84,843	Transcription factors
116	1,369	454917	Cytochrome P450
222	1307	260874	Myosin light chain, Titin
5	1258	275543	Trypsinogen, Thrombin, Serine protease
137	1127	350390	Hemagglutinin
311	1006	296557	GTP-binding proteins
135	886	241207	Fibronectin, Tenascin
494	616	319474	Peptide synthetase

Table 2: **Top 20 clusters.** The (preliminary) family description states the feature common to most or many of the member proteins. Additional clusters, which are among the 20 largest in terms of the number of amino acids, are given below the horizontal line.

### 3.2 Modeling

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. Using our alignments we created 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.

### 3.3 Overlapping clusters

Since the ProtoMap clustering algorithm only merges clusters when the statistical evidence for a relationship

between the member proteins is strong, it may happen that some of our seed type-I clusters (that contain only SCOP sequences) are related through weak but genuine sequence similarity. Consequently it may happen that the final extended (type-I) clusters of two seed clusters which share weak sequence similarity, will include the same sequence as a member. Out of all 1,008,910 type-I cluster pairs, 3559 pairs (0.35%) have at least one protein in common. However, out of these 3559 pairs, 2384 are totally disjoint at the level of the amino acid: the same protein may be classified into 2 different clusters, but each cluster includes a different part (domain) of the protein. One such example is clusters 51 and 127 (see Table 2) which share 3958 proteins but zero amino acids. Out of the remaining 1175 pairs (0.11% of all cluster pairs), in only 246 cases are more than 50% of the amino acids included in both clusters. In no case were the clusters totally overlapping. Moreover, though a protein domain may be classified to more than one cluster, the quality of the alignment may vary. Since a major goal of our analysis was to build reliable models for as many as possible sequences in the protein space, we left the clusters untouched. All 1175 pairs of clusters are marked as "related clusters" for further analysis.

Overlapping is also observed for type-II clusters. Out of all 7,536,903 type-II cluster pairs, 17,508 pairs (0.23%) have at least one protein in common but only 11,426 share one amino acid or more. 501 cluster pairs were more than 50% overlapped, and of these, 18 cluster pairs were identical. This overlapping reflects the fact that 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 (for example, by picking the "center" sequence, as was done for type-I clusters). Databases such as Pfam and ProtoMap which provide pre-defined groups of related sequences can help to reveal these center sequences, and we plan to use of these databases in the future.

Among all possible pairs of type-I type-II clusters (5,517,743), 10,043 cluster pairs share at least one sequence. Surprisingly, 3822 of them share more than one amino acid. While most of these pairs were disjoint, for 297 pairs more than 50% of the amino acids were overlapped. In no case the clusters were totally identical, but several type-I clusters were completely or almost completely contained within type-II clusters. Again, this indicates that PSI-BLAST searches should be treated with care, and more than one run may be required to insure that all possible sequences that could have been detected with PSI-BLAST were indeed detected.

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

Sequence based metrics have limited success when it comes to detecting weak similarities between sequences that have diverged greatly. Moreover, in many cases proteins may have the same fold and close biological function, though significant sequence similarity is not observed [Murzin 1993, Pearson 1997, Brenner et al 1998]. Therefore, using structural information where available to deduce relationship between proteins can greatly increase the accuracy of predictions and functional analysis. Incorporating such information in our map can help to detect higher level regularities within the protein space.

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. However, since these measures are based on different considerations it is not clear how one should judge scores that are assigned by either metric. The same problem exists, for example, when one wants to give the distance between two cities either in land travel distance or in air travel distance. Without knowing the relation between the two metrics it is hard to develop a notion of “close” and “far” which is consistent with both metrics. Moreover, it is not clear how statistical measures that are based on different protein features can be combined and transformed to a single scale. Fortunately, the statistical estimates that we use are based on the same statistical framework as will be clarified below, and, with proper normalization by search space size, are correlated fairly well. To correct biases and shifts, which are due to different random background distributions, one can use the similarities observed that are significant both in sequence and structure to calibrate the scores.

