



Predicting fold novelty based on ProtoNet hierarchical classification

Ilona Kifer^{1,2}, Ori Sasson² and Michal Linial^{1,*}

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

Received on December 19, 2003; revised on October 18, 2004; accepted on November 2, 2004
Advance Access publication ...

ABSTRACT

Motivation: Structural genomics projects aim to solve a large number of protein structures with the ultimate objective of representing the entire protein space. The computational challenge is to identify and prioritize a small set of proteins with new, currently unknown, superfamilies or folds.

Results: We develop a method that assigns each protein a likelihood of it belonging to a new, yet undetermined, structural superfamily. The method relies on a variant of ProtoNet, an automatic hierarchical classification scheme of all protein sequences from SwissProt. Our results show that proteins that are remote from solved structures in the ProtoNet hierarchy are more likely to belong to new superfamilies. The results are validated against SCOP releases from recent years that account for about half of the solved structures known to date. We show that our new method and the representation of ProtoNet are superior in detecting new targets, compared to our previous method using ProtoMap classification. Furthermore, our method outperforms PSI-BLAST search in detecting potential new superfamilies.

Availability: An interactive tool implementing this method, named ProTarget, is available at <http://www.protarget.cs.huji.ac.il>. It can be used interactively to retrieve a list of candidate proteins for Structural genomics projects. Supplementary material is available at <http://www.protarget.cs.huji.ac.il/supplement>

Contact: michall@cc.huji.ac.il

INTRODUCTION

Recent years have seen an explosive growth in the number of publicly available protein sequences, much as a result of many large-scale sequencing projects, including that of the human genome. The number of publicly available protein sequences exceeds one million, yet the number of proteins for which the 3D structure has been determined is significantly smaller. Currently, the number of protein sequences exceeds the number of 3D solved structures by more than 50-fold. During a period of 3 years (March 2000–2003), 8000 new

entries had been added to the PDB, yet, at this period only 410 superfamilies (5% of all these entries) were actually new, according to SCOP classification (Lo Conte *et al.*, 2000). Therefore, it is desirable to develop methods that will lead to an increase in the success rate of identifying new superfamilies (Chance *et al.*, 2002; Sanchez *et al.*, 2002; Vitkup *et al.*, 2001).

The goal of structural genomics (SG) is to cover the protein fold space, and in particular to complete the structural representatives of all proteins in selected model organisms (Burley and Bonanno, 2002). One of the most important tasks in SG is *target selection* (Brenner, 2000). Target selection is the process of choosing protein sequences for structural determination (Sali, 1998). However, the actual number of proteins required for achieving the goal of covering the entire protein structural space remains unknown (Brenner and Levitt, 2000; Elofsson and Sonnhammer, 1999; Liu and Rost, 2002, 2003; Vitkup *et al.*, 2001).

Several complementary strategies were applied to facilitate new superfamily and fold discovery (Zhang and Kim, 2003). According to one approach, all hypothetical proteins that lack homologues in other organisms are selected (Eswaramoorthy *et al.*, 2003; Zarembinski *et al.*, 1998). An alternative naïve strategy is to apply state-of-the-art methods for detecting remote homologues for all presently solved proteins in the PDB such as PSI-BLAST and SAM-T99 (Altschul *et al.*, 1997; Karplus *et al.*, 1998). Proteins that are not included in the hit list of such searches (above a predetermined threshold) are considered as potential targets (Brenner *et al.*, 1998; Elofsson and Sonnhammer, 1999). An exhaustive target list for about 60 genomes was compiled (Carter *et al.*, 2003; Gough and Chothia, 2002). The resulting structural fragments were then proposed as candidates for structural determination. However, no prioritizing methodology that ranks these proposed targets by the probability of being a new fold has been proposed.

We present an alternative approach that is based on a global statistical–computational learning procedure. In a previous work (Portugaly *et al.*, 2002), a target list that is rich in new superfamilies was created based on a map of the protein

*To whom correspondence should be addressed.

sequence space as captured ProtoMap (Yona *et al.*, 2000). Portugaly and Linial (2000) have introduced a naïve measure that captures the minimal volume around a protein at which a solved structure is encountered. This measure was used as a basis for prioritizing proteins as candidates for SG target lists. The shared principle for that study and the present one is the notion that having a scaffold of the protein sequence space is instrumental for identifying parts of the space that are not yet occupied by any solved structure (Linial and Yona, 2000). The number of proteins that is included in the current analysis was updated to cope with recent growth in the SwissProt and PDB databases. However the methodology for prioritizing proteins as SG targets is markedly different. Most significantly, the scheme for organizing the protein space and the statistical model underlying the search for new superfamilies were changed. Herein, we take advantage of ProtoNet—an agglomerative hierarchical clustering of all protein sequences (Sasson *et al.*, 2003).

We compare the performance of our method to prediction for new superfamilies based on a previously published work (Portugaly *et al.*, 2002) and show that the current ProtoNet-based classification is superior in ranking new superfamilies. The new concept for global protein classification that is implemented in ProtoNet markedly improves the prediction for a new superfamily compared to the ProtoMap-based method. Furthermore, we show that such improvement in ranking new superfamilies cannot be attributed to the enlarged set of proteins or to the prediction measure that differs between the ProtoMap-based method and the current one.

