

Probabilities for having a new fold on the basis of a map of all protein sequences

Elon Portugaly
Institute of Computer Science,
The Hebrew University
Jerusalem 91904, Israel
972-2-6585188

elonp@cs.huji.ac.il

Michal Linial
Department of Biological Chemistry,
Institute of Life Sciences,
The Hebrew University
Jerusalem 91904, Israel
972-2-6585425

michall@leonardo.ls.huji.ac.il

ABSTRACT

It is a major problem in the study of protein structure to predict which proteins have new, currently unknown structural folds. In an attempt to address this problem we studied the location of all proteins with solved structures within the map of all known protein sequences provided by ProtoMap. The mutual distances in this map among solved structures are used to derive a probabilistic model from which we infer an estimate for the probability of an unsolved protein to have a new fold. The probabilities were based on data from SCOP release 1.37. The results were evaluated against the more recent SCOP pre-release 1.41. Our predicted probabilities for unsolved proteins to have a new fold are very well correlated with the proportion of new folds among recently released structures. Thus, information about the structure of proteins can be inferred from a global relational view of protein sequences. Finally, the same procedure was applied to estimate probabilities on the basis of SCOP 1.41. A list of the highest scoring proteins is provided: These are about 80 non-membranous proteins that belong to clusters with more than 5 proteins and achieve the highest probability to have a new fold. A rational selection for 3D determination of those targets is expected to accelerate the pace of new fold discovery.

Keywords

global protein organization, structural genomics, clustering, statistical mode, structure prediction.

1. Introduction

Solving the structure of all representative proteins is a major goal of present day biology. The number of known protein sequences already exceeds 300,000, and is rapidly growing. In the foreseeable future it will be impossible to experimentally solve so many structures. Therefore, it is necessary to find ways to predict some key structural properties of a protein based on its

sequence and on data derived from structurally solved proteins. Current attempts to computationally determine a protein's structure based on sequence alone still have a limited success, partly due to the shortage in solved structures that can be used as models. To quote from the structure-based functional genomics meeting in Avelon, New-Jersey 1998: "Structure determination of 10,000 properly chosen proteins should result in useful three-dimensional models for hundreds of thousands of other protein sequences. In other words, structural genomics will put each protein within comparative modeling distance of a known protein structure" [16]. In order to realize this vision, the structural community must first select those properly chosen target proteins. Clearly these target proteins should include representatives of all possible structural folds (discussed in [7], [8], [9], [11] and [17]).

Assignment of structures to proteomes is addressed by computational as well as by experimental approaches. Those proteins, for which no related structure is known, can form a basis for a list of targets likely to have unknown folds [16]. The total number of protein folds in the entire protein universe is unknown [3], [5]. This number was estimated to be around 1,000 [1], but, estimates ranging from 700 to over 10,000 were made [2], [14], [18] and [4]. The total number of currently known protein folds 453 according to the SCOP 1.41 classification [6] and 635 folds (topologies) according to CATH 1.5 [15].

In this study we address the problem of compiling a list of target proteins by estimating the probability for each protein to have a new fold. We use the structures in SCOP 1.37 to estimate the probability of each protein to have a new fold. We verify our estimation procedure by verifying our estimated probabilities against newer structures from SCOP 1.41. Having verified our procedure, we apply it to the structures in SCOP 1.41 to generate up-to-date estimations.

Our target list consists of those proteins whose estimated probability to have a new fold is highest.

2. Prediction Procedure

We employ two classifications of proteins: ProtoMap [19] and SCOP [13]. ProtoMap is an automatically-generated hierarchical and relational classification of all protein sequences in Swissprot. We use the most relaxed level of classification (level e-0) of ProtoMap version 2.0. This version includes 72,623 protein sequences that are classified to 13,354 clusters. ProtoMap can be accessed at <http://www.protomap.cs.huji.ac.il>.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, to republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

RECOMB 2000 Tokyo Japan USA
Copyright ACM 2000 1-58113-186-0/00/04 \$5.00

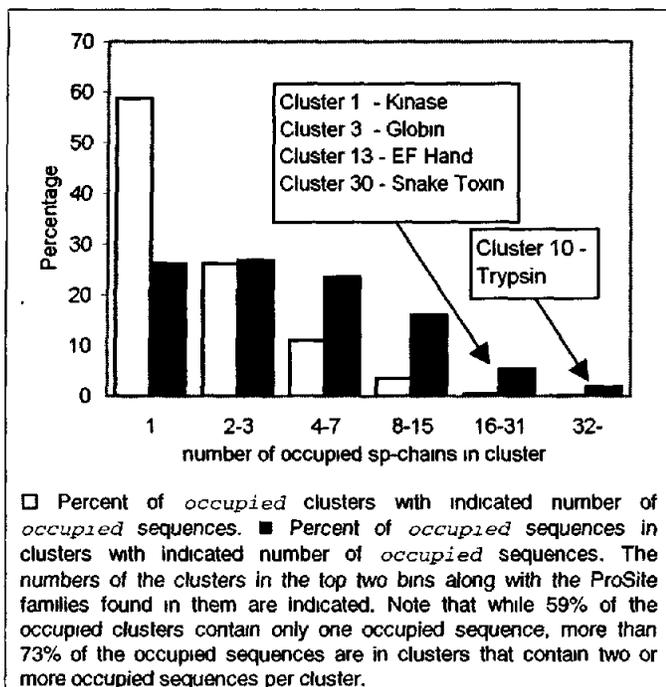


Fig. 1 - Distribution of the number of occupied sequences in each occupied cluster

Each cluster of sequences in ProtoMap has a weighted list of related clusters. Weights (called *quality*) range from 0.0 to 1.0 and reflect the statistical significance of the relatedness. The lists of related clusters encode many biological meaningful relations and form the basis for mapping the protein space [19]. One can define a graph whose vertices are all the clusters in ProtoMap level e-0. Edges connect each cluster to the clusters related to it. This graph can be “clipped” at different thresholds, by eliminating all edges below a given threshold. Consequently, each threshold yields connected components of different size and connectivity.