#### 4.1 Structural comparisons

Our procedure starts by calculating structural similarities between all seed proteins of SCOP clusters. These matches were calculated using **structal** [Gerstein & Levitt 1998, Levitt & Gerstein 1998]. This program uses iterative dynamic programming to find the optimal global alignment. A similarity matrix  $S_{i,j}$  is created at each round based on the inter-atomic distances between each atom  $i$  in the first structure and each atom  $j$  in the second structure. An important advantage of this program is that it gives statistical estimates similar to the statistical estimates of BLAST, i.e., the output gives the probability that the observed structural similarity could have been obtained by chance.

Out of the 1,008,910 type-I cluster pairs, 2695 pairs share significant structural similarity with value bet-

ter than  $1e^{-3}$ . Note that this number of pairs (2695) is more than 2.5 times bigger than the number of overlapping clusters (1175). Moreover, only 407 pairs of the 1175 overlapping pairs are detected as structurally similar. An additional 6341 pairs were found with structural similarity value better than  $1e^{-2}$ ; only 155 of these are overlapping clusters. Hence, overlap does not entail structural similarity and the contrary does not hold either. I.e., most structural similarities cannot be explained as a result of overlap. The structural alignments used here will be made available on the BioSpace website (and their detailed analysis will appear elsewhere).

#### 4.2 Comparing profiles

Our ultimate goal is to map the whole protein space. 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 type-II clusters, this relationship could have been detected. However, in the absence of structural information for these clusters we must rely on sequence information. Although we already used the most powerful tools available for database searches, we have not yet used the profiles obtained for our clusters. These profiles can provide a new, potentially powerful measure of similarity between clusters. Moreover, by comparing profiles, and using the same statistical framework as for structural similarities to assess their significance, we are able to obtain a complete protein map that is consistent in terms of both structure and sequence comparison. A detailed description of our novel procedure for profile comparison is given in appendix A.

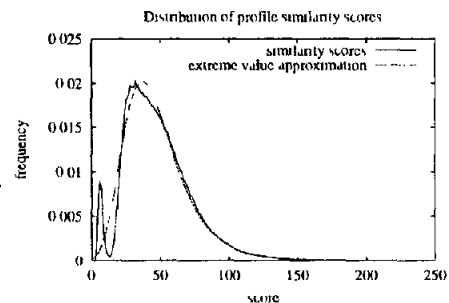


Figure 2: Distribution of profile similarity scores

In Fig. 2 we show the distribution of profile similarity scores for pairs of unrelated type-I clusters. Pairs of clusters in this subset are selected following three criteria: (1) they do not overlap, (2) they do not share significant structural similarity and (3) they belong to differ-

ent SCOP classes. Effectively, this distribution reflects the distribution of similarity scores of “random” profiles. An extreme-value distribution [Gumbel 1958] was fitted to this distribution, so as to estimate the statistical significance (e-value) for each raw similarity score. Note that the theoretical fit deviates from the empirical distribution for low similarity scores. Matches along secondary structure elements can explain this, and we are currently testing this hypothesis. This deviation does not affect our probability calculations as it is confined to the region of the plot where score are not significant.

To evaluate the sensitivity and selectivity of our new approach for profile-profile comparison we are currently testing it on profiles of type-I clusters that are associated with the same SCOP family, superfamily, fold or class.

### 4.3 The unified map - detection of super-clusters

The pairwise cluster similarity scores (based either on the structural similarities or the profile similarities) are used to cluster the clusters to superfamilies and “fold” families, using the ProtoMap clustering algorithm that is described in section 2.2.

Based on our preliminary data, the 5,304 clusters form 3,790 superclusters of which 333 contain more than one cluster. Many relations are observed between superclusters but 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, but these relationships already provide us with local maps of the protein space (as described in [Yona et al. 1999]). Application of a multidimensional scaling procedure currently underway is expected to lead to the first global map of the protein space that combines both sequence and structure.