We introduce a new measure based on the tree of ProtoNet to identify candidate proteins for structural determination. We take advantage of the tree hierarchy of ProtoNet to define the best separation level between proteins that belong to already known and previously unknown (new) superfamilies. The separator we have found identifies 85% of the unsolved proteins, and labels correctly 33% of the proteins predicted as new superfamilies. Finally, a value that reflects the confidence in our prediction was assigned to each query protein. In practice, we have developed a website, ProTarget, that can be used to predict the likelihood of any protein sequence to contain a new structural superfamily. In ProTarget the input sequence is further filtered against sequences that have already been solved. This facilitates the evaluation of sequences, looking for a new superfamily, even in the case of a multi-domain protein which contains previously solved domains as well as new (unsolved) domains.

METHODOLOGY

Databases

A computational–statistical method was developed to estimate how likely a protein is to represent a new superfamily. This approach is based on a synthesis of three

classifications—SCOP, a structure-based classification (Lo Conte *et al.*, 2000); ProtoNet, a sequence-based hierarchical tree (Sasson *et al.*, 2003) and Proto3D. The latter is a ProtoNet-like tree in which all sequences of domains from the PDB are included. Proto3D includes all 114,033 SwissProt sequences (release 40.28) and all domains from PDB dated September 2002. A total of ~36 700 sequences of structural domains from the PDB were included after excluding those belonging to ‘non-true’ SCOP classes. SCOP versions used for validation tests are from March 2000 (SCOP 1.55) to March 2003 (SCOP 1.63). SCOP can be accessed at <http://www.scop.mrc-lmb.cam.ac.uk/scop/>; ProtoNet is accessible at <http://www.protonet.cs.huji.ac.il>.

ProtoNet is an automatically generated hierarchical classification of all sequences in SwissProt. ProtoNet allows navigation from the individual proteins to the root of the tree, as well as the consideration of several classification algorithms. The results herein refer to the default mode of classification used in ProtoNet (Sanchez *et al.*, 2002). There are over 110 000 merging steps in the entire ProtoNet tree. At an arbitrary horizontal level of the tree (when ~2% of the proteins are singletons), the rest of the proteins (~112 000) are classified to 7800 clusters, 1500 of which contain at least 15 proteins each. Recently, ProtoNet (version 4.0) was extended to cover over one million proteins that include all proteins from SwissProt and from TrEMBL. For the sake of simplicity, the presentation in this paper is limited to the Proto3D database, and does not analyze the last version of ProtoNet.

Clustering method

The clustering method we use is a variant of the ProtoNet clustering algorithm. The SwissProt dataset used in this work has a great deal of redundancy. This problem worsens when looking at sequences of solved structures taken from the PDB database. To avoid bias in the clustering process, we generate a non-redundant database derived from the union of SwissProt and PDB. The outcome of the clustering process was tested to be a valid clustering stage to produce a non-redundant set of sequences. A single protein was chosen from each such cluster and was used as a leaf in our clustering tree. From this point, the default ProtoNet clustering algorithm was performed (details omitted, see supplement).

Navigating the ProtoNet clustering hierarchy

To make practical use of the ProtoNet clustering we need to measure the level in the hierarchy of each cluster. Since the clustering process is agglomerative, each cluster is created out of the merger of two other clusters. We define the ‘birth-time’ of the cluster to capture its level. We introduce two alternative measures for the hierarchy of a cluster in the ProtoNet tree: The first measure captures the progress of the clustering process, through a measure called *ProtoLevel*. The *ProtoLevel* of a cluster is scaled between 0 (all proteins are singletons) and 100 (all proteins are in one cluster).

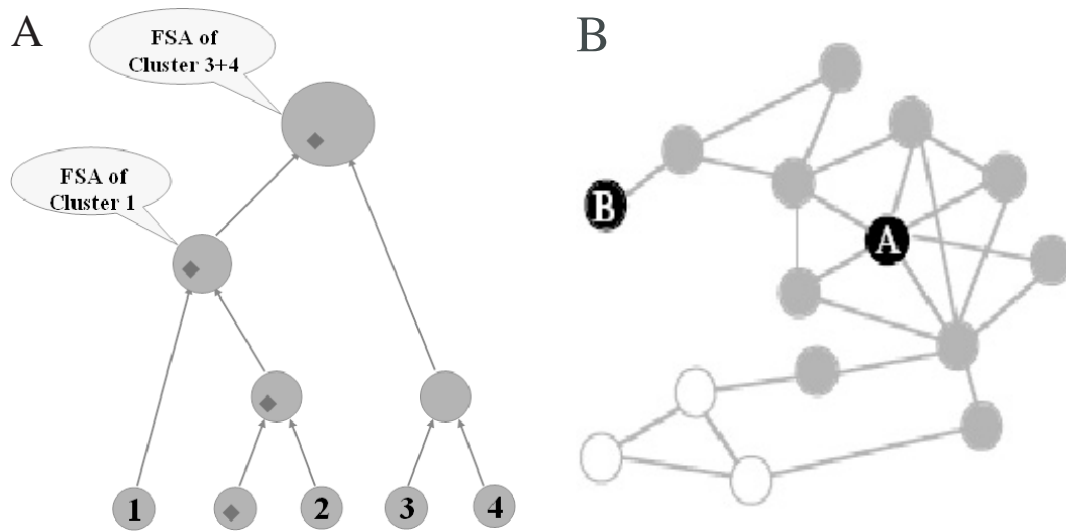


Fig. 1. Schematic illustration of the ProtoNet LSA (A) and the ProtoMap VSA (B) measures. In (A), it shows the progression of hierarchical clustering from bottom to top. The nodes marked 1,2,3,4 indicate protein sequences from SwissProt, whereas the diamond indicates a PDB solved structure. The LSA of a sequence is defined as the lowest cluster in the hierarchy that contains a solved structure. In (B), starting from a cluster of interest A and assuming the closest solved structure is B, the VSV is defined as 11. This value is derived by counting how many clusters can be reached from cluster A within two steps, which is the maximal number of steps which can be traversed from A without reaching cluster B.