SCOP is a hierarchical classification of all known protein structural domains [13]. We have studied SCOP release 1.37 (5,741 natural protein entries that were registered at the PDB database prior to 1997, 20 Oct). This release comprised 11,748 records of 2,264 domains. The transformation from the number of PDB entries to the number of SCOP records and SCOP domains reflects (i) parsing of the proteins to their structural domains (ii) grouping of entries in SCOP records that reflects the redundancy within PDB. These 2,264 domains are classified to 834 families, 593 super-families, 427 folds and 8 classes. Two more classes – “designed proteins” and “non-protein” are not considered in this study. SCOP can be accessed at <http://scop.mrc-lmb.cam.ac.uk/scop>.

Our hypothesis is that distances on the ProtoMap graph are consistent with distances between protein features, including their structure. This hypothesis is based on numerous biological test cases that were manually evaluated [10] and [19]. We describe below an exploration of the distances in the ProtoMap graph between known structures, and extract from this exploration a statistical estimation of a protein’s probability to have a new fold.

2.1 Mapping SCOP Domains to Swissprot Protein Chains

We start by mapping each domain in SCOP to the sequence(s) in Swissprot that form that domain.

Of the 2,264 domains, 1,986 mapped successfully. Most (170 domains) of the unmatched domains are variable regions in immunoglobulins that have no corresponding sequence. The rest (108 domains) are well distributed and are the result of inconsistencies between records in SCOP, PDB, and Swissprot. Our discussion ignores all these unmatched domains.

Mapping of domains to sequences can be viewed in both directions. We say that a sequence is *occupied* if it maps to at least one domain, and is *vacant* otherwise. Of the 72,623 sequences, 1,688 are occupied.

We extend the mapping of sequences to folds, by saying that a sequence maps to a fold if it maps to a domain that is classified to that fold. Of the 427 folds in SCOP, 411 mapped to sequences.

We extend the mapping of folds to clusters, by saying that a fold maps to a cluster if it maps to a sequence that is classified to that cluster. We say that a cluster is *occupied* if it maps to any fold, and *vacant* otherwise.

Of the 13,354 clusters in ProtoMap (at the most relaxed level of classification, e-0), 756 are occupied. The distribution of the number of occupied sequences in each occupied cluster is shown in Fig. 1.

2.2 Assigning Representative Folds for ProtoMap Clusters

Since many proteins are multi-domain, a sequence may map to several domains, which usually have distinct folds. Of the 1,688 occupied sequences, 352 (21%) map to more than one domain. This creates a problem when one tries to investigate structure, which is domain oriented, using ProtoMap, which is whole-protein oriented. The problem is illustrated in Fig. 2. Panel B illustrates a problematic ProtoMap cluster. Sequences no. 1 and 5 are in that cluster due to false transitivity - they have different

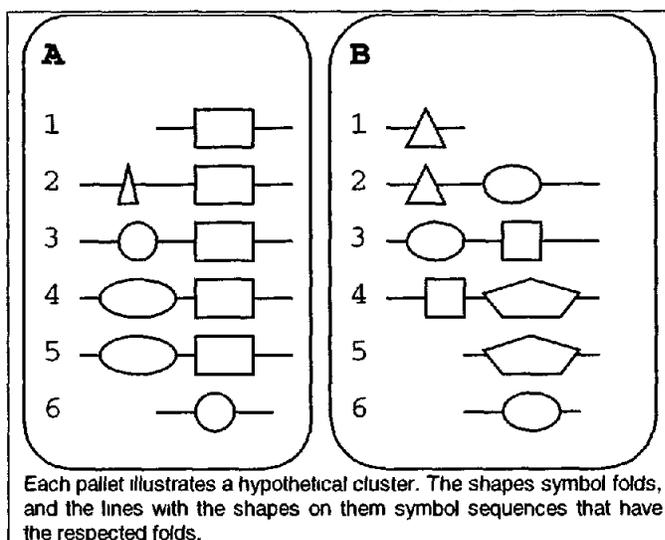


Fig. 2 - Multi-domain problem illustration 1

fold, and furthermore, it is possible that there is no fold x that appears on any protein with the triangular fold and on another protein with the pentagonal fold. Panel A illustrates a more regular cluster. Almost all the proteins in the cluster share a common fold, but some of them have other folds in addition to that fold. In this study, we chose to address with the multi-domain problem by assigning a single representative fold for each occupied cluster.

For each occupied cluster, we chose as a representative, the fold that maps to the maximum number of sequences in that cluster (if more than one fold was maximal in that respect, one was chosen arbitrarily). Thus, if we now look at Fig. 2 as if the illustrated sequences are the occupied sequences in the hypothetical clusters, then the representative folds for the clusters in panel A and B would be the square fold and the oval fold, respectively.