SCOP sequences (or type-I clusters) are scattered among 584 superclusters. This number lies between the number of folds (440) and the number of superfamilies (640) in the SCOP database. Only 15 superclusters contain SCOP entries from different classes. These cases will be studied carefully, and our parameters and algorithms will be refined accordingly. Within superclusters that contain at least one type-I cluster we find 677 type-II clusters. Hence we believe that this map can extend predictions of structural and functional similarities beyond that obtainable with existing methods.

## 5 Discussion

One of the most important problems in genomics is the clustering of related proteins in the protein space. Usually proteins are grouped based on either sequence sim-

ilarity or structure similarity (when structural information is available). Here we have tried to combine both measures of similarity in order to build a map of the protein space that is consistent with both sequence similarity and structure similarity.

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 used to group these clusters into superclusters. Specifically, we made use of efficient and sensitive structural comparison algorithm, and developed a novel method for profile-profile comparison. Taken together, these tools enable us to establish a consistent and unified framework for sequence and structure analysis. By combining sequence information and structure information (when available) it is believed that a better, more accurate map can be created.

The work described here is the first stage in 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, and to build a 3D model for a substantial part of the protein space.

## 6 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  $\mathbf{P} = \mathbf{p}_1 \mathbf{p}_2 \mathbf{p}_3 \dots \mathbf{p}_n$  and  $\mathbf{Q} = \mathbf{q}_1 \mathbf{q}_2 \mathbf{q}_3 \dots \mathbf{q}_m$ , where  $n$  and  $m$  are the lengths of the profiles (the number of positions or columns) and  $\mathbf{p}_i, \mathbf{q}_j$  are probability distributions over the 20 letter alphabet of amino acids, the score of a “match” between two columns  $\mathbf{p}_i$  and  $\mathbf{q}_j$  is based on their statistical similarity.

A possible measure of statistical similarity between two (empirical) probability distributions  $\mathbf{p}_i(x)$  and  $\mathbf{q}_j(x)$ ,

is the Kullback-Leibler (KL) divergence [Kullback 1959] which is defined as

$$D^{KL}[\mathbf{p}||\mathbf{q}] = \sum_k p_{ik} \log_2 \frac{p_{ik}}{q_{ik}}$$

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  $\mathbf{p}$  and  $\mathbf{q}$ , for every  $0 \leq \lambda \leq 1$ , their  $\lambda$ -JS divergence is defined as

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

where

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

can be considered as the most likely common source of both  $\mathbf{p}$  and  $\mathbf{q}$ , with  $\lambda$  as a prior. Without a priori information, a natural choice is  $\lambda = 1/2$ . We denote the corresponding measure by  $D^{JS}$ . 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 a priori probability of the source distribution. This information can help to assess the significance of a “match”. Say that the two given empirical distributions resembles 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 resembles 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  $\mathbf{p}$  and  $\mathbf{q}$  is defined as

$$\begin{aligned} \text{Score}(\mathbf{p}, \mathbf{q}) &= \frac{1}{2}(1 - D)(1 + S) \\ &= \frac{1}{2}(1 - D^{JS}[\mathbf{p}||\mathbf{q}])(1 + D^{JS}[\mathbf{r}||\mathbf{P}_0]) \end{aligned}$$

Thus, the score for two similar distributions (i.e.  $D = D^{JS}[\mathbf{p}||\mathbf{q}] \rightarrow 0$ ) whose common source is far from the background distribution (i.e.  $S = D^{JS}[\mathbf{r}||\mathbf{P}_0] \rightarrow 1$ ), 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)  $s^* = \max\{s(a, b)\} > 0$ . 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 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.

## References

- [Altschul et al. 1994] Altschul, S. F., Boguski, M. S., Gish, W. G. & Wootton, J. C. (1994) Issues in searching molecular sequence databases. *Nature Genetics* **6**, 119-129.
- [Altschul et al. 1997] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucl. Acids Res.* **25**, 3389-3402.