The second measure, called *Pair-time*, counts the number of protein pairs within clusters at a certain level, normalized by the number of potential protein pairs. Both measures attempt to capture the progress of the clustering process. The first measure is arbitrary and superimposes the clustering process on a scale of 0–100. Counting the number of pairs within clusters captures the essence of the clustering. The reason for counting pairs as opposed to just counting proteins is that it provides a combinatorial view of the clustering process, and indicates how many of the potential pairs of proteins were brought together into a single cluster.

If n is the number of proteins being clustered, the number of potential protein pairs is $n \times (n - 1)/2$. The birth-time of a cluster is defined by this value, counting the number of protein pairs at the time of its creation. It is easy to see that the pair-time of a singleton is 0, since no mergers have yet occurred, and that the pair-time of the root is 1, since the number of protein pairs within clusters is equal to the number of potential protein pairs.

The two measures are interchangeable as there is a one-to-one and unique mapping between a ProtoLevel value and the corresponding Pair-time. Yet, they provide a different scaling of the Pair-clustering process (the Pair-time measure assigns a large amount of clusters with very low birth-time values, while the ProtoLevel measure assigns a large amount of clusters with very high birth-time values). For a better visualization of the data we have applied either one of the hierarchical scaling representation: ProtoLevel (ranges 0–100) or \log_{10} (Pair-time).

We use these measures to define the distance between two leaves in the clustering tree (i.e. two protein sequences). The distance is defined as the birth-time of the lowest common ancestor (LCA) of the two leaves. Since the hierarchical clustering progress is based on sequence similarity, the most similar sequences are merged together first. Thus, the higher the birth-time of the LCA of two sequences is, the further away they are.

We apply this distance measure to the clustering tree to determine how far a given sequence is from any ‘solved’ sequence (i.e. a sequence for which the structure is already known). We define the distance of a leaf from a ‘solved’ leaf (a leaf containing at least one solved PDB domain) as the birth-time of the *lowest solved ancestor* (LSA)—the lowest ancestor cluster containing a solved PDB structure. Figure 1A illustrates the notion of LSA.

Predicting structural novelty

Our view is that there is a correspondence between the LSA birth-time of a protein and the probability of it containing a new superfamily domain. In other words, we expect that the later the birth-time of the LSA cluster of a protein, the higher the chance that it has an unsolved structure.

We propose a method called ProTarget for predicting whether a protein has a new structure. This method measures the LSA birth-time of the protein and compares it to a fixed threshold. The choice of this threshold is described in the Results section.

Testing

We test out prediction method using different versions of SCOP. To obtain statistically significant results, we include in our test set all domains solved in a consecutive series of SCOP versions (from 1.55 to 1.63, dated March 2001–March 2003). Each domain in this combined set is labeled 1 (positive sample) if at the time it was solved it uncovered a new superfamily of SCOP and -1 (negative sample) if it joined an already existing one.

We compare the performance of our method with two other methods: PSI-BLAST and VSV. PSI-BLAST (Altschul *et al.*, 1997) is frequently used for finding structural similarities based on sequence information (e.g. Jones, 1999). VSV stands for ‘vacant surrounding volume’, and is the measure introduced in Portugaly *et al.* (2002). The VSV measure attempts to quantify the distance between a given cluster and any cluster which contains a solved sequence. This is achieved by counting how many clusters around a certain cluster do not contain a solved sequence (Fig. 1B).

Comparing the LSA method with the VSV method is not a straightforward affair since the latter is based on the ProtoMap classification scheme (Yona *et al.*, 2000). We studied ~~an earlier version of SCOP. We considered~~ SCOP version 1.37 to create the base samples and version 1.50 ~~for the~~ test set. Our set of samples consists of clusters of proteins at the bottom level of the hierarchy. An additional step was needed to map the labeled domains onto these clusters. Each domain in the test set was associated with the SwissProt proteins that showed a significant sequence similarity (BLAST E-score $\leq 1e-30$).

In testing the LSA method, our sample set includes clusters that contained at least one solved SCOP domain from the test set. We excluded from this set all clusters containing a domain that was solved before the SCOP versions used in the test set. This exclusion was necessary as the elementary unit for classification in ProtoNet is a whole protein (which often is a multi-domain protein), while structural determination (as in PDB and SCOP) is often performed on a single domain of a protein. A sample was labeled 1 if it contained at least one domain from the test set which was labeled 1 and -1 otherwise.

For each sample the birth-time of the LSA was calculated, yielding a set of labeled samples. Sorting this set according to the birth-time of the LSA values allows us to examine our assumption as we should be able to separate the samples labeled 1 from the samples labeled -1 , using a 1D linear separator.

RESULTS

Using the hierarchical navigation of the ProtoNet tree described above, we test the validity of our view that the later the birth-time of the LSA of a protein, the higher the chance it is a new, unsolved structure.