For panel A, if the occupied sequence faithfully represent all the sequences in the cluster, the assignment of the square fold as a representative, while simplifying the data, is justified. For the hypothetical cluster in panel B, the oval fold is clearly not a satisfying representative, however, no better choice is available.

The following statistics describe the distortion generated by the above process of assigning representative folds to cluster:

In 82% of the occupied clusters, the second best candidate fold mapped to less than 0.75 times the number of sequences as the fold that was chosen as representative.

Off the 411 folds that mapped to sequences, 329 folds were chosen as clusters' representatives. Of the remaining 82 folds that were never chosen as representatives, 44 are coupled to a representative fold (i.e., every sequence to which they map, also maps to that fold). These folds add little information to the information supplied by their coupled fold. Therefore, about 10% of the folds are not represented in the group of representative folds.

Only 80 sequences do not map to the representative fold of their cluster. These 80 sequences constitute less than 6.5% of the occupied sequences in clusters that include more than one occupied sequence. Therefore, ProtoMap is selective for SCOP folds, and the representative fold represents at least some part of almost all the sequences in the cluster.

2.3 Predicting a Proteins Probability to Have a New Fold

Having observed that ProtoMap is selective for folds, we can speak of the "Fold of a cluster" and about the probability that a cluster's fold is a new, unknown fold. We say that a cluster is *new* when its (presently undetermined) corresponding fold is absent from SCOP, and *old* otherwise. We wish to predict for each cluster the probability that it is new. This is done as follows: Fix a threshold t on the quality of relatedness between clusters. This defines a graph G_t whose vertices are clusters and where edges connect pairs of clusters whose relatedness is of quality t or more. Let $d_t(a,b)$ define the distance between clusters a and b in G_t (i.e. the minimal number of edges on any path from a to b in G_t). The i -neighborhood of cluster c is the group of clusters whose distance from c in G_t is no more than i :

$$N_i(c;i) \equiv \{a \mid d_t(a,c) \leq i\}$$

We say that an i -neighborhood of c is *vacant* if all the clusters in it are vacant and is *occupied* otherwise. The *vacant-neighborhood* of cluster c is the maximal i -neighborhood of c that is vacant. The *vacant-neighborhood-width* is the i of the vacant-neighborhood:

$$VNW_i(c) \equiv \max_i \{i \mid N_i(c;i) \in \text{vacant}\}$$

$$VN_i(c) \equiv N_i(c;VNW_i(c))$$

G_t is not a connected graph, and there are cases where all the clusters in a connected component are vacant. For a cluster in such a connected component, $VNW_i(c) = \infty$. We define the *isolation size* of cluster c as the size of its vacant-neighborhood, and the *isolation type* of cluster c as *relative* if it is in a connected component that is not all vacant, and as *full* otherwise. These two values together are the *isolation value* of c :

$$IS_i(c) \equiv |VN_i(c)|$$

$$IT_i(c) \equiv \begin{cases} \text{relative} & \text{if } VNW_i(c) < \infty \\ \text{full} & \text{if } VNW_i(c) = \infty \end{cases}$$

$$I_i(c) \equiv (IS_i(c), IT_i(c))$$

We define two base distributions other the uniform distribution of all clusters: Let c be a random variable uniformly distributed over the group of all clusters. Than define:

$$D_i^{old}[I] \equiv \Pr[I = I(c) \mid c \in \text{old}]$$

$$D_i^{new}[I] \equiv \Pr[I = I(c) \mid c \in \text{new}]$$

$D_i^{old}[I]$ is the distribution of isolation values of old clusters, and $D_i^{new}[I]$ is the distribution of isolation values of new clusters. Our training set for estimating the base distributions consisted of the entire set of occupied clusters (756 clusters).

To estimate $D_i^{old}[I]$, we calculated for each occupied cluster the isolation it would have had, had it been vacant. To estimate $D_i^{new}[I]$, we calculated for each occupied cluster with a representative fold f , the isolation it would have had, had all the occupied clusters represented by the same fold f been vacant. We then gathered the data to bins of isolation values. The data immediately suggest partitioning the range of isolation values into three bins: (z , *full*), (1 , *relative*), and the rest of the values, where z stands for any value. We define the bins of isolation values as follows:

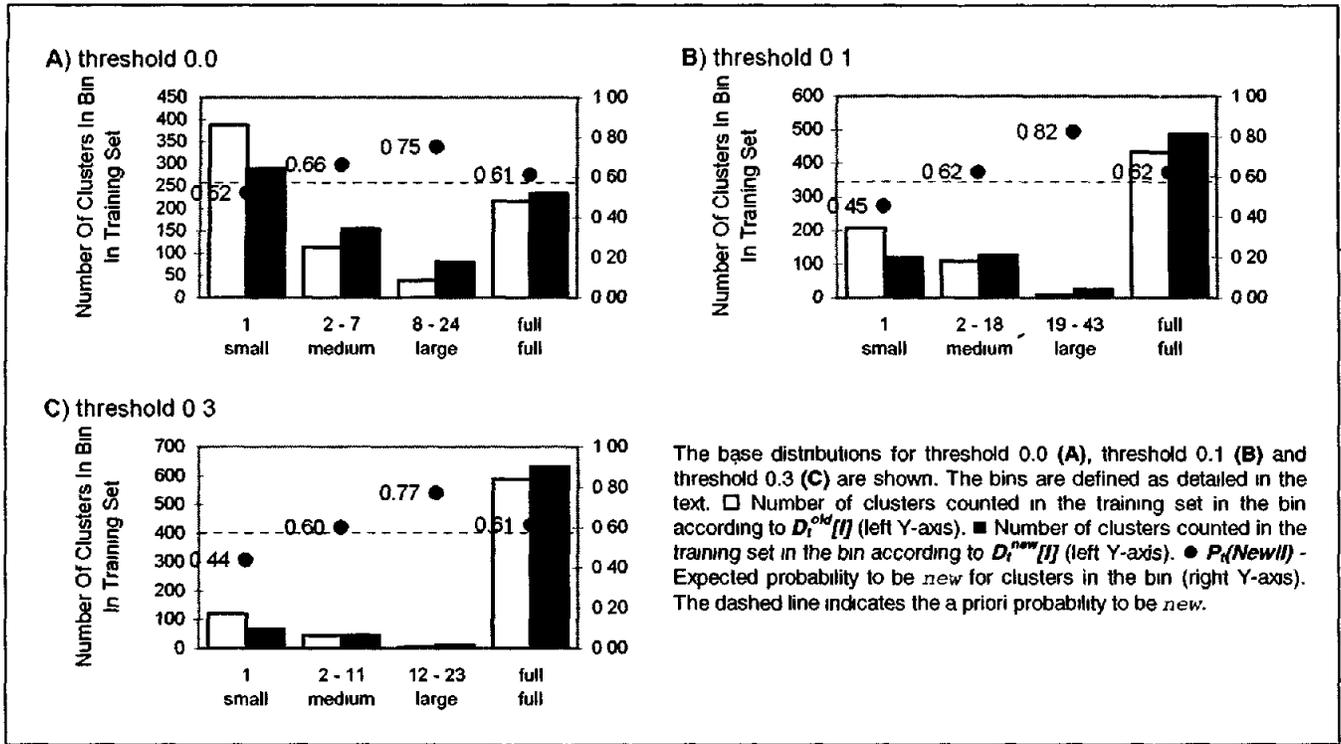


Fig. 3 - Base distributions

$$\hat{I}_i(c) \equiv \begin{cases} s & \text{if } [IT_i(c) = \text{relative}] \wedge [IS_i(c) = 1] \\ m & \text{if } [IT_i(c) = \text{relative}] \wedge [2 \leq IS_i(c) < x] \\ l & \text{if } [IT_i(c) = \text{relative}] \wedge [x \leq IS_i(c)] \\ f & \text{if } IT_i(c) = \text{full} \end{cases}$$

where *s* stands for *small*, *m* for *medium*, *l* for *large* and *f* for *full*.

The parameter *x* was determined to maximize the difference between $D_i^{old}[\hat{I}]$ and $D_i^{new}[\hat{I}]$, by maximizing the Kullback-Leiber divergence between the two distributions ($D_{kl}[D_i^{old} | D_i^{new}]$).

We performed the estimation process for three thresholds: 0.0, 0.1 and 0.3. The values determined for the parameter *x* were 8, 19 and 12 respectively. Fig. 3 describes the three pairs of base distributions that derived from the estimation process.

In order to transform the distributions to probabilities we have to include the prior probability of a cluster to be new. This prior probability for a new fold is based on the number of currently discovered folds (427, according to SCOP 1.37) and on the estimation of total number of folds, for which a rather conservative estimate of 1,000 is taken [1]. Based on this assumption, we calculated the overall probability of any fold to be new:

$$Pr[new] = 1 - (\text{total number of known folds}) / (\text{total number of folds}) = 1 - 427 / 1000 = 0.573.$$

We defined the distribution of Isolation values of all vacant clusters:

$$D_i[\hat{I}] \equiv Pr[\hat{I} = \hat{I}_i(c) | c \in \text{vacant}] = Pr[new] * D_i^{new}[\hat{I}] + (1 - Pr[new]) * D_i^{old}[\hat{I}]$$

The probability that a cluster is new, given its isolation value is:

$$Pr_i[new | \hat{I}] = \frac{Pr_i[\hat{I} | new] * Pr[new]}{D_i[\hat{I}]} = \frac{D_i^{new}[\hat{I}] * Pr[new]}{D_i[\hat{I}]}$$

For each cluster *c*, the probability that it is new is given by:

$$Pr_i[c \in new] = Pr_i[new | \hat{I}_i(c)]$$

Fig. 3 shows the values calculated for the probability that cluster be new for each bin of isolation values, for all three threshold we used ($Pr_{0.0}[new|\hat{I}]$, $Pr_{0.1}[new|\hat{I}]$, $Pr_{0.3}[new|\hat{I}]$). There seems to be a paradox in the probabilities assigned to the *full* isolation bin. One would expect that the probability of a cluster in the *full* isolation bin to be new be higher than the probability of a cluster in the *large* isolation bin. This paradox is discussed later.

We now wanted to choose one final probability function. Since $D_{kl}[D_{0.3}^{old} | D_{0.3}^{new}]$ was significantly lower than $D_{kl}[D_{0.1}^{old} | D_{0.1}^{new}]$ and $D_{kl}[D_{0.0}^{old} | D_{0.0}^{new}]$, we continued analysis only for thresholds 0.0 and 0.1.