- [Attwood et al 1999] Attwood, T K, Flower, D R, Lewis, A P, Mabey, J E, Morgan, S R, Scordis, P, Selley, J & Wright, W (1999) PRINTS prepares for the new millennium *Nucl Acids Res* **27**, 220-225
- [Barker et al 1996] Barker, W C, Pfeiffer, F & George, D G (1996) Superfamily classification in PIR-international protein sequence database *Methods Enzymol* **266**, 59-71
- [Bateman et al 1999] Bateman, A, Birney, E, Durbin, R, Eddy, S R, Finn, R D, & Sonnhammer, E L (1999) Pfam 3.1 1313 multiple alignments and profile HMMs match the majority of proteins *Nucl Acids Res* **27**, 260-262
- [Brenner et al 1998] Brenner, S E, Chothia, C & Hubbard, T J P (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships *Proc Natl Acad Sci USA* **95**, 6073-6078
- [Corpet et al 1999] Corpet, F, Gouzy, J, & Kahn, D (1999) Recent improvements of the ProDom database of protein domain families *Nucl Acids Res* **27**, 263-267
- [Dayhoff 1976] Dayhoff, M O (1976) The origin and evolution of protein superfamilies *Fed Proc* **35**, 2132-2138
- [Elofsson & Sonnhammer 1999] Elofsson, A & Sonnhammer, E L (1999) A comparison of sequence and structure protein domain families as a basis for structural genomics *Bioinformatics* **15**:6, 480-500
- [El-Yaniv et al 1997] El-Yaniv, R, Fine, S & Tishby, N (1997) Agnostic classification of markovian sequences *Advances in Neural Information Processing Systems* **10**, 465-471
- [Gerstein & Levitt 1998] Gerstein, M & Levitt, M (1998) Comprehensive assessment of automatic structural alignment against a manual standard, the SCOP classification of proteins *Protein Sci* **7**, 445-456
- [Gonnet et al 1992] Gonnet, G H, Cohen, M A & Benner, S A (1992) Exhaustive matching of the entire protein sequence database *Science* **256**, 1443-1445
- [Gracy & Argos 1998] Gracy, J & Argos, P (1998) 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* **14**:2, 164-187
- [Gumbel 1958] Gumbel, E J (1958) "Statistics of extremes" Columbia University Press, New York
- [Han & Baker 1996] Han, K F & Baker, D (1996) Global properties of the mapping between local amino acid sequence and local structure in proteins *Proc Natl Acad Sci USA* **93**, 5814-5818
- [Harris et al 1992] Harris, N L, Hunter, L & States, D J (1992) Mega-classification: Discovering motifs in massive datastreams In *Proc of the 10th national conf on AI*, 837-842, AAAI press/The MIT Press, Menlo park/Cambridge
- [Henikoff & Henikoff 1992] Henikoff, S & Henikoff, J G (1992) Amino acid substitution matrices from protein blocks *Proc Natl Acad Sci USA* **89**, 10915-10919
- [Henikoff & Henikoff 1993] Henikoff S & Henikoff, J G (1993) Performance evaluation of amino acid substitution matrices *Proteins* **17**, 49-61
- [Henikoff et al 1999] Henikoff, J G, Henikoff, S & Pietrovski, S (1999) New features of the Blocks Database servers *Nucl Acids Res* **27**, 226-228
- [Hofmann et al 1999] Hofmann, K, Bucher, P, Falquet, L & Barroch, A (1999) The PROSITE database, its status in 1999 *Nucl Acids Res* **27**, 215-219
- [Holm & Sander 1997a] Holm, L & Sander, C (1997) Dali/FSSP classification of three-dimensional protein folds *Nucl Acids Res* **25**, 231-234
- [Hubbard et al 1999] Hubbard, T J, Ailey, B, Brenner, S E, Murzin, A G & Chothia, C (1999) SCOP: a Structural Classification of Proteins database *Nucl Acids Res* **27**, 254-256