Verifying the LSA concept

We consider several combinations of base and test sets, using SCOP versions that were released from March 2001 till March 2003. About 6500 new solved structures with 18000 domains were added during this period to the PDB. While we repeated the tests for all possible independent pairs of SCOP releases (such as 1.55 to 1.57; 1.57 to 1.59; 1.59 to 1.61; 1.61 to 1.63 and combinations of these five major SCOP releases), the results for those different datasets were very much similar. In order to obtain statistical significance, the dataset presented here contains all domains solved between (and including) versions 1.55 and 1.63 of SCOP (referred to as 1.55–1.63). This test set contains 1111 sample clusters, of them 241 being (22%) positive samples, and 870 (78%) negative ones.

We studied the distribution of the LSA birth-time for both positive and negative samples in different ProtoLevels, seeking a correlation between the birth-time of the LSA and the likelihood of a protein belonging to a new superfamily. A crucial point in this analysis was to determine the LSA of a protein only among clusters containing structures solved in earlier versions (i.e. previous to the ones used for the test set). The correlation we found is presented in Figure 2. For each Protolevel value p , we show the percentage of positive samples that were assigned a Protolevel value $\leq p$, and likewise for the negative samples. As seen in Figure 2, $\sim 40\%$ of the negative samples and only $\sim 10\%$ of the positive samples accumulate at ProtoLevel ≤ 0 (extreme left of the graph), while $\sim 60\%$ of the positive samples and $\sim 25\%$ of the negative samples are assigned a Protolevel ≥ 95 (extreme right of the graph, accumulating all samples at 95 and above). Noticeably, the area between these two regions contains only a small fraction of the samples. The extreme left of the graph is mostly composed of negative samples, suggesting that a low ProtoLevel is indicative of belonging to a known superfamily. The extreme right of the graph consists mostly of positive samples, suggesting that a high Protolevel is indicative of belonging to new superfamilies.

Performance evaluation

In order to gauge the performance of our method in comparison to the VSV method (Portugaly *et al.*, 2002), we had to use older data to replicate the reported results, namely SwissProt release 36 ($\sim 94\,000$ proteins) and SCOP 1.50 (10 650 PDB, $\sim 24\,000$ domains, February 2000). The domains that were solved until SCOP 1.37 (~ 6500 PDB, $\sim 13\,000$ domains, October 1997) were the train set, while the domains solved between 1.37 and 1.50 composed the test set. The samples we used were the result of the ~~first-stage~~ clustering algorithm ~~with a~~ threshold of $1e-100$. Samples were labeled $+1$ if they belonged to a superfamily from 1.37 to 1.50 and -1 otherwise.

To compare the methods, each sample was associated with two different values—the VSV corresponding to its distance

Author: pl.
confirm if it is
five or four.

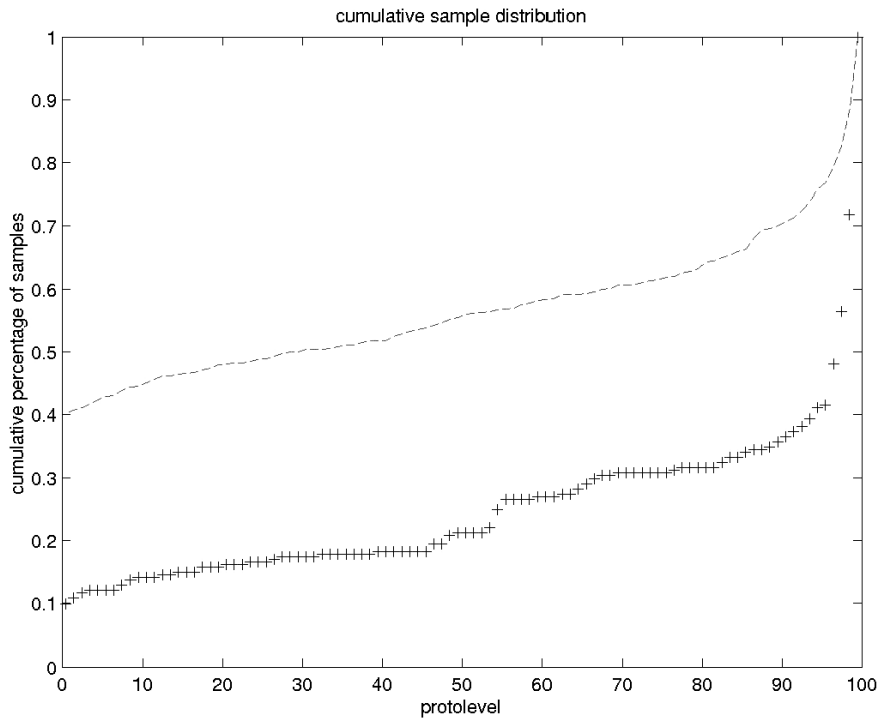


Fig. 2. Cumulative distribution of positive and negative samples for different ProtoLevels. Samples of known superfamilies (negative) and new ones (positive) are marked by a dashed line and a plus symbol, respectively. The 40% negative samples and 10% positive samples at ProtoLevel = 0 indicate the large amount of samples which were already assigned at the beginning of the clustering procedure (at ProtoLevel ranges 0–1). Almost a third of the positive samples have a ProtoLevel value that is in the range of 99–100. Both graphs merge at the upper right corner.

in the ProtoMap graph at the $1e-100$ granularity level and the birth-time of its LCA in the ProtoNet tree. We use the methodology presented in Portugaly *et al.* (2002) to compare these values. The idea here is that a labeling of clusters imposes an ordering of the sample space. The quality of an ordering can be measured as to how far it is from the perfectly correct ordering (i.e. of samples marked by -1 associate with lower VSV than those marked $+1$). Thus, we define a score for an ordering as the number of swaps of adjacent samples that need to be performed in order to transform it into a perfect order.