We next used the membranous proteins as a test case. So far, there are only very few membranous protein domains whose structure is solved (mostly classified as SCOP class 6). Therefore

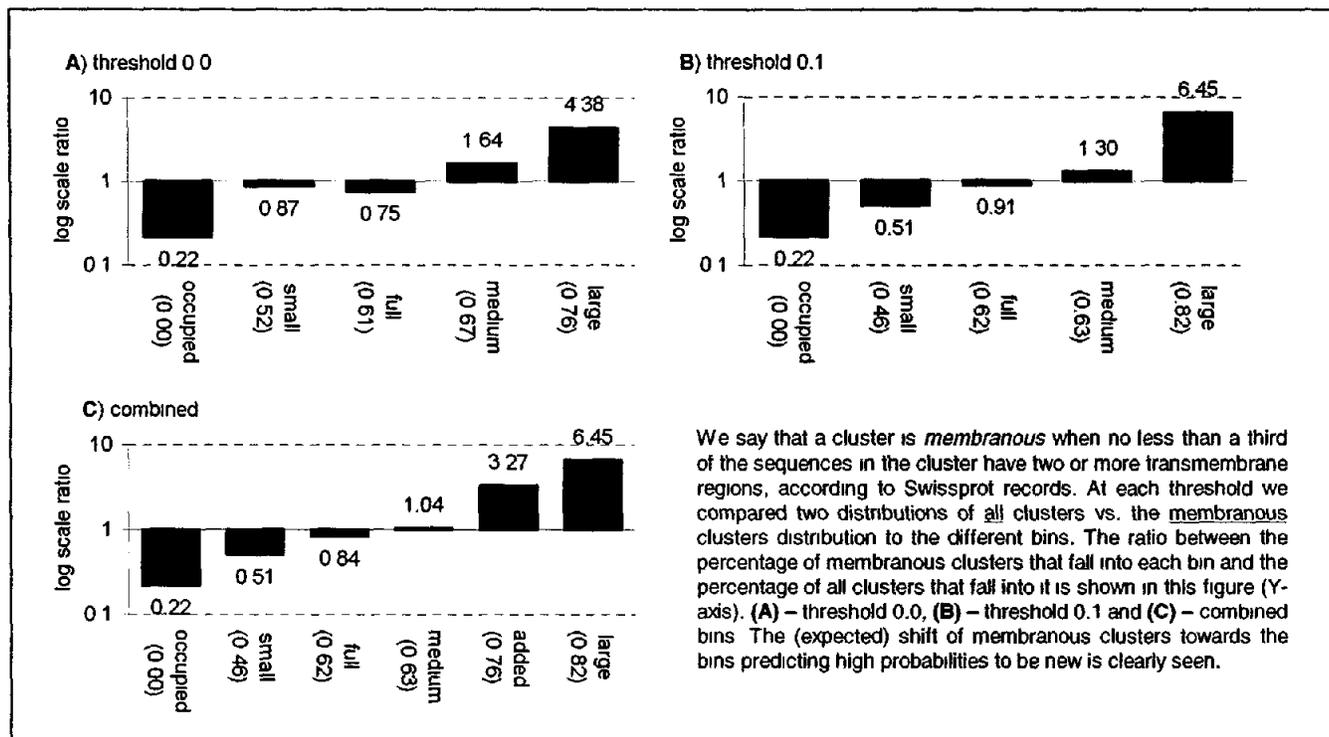


Fig. 4 - Membranous clusters test

we expect clusters of membranous proteins to have higher probabilities to be new than the rest of the clusters. For this test only proteins with multiple membrane spanning regions were considered. The analysis described in Fig. 4A-B shows that the probability functions we calculated indeed assign higher probabilities to membranous clusters. The probability function calculated using threshold 0.1 shows best this feature. For threshold 0.1, the ratio ranges from 0.22 in the lowest probability bin to 6.45 in the highest bin; a factor of 29.5. For threshold 0.0, this factor is 20. The higher factor that was found for threshold 0.1 is in accord with the higher probability assigned to its highest probability bin (0.82 vs. 0.76).

In view of the membranous proteins test results, we decided to combine $Pr_{0.0}$ and $Pr_{0.1}$. Our final isolation value bins and probability function $Pr_{combined}$ were defined as follows:

$$\hat{I}_{combined}(c) \equiv \begin{cases} \hat{I}_{0.1}(c) & \text{if } \hat{I}_{0.0}(c) \neq 1 \vee \hat{I}_{0.1}(c) = 1 \\ a & \text{if } \hat{I}_{0.0}(c) = 1 \wedge \hat{I}_{0.1}(c) \neq 1 \end{cases}$$

$$Pr_{combined}[c \in new] \equiv \begin{cases} Pr_{0.1}[c \in new] & \text{if } \hat{I}_{combined}(c) \neq a \\ Pr_{0.0}[c \in new] & \text{if } \hat{I}_{combined}(c) = a \end{cases}$$

where a stands for *added*.

Table 1 lists the probability to be new $Pr_{combined}$ assigned to each cluster according to its isolation value bin.

Bin:	<i>occupied</i>	<i>small</i>	<i>full</i>	<i>medium</i>	<i>Added</i>	<i>large</i>
Prob.:	0.00	0.45	0.62	0.62	0.75	0.82

Table 1 - $Pr_{combined}$ based on structures in SCOP 1.37

3. Evaluation of Prediction

At this point we had fixed $Pr_{combined}$ as our prediction and had not change it any further.

Our prediction and the use of a statistical-computational approach are based on all records in SCOP release 1.37. We tested our prediction on new structures that are included in the currently pre-released SCOP 1.41.