- [Hughey & Krogh 1998] Hughey, R & Krogh, A (1998) "SAM: Sequence alignment and modeling software system" Technical Report UCSC-CRL-96-22, University of California, Santa Cruz, CA, July 1998
- [Karlin & Altschul 1990] Karlin, S & Altschul, S F (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes *Proc Natl Acad Sci USA* **87**, 2264-2268
- [Koonin et al 1996] Koonin, E. V., Tatusov, R L & Rudd, K E (1996). Protein sequence comparison at genome scale *Methods Enzymol* **266**, 295-321.
- [Krause & Vingron 1998] Krause, A. & Vingron, M (1998) A set-theoretic approach to database searching and clustering *Bioinformatics* **14**:5, 430-438.
- [Kullback 1959] Kullback, S (1959). "Information theory and statistics" John Wiley and Sons, New York
- [Levitt 1992] Levitt, M (1992) Accurate modelling of protein conformation by automatic segment matching *J Mol Biol* **226**, 507-533
- [Levitt & Chothia 1976] Levitt, M & Chothia, C (1976) Structural Patterns in Globular Proteins. *Nature* **261**, 552-558
- [Levitt & Gerstein 1998] Levitt, M & Gerstein, M (1998) A Unified Statistical Framework for Sequence Comparison and Structure Comparison *Proc. Natl. Acad. Sci. USA* **95**, 5913-5920
- [Levitt et al 1995] Levitt, M., Hirshberg, M, Sharon, R & Daggett, V (1995) Potential energy function and parameters for simulations of the molecular dynamics of proteins and nucleic acids in solution *Comp Phys. Comm.* **91**, 215-231
- [Lin 1991] Lin, J (1991) Divergence measures based on the Shannon entropy *IEEE Trans Info. Theory* **37**:1, 145-151
- [Murzin 1993] Murzin, A G (1993) OB(oligonucleotide/oligosaccharide binding)-fold common structural and functional solution for non-homologous sequences *EMBO J.* **12**:3, 861-867
- [Orengo et al 1997] Orengo, C. A., Michie, A D, Jones, S, Jones, D T, Swindells, M. B & Thornton, J M (1997) CATH-a hierarchical classification of protein domain structures *Structure* **5**, 1093-1108
- [Park et al 1998] Park, J, Karplus, K, Barrett, C, Hughey, R, Haussler, D, Hubbard, T. & Chothia, C (1998) Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. *J Mol Biol* **284**, 1201-1210
- [Pearson 1995] Pearson, W. R. (1995). Comparison of methods for searching protein sequence databases *Protein Sci* **4**, 1145-1160
- [Pearson 1997] Pearson, W R (1997) Identifying distantly related protein sequences *Comp App Biosci* **13**:4, 325-332
- [Rigoutsos et al 1999] Rigoutsos, I, Gao, Y., Floratos, A & Parida, L (1999) Building dictionaries of 1D and 3D motifs by mining the unaligned 1D sequences of 17 archaeal and bacterial genomes *In the proceedings of ISMB 99*, 223-233
- [Tatusov et al 1997] Tatusov, R. L., Eugene, V K & David, J L (1997) A genomic perspective on protein families *Science* **278**, 631-637
- [Watanabe & Otsuka 1995] Watanabe, H. & Otsuka, J (1995) A comprehensive representation of extensive similarity linkage between large numbers of proteins *Comp App Biosci* **11**:2, 159-166
- [Wootton & Federhen 1993] Wootton, J C & Federhen, S (1993) Statistics of local complexity in amino acid sequences and sequence databases. *Comp Chem.* **17**, 149-163
- [Yona et al 1999] Yona, G., Lital, N & Lital, 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 1999] Yona, G (1999). "Methods for global organization of the protein sequence space" *Ph D Thesis*, The Hebrew University, Jerusalem, Israel