With this measure we can proceed to define a P -value. This P -value measures the chance of obtaining this score or higher from a random ordering. The motivation for this measure is that we want the ordering obtained to be far from random, since the further the ordering is from a random ordering, the difficulty in calculating this P -value for VSV labeling is that VSV is not a unique measure. Therefore, distinct samples are frequently assigned with the same VSV, hence lacking strict ordering between samples. This prevents us from counting the number of swaps needed to reach the perfectly correct ordering. To choose a strict ordering of the samples, a random ordering was imposed for all samples with identical VSV. To compensate for the random choice, this process was repeated 200 times (yielding 200 orders).

Figure 3 shows the P -values obtained for three methods—PSI-BLAST, VSV and the ProTarget-based LSA method described herein. In our analysis we only used samples that were not ‘solved’—did not contain a structure that had been solved until SCOP 1.37. To focus on the performance of our method in non-trivial cases, we studied the case (Fig. 3B) where we removed the samples that were neighbors of solved samples as well (which were, with high probability, samples labeled -1). The set that was left is assumed to be harder for prediction. A comparison with PSI-BLAST is also presented, and it follows the calculation made in Portugaly *et al.* (2002). It is clear that the P -values for ProTarget are much better than the P -values achieved by the other two methods and thus.

Figure 3 clearly shows that the performance of the ProTarget method on this data is superior to the previous ones.

Measuring ProTarget performance

Compared to PSI-BLAST, the VSV-based method and the ProTarget method indicate superior results for the ProTarget for the SCOP 1.37–1.50 validation test. ProTarget results yielded a P -value close to zero, indicating they were almost impossible to achieve with random ordering.

Please check completeness of sense in “The motivation...”

Please check completeness of sentence “It is clear...”

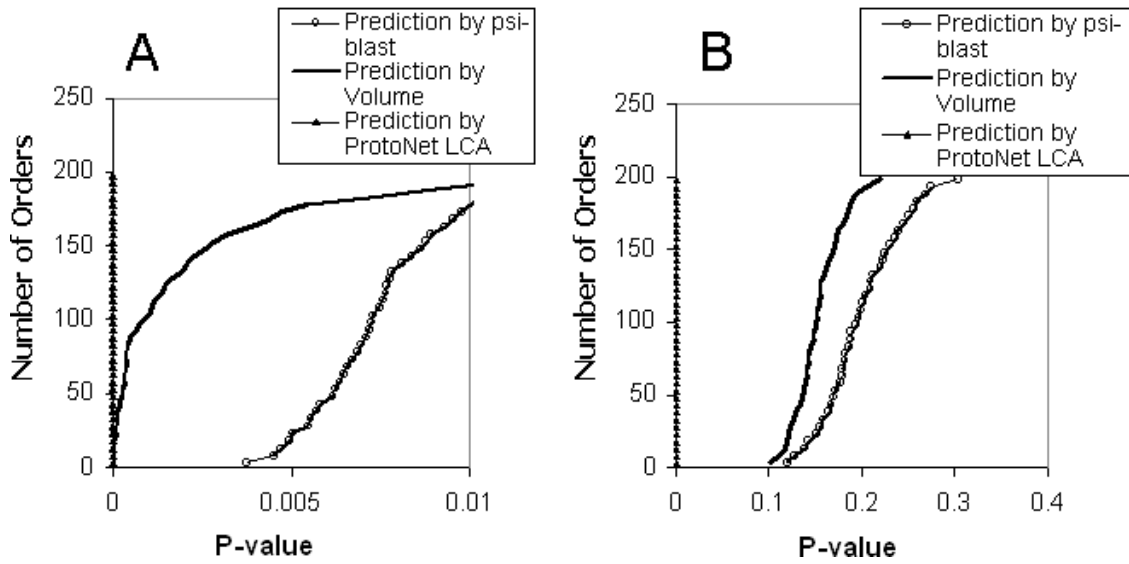


Fig. 3. Validation test for SCOP 1.37–1.50, showing P -values for 200 different choices of ordering. P -value here measures the probability of a random ordering to match or outperform the given ordering. **(A)** Results of all samples that were not marked as ‘solved’ (contained a solved structure from version 1.37 and before) are shown. **(B)** Only samples that are not ‘solved’ and that are not neighbors of ‘solved’ clusters in ProtoMap are shown. Note that PSI-BLAST performs worse than the VSV-based method (Portugaly *et al.*, 2002) and ProTarget outperforms both methods (the P -value for ProTarget overlaps the Y -axis). The P -values achieved by ProTarget are indicated by triangles, and were 0 in (A) and 0.00113–0.00129 in (B). PSI-BLAST parameters used: BLOSUM62, gap penalty and gap extension are 14 and 1, iteration threshold of E-0.001, maximal number of iteration is 10 and E-score threshold is 100.