We repeated the process of mapping domains to sequences using the records of SCOP 1.41. Following the procedure 2,428 domains that constitute 438 folds were successfully mapped. Out of these 532 domains and 60 folds are new. We keep the definitions of *occupied* and *vacant* sequences to indicate the status of sequences regarding SCOP 1.37. We say that a sequence is *1.41 occupied* if it is *vacant* and maps to a domain in SCOP 1.41. We say that a cluster is *1.41 occupied* if it contains at least one sequence that is *1.41 occupied*. Our test set was the group of *1.41 occupied* clusters. We assigned each of these clusters a fold based on both the *occupied* and the *1.41 occupied* sequences in it (we applied the same process we used for assigning representative folds from SCOP 1.37). Due to ProtoMap's selectivity for folds, a *1.41 occupied* cluster is most likely *new* if its (1.41) representative fold is new.

Fig. 5 describes the proportion of *new* clusters among the *1.41* clusters in each $Pr_{combined}$ bin. Our expectation from the

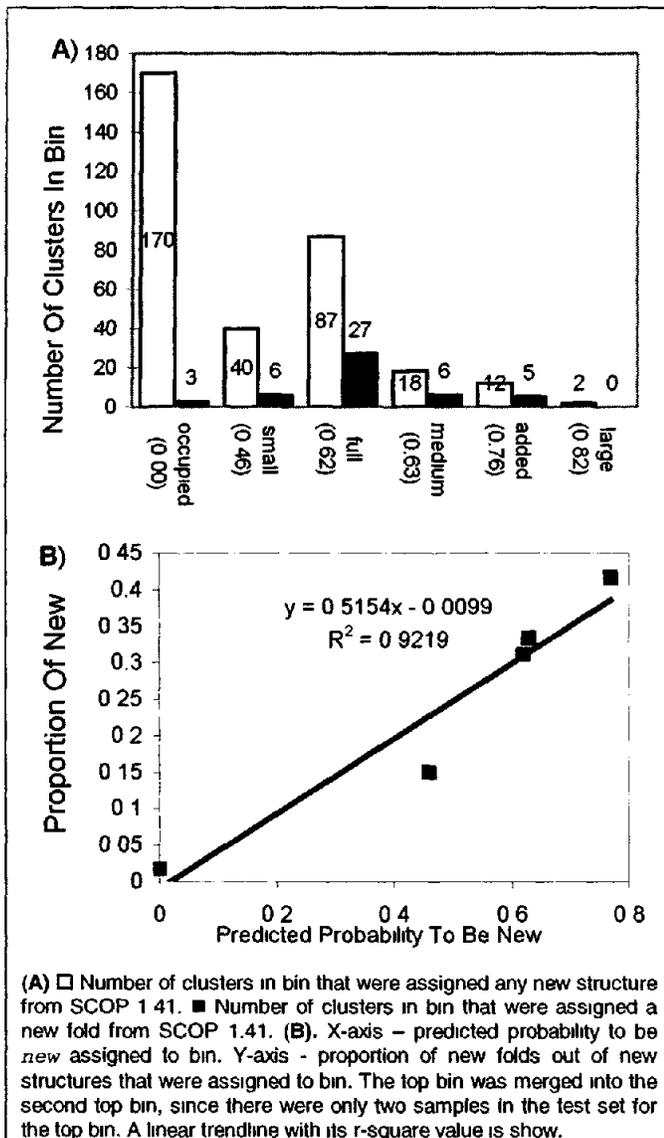


Fig. 5 - Proportion of new folds in each bin

probability function is that it should keep order with the real distribution describing the probability of clusters to be new. That is if for bins A and B, $Pr_{combined}[A] > Pr_{combined}[B]$, we would like that the proportion of new clusters in bin A be higher than the proportion of new clusters in bin B. The results in Fig. 5 fully support that expectation.

4. Splitting the *full* Isolation Bin

The estimated probability to be new for the *full* isolation bin seems like a paradox: Let *b* be a cluster with *isolation type full*. Let *a* be a cluster with *isolation type relative*. It would have been more reasonable had the probability to be new of cluster *b* been higher than the probability to be new of cluster *a*, regardless of the isolation size of either cluster.

There probability are transitive relations between clusters that ProtoMap misses, some because the intermediate proteins between the clusters are unknown. Because of this, we cannot