To allow us to measure the effectiveness of the ProTarget method, we define a function that captures the quality of separation between positive and negative samples for a given ProtoLevel p , and seek to minimize it:

$$Q(p) = \alpha \frac{FN}{\# \text{ positive}} + (1 - \alpha) \frac{FP}{\# \text{ negative}}$$

We use the value 0.5 , $\alpha = 95$, meaning that both types of errors (i.e. FP and FN) are equally important. This is not necessarily the case (see Discussion). Figure 4 shows the behavior of Q for different $Pair\text{-}time$ values, obtained for all different non-overlapping test sets of SCOP versions, and their union. Note that we included versions of SCOP that include $\sim 50\,000$ domains (version 1.63), as opposed to only $\sim 24\,000$ domains (version 1.50) that were considered in the comparative test.

The minimum of this function was obtained at different values of ProtoLevel for different test sets [transformed to $\log_{10}(Pair\text{-}time)$, see Methodology]. For all test sets the minima were very similar. Thus, we chose the union of all datasets (SCOP 1.55–1.63) and applied a smoothing technique to its error function. Smoothing is attained by fitting a parabola that minimizes the sum-of-square error to the data. The minimum following such smoothing was equal to $\log_{10}(Pair\text{-}time) = -2.9695$ which corresponds to ProtoLevel ~ 44 . At the minima, the error function was equal to 32.9%. This error is composed of $\sim 47\%$ FP and

$\sim 18\%$ FN. The high error of FP reflects the fact that many proteins with undetected sequence similarity may still belong to the same 3D structure at the superfamily level. The results are summarized in Table 1.

Prediction confidence level

Our prediction method is based on simple linear separation. In other words, if the LSA for a certain sequence is above a certain value, we conclude the sequence is likely to be a new superfamily. Figure 4 suggests that there is a high level of uncertainty in this prediction, especially when the LSA is close to the threshold level. Consequently, the further the LSA is from the threshold, the higher is the confidence in the prediction. This observation can be used for obtaining a statistical measure of confidence in our prediction. Such a value is useful in practical applications of target selection, in order to rank the most promising targets.

We define the confidence for a specific ProtoLevel p as:

- For a positive prediction,

$$\begin{aligned} \text{confidence}(p > \text{threshold}) &= 1 - P(-1 | p > \text{threshold}) \\ &= 1 - \frac{\sum_{s=\text{th}}^p I\{s = -1\}}{\sum_{s=\text{th}}^p 1}, \end{aligned}$$

where I is the indicator function. The expression $I\{s = -1\}$ is 1 if the true label of the sample is -1 and

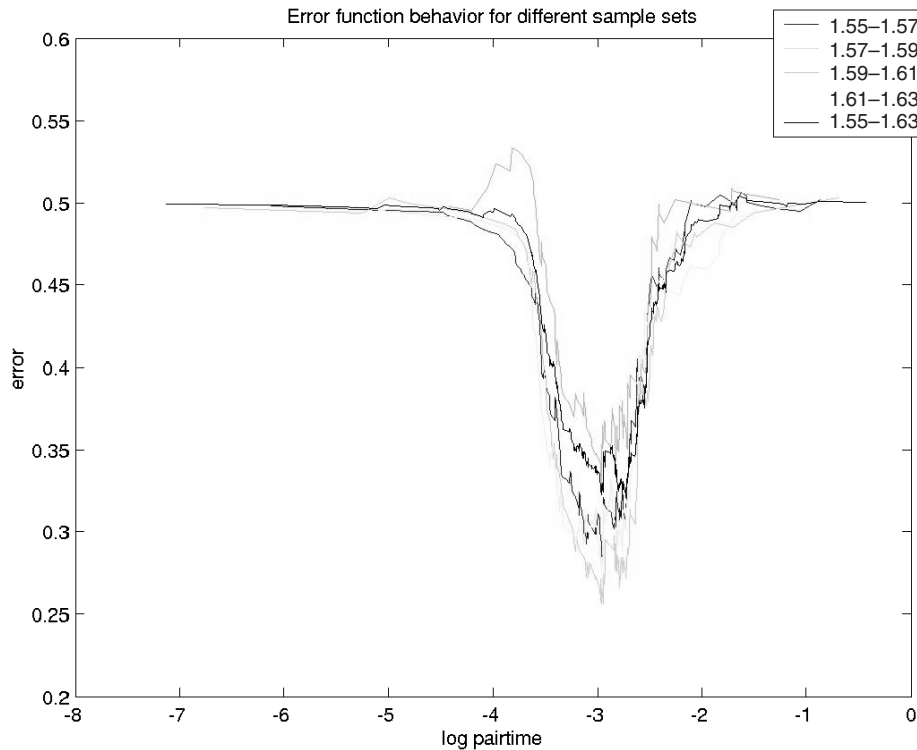


Fig. 4. The quality of separation Q for different hierarchical thresholds presented by the Pair-time on four independent test sets (marked 1.55–1.57; 1.57–1.59; 1.59–1.61; 1.61–1.63 in grey color) and a test set combining all (marked as 1.55–1.63 in black). ~18 000 redundant domains were added from SCOP 1.55 to 1.63.

Table 1. Summary of error for different SCOP sample sets

Test set SCOP pairs	Number of samples (positive/negative)	Minimizing log(PT)/ ProtoLevel	Total error (Q) (%)	FP fraction (%)	FN fraction (%)
1.55–1.57	329 (70/259)	–2.95054/52.322	28.2	37.8	18.6
1.57–1.59	216 (55/161)	–2.94883/54.1925	25.6	34.8	16.4
1.59–1.61	251 (35/216)	–2.96236/48.5372	32.5	44.9	20
1.61–1.63	266 (55/211)	–2.95927/49.4395	27.6	37	18.2
1.55–1.63	1111 (241/870)	–2.73122/95.911	32.1	23.1	18.3
Smoothing parabola	—	–2.96586/46.6186	32.2	46.2	41.1
		–2.9695/~44	32.9	47	18.6

Note that for the combining test set, two very close minima were observed. The smoothing enforces the minima at the region of $\log_{10}(\text{Pair-time})$ of –2.9695, which is translated to ProtoLevel ~44, a value very similar to the minima obtained by all other independent subsets.

0 if it is not. The sum in the quotient is on all samples with a value between th and p , while the denominator sums only on the FP samples between th and p .

- For a negative prediction,

$$\begin{aligned}
 & \text{confidence}(p < \text{threshold}) \\
 &= 1 - P(1|p < \text{threshold}) \\
 &= 1 - \frac{\sum_{s=th}^p I\{s = -1\}}{\sum_{s=th}^p 1}.
 \end{aligned}$$

Separation by other ProtoNet features

The analysis described above indicates that we have achieved a certain level of separation using the ProtoLevel birth-time of the LSA. The choice of this parameter may seem arbitrary. In order to ensure that we are using the best possible measure for separation, we considered other features of the ProtoNet tree, and their combinations. The results show that this is in fact the best feature to use for separation (details omitted).

One assumption we made when learning from the data is that the separation is indeed linear. This assumption reflects our

Table 2. A comparison between SVM on all ProtoNet features listed in Table 2 and an exhaustive search on birth-time alone. (for definition, see text)

Method	Total error (%)	Sensitivity (coverage) (%)	Specificity (accuracy) (%)
Exhaustive search (smoothed)	32.1	82	33
SVM	31.9	74.3	35.1

perception of the ProtoNet tree: we expect a linear correlation (if any) between each feature of the tree and the probability of it belonging to a new superfamily. While this assumption is intuitive, it still lacks affirmation. We compared the exhaustive search method (on the LSA birth-time feature) to an SVM with an RBF kernel using the SVM-light package (Joachims, 1999; Zavaljevski *et al.*, 2002) that was learned on all features together. SVM is a method for learning not necessarily linear separators, while attempting to minimize the percentage of mistakes (approximately). Specificity and sensitivity are defined as follows: specificity is defined as TP/(TP + FP) and accounts for the percentage of positive samples out of all samples predicted as positive (accuracy); sensitivity is defined as TP/(TP + FN) and represents the percentage of positive samples predicted correctly (coverage). Since the negative samples highly outnumber the positive samples, each FN error was weighed appropriately higher than an FP error. The learning was done using the leave-one-out method, and the results for both methods are presented in Table 2.

As seen from Table 2, the SVM does not improve the results of the exhaustive search using birth-time alone in any significant way. ~~This strengthens our assumption that if a separation exists it is a linear one.~~ The results reinforce our view that the additional features of the ProtoNet tree do not add to the overall quality of separation.

ProTarget Web tool

The above analysis led to the development of a tool, ProTarget (<http://www.protarget.cs.huji.ac.il>), for ranking user-supplied targets according to their propensity of belonging to new superfamilies.

Conceptually, the process we need to follow when analyzing a new protein consists of three steps. First we perform a BLAST search of the query protein against the database of all solved structures (i.e. PDB). If the query protein is similar enough (BLAST e-score $\leq 1e-5$) to one or more solved domains, it is broken down into several pieces (a procedure referred to as ‘crop’). All fragments that overlap with a known domain are filtered out, while each of the remaining fragments that are long enough (>30 aa) is subjected to the following steps.

We insert a new protein or a partial sequence into the ProtoNet clustering, by associating it with the most suitable ProtoNet cluster, based on its BLAST similarity to other sequences. In some cases, there is no apparent similarity to any protein in ProtoNet; in such cases the protein is marked as ‘isolated’ and is not treated any further. The next step involves calculating the birth-time of the LSA of the cluster, and finally a prediction is given according to whether this value is above or below the predetermined threshold (the minima of the smoothed Q function), together with a confidence score.

Marking proteins, or protein fragments as ‘isolated’, may have several interpretations. On the one hand, their sequence is distant from any other known protein and thus suggests a strong candidate for a new structure. On the other hand, we cannot provide any prediction for those proteins using our analysis.

DISCUSSION

The first step in any structural genomics project is the selection of appropriate targets. The effort to compose a filtered, non-overlapping target list for the entire SG community has been presented (Bray *et al.*, 2004; Goldsmith-Fischman and Honig, 2003). Such a synthesized target list contains over 40 000 potential targets originating in international SG centers and has been deposited in the PDB (Westbrook *et al.*, 2003). The synthesis of all targets originating from the different SG centers suffers from a large degree of redundancy, as well as over-representation of targets in certain families while other families are ignored (Sasson and Linial, unpublished results).

We present a methodology for prioritizing target proteins for structure determination. The ProTarget method does not rely on considerations such as methodology of choice (NMR, X-ray crystallography), the status of the proteome (a complete genome or not), protein length and protein taxonomical origin. Thus, as long as the query protein is not entirely covered by already solved domains, it is a valid candidate for our analysis. But how valid is the ProTarget method to suggest new fold or superfamilies considering that the number of solved structures in mid-2004 already reached 25 000? *A priori*, the impact of accumulating new structures on the predicting power is not known. Figure 4 and Table 1 indicate that the method is very robust with respect to the addition of new structural data. Almost identical prediction results are shown for all independent test sets used in our study. The other component in our prediction method is the ProtoNet itself. We show that the ProtoNet that is based on only 94 000 proteins (Swissprot 36.0, Figure 4) produces results that are similar to those obtained with larger sets of 1 14 000 (Swissprot 40.28, ProtoNet 2.4, Table 2). These results extend to larger databases, such as ProtoNet 4.0 (details omitted).