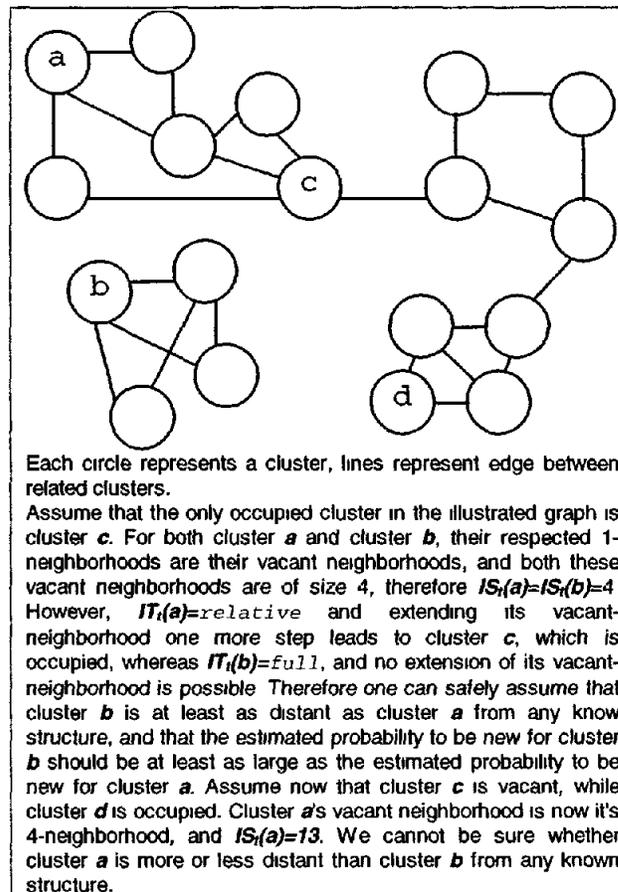


Fig. 6 - Multi-domain problem illustration 2

say that if a cluster's *isolation type* is *full*, its probability to be new is higher than that of any cluster whose *isolation type* is *relative*. Looking back at clusters *a* and *b* from our example, we can expect only that if the *isolation sizes* of cluster *a* and cluster *b* are the same, the probability to be new for cluster *b* be higher than the probability to be new for cluster *a* (Illustrated in Fig. 6).

We expect our probability to be new function to reflect this reasoning, however $Pr_{combined}$ assigns a lower probability to be new for clusters with *isolation type full* than to any cluster except for those that are related to occupied clusters ($I_1(c)=(1,relative)$). This problem is solved if the *full* isolation bin is split to two bins: *small full*, for clusters in vacant connected components of size 1 (parallel to bin *small*), and *medium full*, for clusters in vacant connected components of size 2-18 (parallel to bin *medium*). The size of the largest vacant connect component in G_{D_i} is 10, therefore there is no need for a *large full* bin. The isolation value bins should therefore be:

$$\hat{I}_i(c) \equiv \begin{cases} s & \text{if } [IT_i(c)=\text{relative}] \wedge [IS_i(c)=1] \\ m & \text{if } [IT_i(c)=\text{relative}] \wedge [2 \leq IS_i(c) < x] \\ 1 & \text{if } [IT_i(c)=\text{relative}] \wedge [x \leq IS_i(c)] \\ sf & \text{if } [IT_i(c)=\text{full}] \wedge [IS_i(c)=1] \\ mf & \text{if } [IT_i(c)=\text{full}] \wedge [2 \leq IS_i(c)] \end{cases}$$

where *sf* stands for *small full* and *mf* for *medium full*.

Estimating $Pr_{combined}$ using these bins results in the probability function described in Table 2. One can see that both the estimated probability and the observed proportion of new clusters for bin *small full* are higher than for bin *small*, and for bin *medium full* are higher than for bin *medium*.

Isolation	Estimated Probability	Proportion of New in SCOP 1.41
<i>occupied</i>	0.00	0.02
<i>small</i>	0.45	0.15
<i>small full</i>	0.59	0.27
<i>medium</i>	0.62	0.33
<i>medium full</i>	0.66	0.40
<i>added</i>	0.75	0.42
<i>large</i>	0.82	0.00

Table 2 - $Pr_{combined}$ with split *full* bin, based on structures from SCOP 1.37

This analysis was performed after the evaluation against SCOP 1.41 and is therefore presented here as a post modification to the estimation process.

5. Up-to-date Estimations Based on SCOP 1.41, and Proposed List of Selected Targets

Having evaluated the estimations based on SCOP 1.37, we concluded that our procedure for generating these estimations is reliable. We then applied the same procedure to the structures from SCOP 1.41. The only modification made to the procedure is the splitting of the *full* isolation bin, as discussed in the previous section. The new $Pr_{combined}$ probability is shown in Table 3.

Isolation	Estimated Probability
<i>occupied</i>	0.00
<i>small</i>	0.44
<i>small full</i>	0.56
<i>medium</i>	0.63
<i>medium full</i>	0.64
<i>added</i>	0.75
<i>large</i>	0.88

Table 3 - $Pr_{combined}$ based on structures from SCOP 1.41

The boundaries of the bins have not changed. The changes in the probabilities estimated for each bin are mostly due to the

changed in the prior. However, the probabilities assigned to clusters in the neighborhood of newly discovered structures were changed.

We regard the clusters assigned to the top two bins of the new $Pr_{combined}$ as possible targets for structural determination. 599 clusters (4.5% of all clusters) that account for 5.8% of the sequences are in this list. Following subtraction of membranous clusters and considering clusters that have more than 5 sequences in each, the number of clusters in this target list is reduced to 81. Among these, 67 clusters are still to be solved as no significant homologues are found in the current PDB (dated up to 6th November 1999). SCOP classification of some of these recently solved structures is still pending. These clusters compose our proposed list of targets. The list of target proteins associated with their probability to be new and additional data regarding this study are available at <http://www.cs.huji.ac.il/~elomp/Targets>.

6. Discussion

The discovery of a novel fold may contribute to understanding functional details of entire protein families, thus, a scheme for discovering those currently missing folds is desirable [5] and [12]. Here we present a systematic statistical-computational approach according to which the pace of discovering new folds may be accelerated. High probability for having new folds is assigned for 5.8% of the sequences. While our target proteins are not restricted to any specific organism, they may be incorporated to protein targets associated with numerous structural genomics projects. The occurrence of our target proteins in various organisms is thus provided.

Our statistical analysis, which produced results at the structural level of folds, has shown that the related cluster lists of ProtoMap capture information about relationship between folds. ProtoMap is based only on sequence information. Thus we have proved that through a global approach considering relationships between all known sequences, one can draw information about structure at the level of protein folds from sequence alone. Our analysis includes several heuristic decisions: The choice of ProtoMap's most relaxed level (1e-0); The assignment of a single representative fold to each occupied cluster; The choice of isolation values as we defined them as our basic parameter describing each cluster; the thresholds 0.0, 0.1, 0.3 on relatedness of clusters. Such decisions are required in essentially any automated learning process. A poor choice of features results in the failure of the learning process, and features that lead to successful learning, as indicated by tests, are legitimate.

The multi-domain problem is inherent to any analysis that is whole-protein oriented. When we consider a certain cluster, we choose one fold to represent it. If this cluster consists of multi-domain proteins, then this fold represents only one of its domains. The multi-domain problem can also lead to false transitivity of relatedness of clusters: Let cluster *a* have domain 1, cluster *b* have domains 1 and 2, and cluster *c* have domain 3. Cluster *a* might be connected via cluster *b* to cluster *c* even though they do not share any true relatedness. In this study, these problems add noise to our results, but do not bias them one way or the other since they occur when we estimate both $D_i^{new}[I]$ and $D_i^{old}[I]$.

The statistical model that was considered in determining the probability of a protein to have new fold is very naive. The information in ProtoMap's related clusters lists, if mined by more sophisticated information mining techniques (i.e. improved algorithms and better choice of data features) will probably provide predictions of the fold identity of yet unsolved proteins. This identity could either be a known fold or an unknown fold. In the latter case, the structure of the fold would not be determined, but a list of proteins with a high probability to have this unknown fold would be supplied.

The results of the prediction success evaluation (Fig. 5) show a strong correlation between predicted probability to be new and proportion of new clusters among tested clusters in each bin. Our predicted probabilities depend on the estimated total number of folds. Using any value for the estimated total number of folds (700 to 10,000) does not change the order of the predicted probabilities assigned to the bins, therefore the prediction success results validate our analysis regardless of the total number of folds.

7. Acknowledgements

We thank Nati Linal for his mathematical advice and suggestions and Hanah Margalit for critical reading. We thank Golan Yona for his advice and support throughout this study and the SCOP team for their help with their files. This study was partially supported by the Israeli Academy of Science (Initiatives in Res. in Sc. & Technology) and the Horowitz Fund.

8. References

- [1] Chothia, C. (1992). One thousand protein families for the molecular biologist. *Nature* 357, 543-544.
- [2] Finkelstein, A. V., and Ptitsyn, O. B. (1987). Why do globular proteins fit the limited set of folding patterns? *Prog. Biophys. Mol. Biol.* 50, 171-190.
- [3] Gerstein, M., and Hegyi, H. (1998). Comparing genomes in terms of protein structure: surveys of a finite parts list. *FEMS Microbiol Rev* 22, 277-304.
- [4] Govindarajan, S., Recabarren, R. & Goldstein, R. A. (1999) *Proteins* 35, 408-414.
- [5] Holm, L., and Sander, C. (1997). New structure-novel fold? *Structure* 5, 165-171.
- [6] Hubbard, T. J., Ailey, B., Brenner, S. E., Murzin, A. G., and Chothia, C. (1999). SCOP: a Structural Classification of Proteins database. *Nucleic Acids Res* 27, 254-256.
- [7] Kim, S. H. (1998). Shining a light on structural genomics. *Nat. Struct. Biol.* 5 Suppl, 643-645.
- [8] Koehl, P., and Levitt, M. (1999). A brighter future for protein structure prediction. *Nat Struct Biol* 6, 108-111.
- [9] Koonin, E. V., Tatusov, R. L., and Galperin, M. Y. (1998). Beyond complete genomes: from sequence to structure and function. *Curr. Opin. Struct. Biol.* 8, 355-363.
- [10] Linal, M., and Yona, G. (1999). Methodologies for target selection in structural genomics. *Prog. Biophys. Mol. Biol* in press.
- [11] Montelione, G. T., and Anderson, S. (1999). Structural genomics: keystone for a Human Proteome Project. *Nature Struct. Biol.* 6, 11-12.
- [12] Murzin, A. G. (1996). Structural classification of proteins: new superfamilies. *Curr. Opin. Struct. Biol.* 6, 386-394.
- [13] Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J Mol Biol* 247, 536-540.
- [14] Orengo, C. A., Jones, D. T., and Thornton, J. M. (1994). Protein superfamilies and domain superfolds. *Nature* 372, 631-634.
- [15] Orengo, C. A., Michie, A. D., Jones, S., Jones, D. T., Swindells, M. B., and Thornton, J. M. (1997). CATH--a hierarchic classification of protein domain structures. *Structure* 5, 1093-1108.
- [16] Sali, A. (1998). 100,000 protein structures for the biologist. *Nature Struct. Biol.* 5, 1029-1032.
- [17] Terwilliger, T. C., Waldo, G., Peat, T. S., Newman, J. M., Chu, K., and Berendzen, J. (1998) *Class-directed structure determination: foundation for a protein structure initiative.* *Protein Sci.* 7, 1851-1856.
- [18] Wang, Z. X. (1998). A re-estimation for the total numbers of protein folds and superfamilies. *Protein Eng.* 11, 621-626.
- [19] Yona, G., Linal, N., and Linal, M. (1999). ProtoMap -- Automated classification of all proteins sequences: a hierarchy of protein families, and local maps of the protein space. *Proteins*, (in press).
- [20] Zhang, C., and DeLisi, C. (1998). Estimating the number of protein folds. *J. Mol. Biol.* 284, 1301-1305.