Our methodology presented here has several limitations due to the current nature of the ProtoNet clustering. The

first limitation stems from the fact that ProtoNet is a sequence-similarity based clustering, and sequence similarity does not account for all structural similarity. This limitation underlies the high percentage of FP in our prediction. Still, ProtoNet provides a solid scheme to detect functionally and structurally remote homologues (Shachar and Linial, 2004). The second limitation originates from ProtoNet being a protein-based classification, rather than a domain-based one. This forced us to disregard multi-domain proteins as samples. When analyzing such proteins, some of the domains could be misrepresented in the clustering process, and may cause a misleading prediction. Thus, we cannot be certain whether our sample space truly represents the protein structural-space.

The multi-domain issue is partially addressed by using the ‘crop’ procedure. This procedure allows a domain in a protein to be selected as a target for SG despite the fact that an additional domain in the same polypeptide chain had been already structurally solved. Still, the ProTarget method suffers from the need to use arbitrary parameters for considering a region in the protein as a target (i.e. threshold of BLAST similarity to PDB, and minimal fragment length that is considered significant). Another outcome of the ‘crop’ step is the potential creation of artificial short fragments that are labelled as ‘isolated’; many of them are linkers, loops and unstructured regions. A proper treatment of the multi-domain issue is expected by using a domain-based clustering of the protein space. This work is currently in progress.

The ProTarget tool provides a prioritizing method for target selection in the scope of SG. Among the top list of targets we encounter many clusters consisting of membranous proteins (with two or more transmembrane domains). This is in accord with the current state of very low structural coverage of membranous proteins. In our view, the membranous proteins are ‘remote’ from other solved structures and thus their confidence of being a new superfamily is high. Different users in the SG community may have different considerations when selecting targets. For instance, one may require higher coverage at the expense of solving many already known structures. The α factor in function Q gives the option of weighting the importance of the accuracy and the coverage according to the experimentalist’s requirements. Different choices of α will result in different minima for Q , thus producing a different separator threshold (and consequently, altered confidence values).

ACKNOWLEDGEMENTS

We thank Nati Linial and Elon Portugaly for their valuable suggestions, ideas and fruitful discussions throughout this study. The authors wish to thank the outstanding ProtoNet team and Alex Savenok for ProTarget web design. This study was partially supported by the CESG consortium (NIMSG, NIH) and the European SPINE consortium. I.K. is a fellow student of SCCB—The Sudarsky Center for Computational Biology in the Hebrew University of Jerusalem.

REFERENCES

- Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
- Bray,J.E., Marsden,R.L., Rison,S.C., Savchenko,A., Edwards,A.M., Thornton,J.M. and Orengo,C.A. (2004) A practical and robust sequence search strategy for structural genomics target selection. *Bioinformatics*, **20**, 2288–2295.
- Brenner,S.E. (2000) Target selection for structural genomics. *Nat. Struct. Biol.*, **7** (Suppl), 967–969.
- Brenner,S.E., Chothia,C. and Hubbard,T.J. (1998) Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships. *Proc. Natl Acad. Sci. USA*, **95**, 6073–6078.
- Brenner,S.E. and Levitt,M. (2000) Expectations from structural genomics. *Protein Sci.*, **9**, 197–200.
- Burley,S.K. and Bonanno,J.B. (2002) Structural genomics of proteins from conserved biochemical pathways and processes. *Curr. Opin. Struct. Biol.*, **12**, 383–391.
- Carter,P., Liu,J. and Rost,B. (2003) PEP: predictions for entire proteomes. *Nucleic Acids Res.*, **31**, 410–413.
- Chance,M.R., Bresnick,A.R., Burley,S.K., Jiang,J.S., Lima,C.D., Sali,A., Almo,S.C., Bonanno,J.B., Buglino,J.A., Boulton,S. *et al.* (2002) Structural genomics: a pipeline for providing structures for the biologist. *Protein Sci.*, **11**, 723–738.
- Elofsson,A. and Sonnhammer,E.L. (1999) A comparison of sequence and structure protein domain families as a basis for structural genomics. *Bioinformatics*, **15**, 480–500.
- Eswaramoorthy,S., Gerchman,S., Graziano,V., Kycia,H., Studier,F.W. and Swaminathan,S. (2003) Structure of a yeast hypothetical protein selected by a structural genomics approach. *Acta Crystallogr. D Biol. Crystallogr.*, **59**, 127–135.
- Goldsmith-Fischman,S. and Honig,B. (2003) Structural genomics: computational methods for structure analysis. *Protein Sci.*, **12**, 1813–1821.
- Gough,J. and Chothia,C. (2002) SUPERFAMILY: HMMs representing all proteins of known structure. SCOP sequence searches, alignments and genome assignments. *Nucleic Acids Res.*, **30**, 268–272.
- Joachims,T. (1999) Making Large-Scale SVM Learning Practical. MIT Press.
- Jones,D.T. (1999) Protein secondary structure prediction based on position-specific scoring matrices. *J. Mol. Biol.*, **292**, 195–202.
- Karplus,K., Barrett,C. and Hughey,R. (1998) Hidden Markov models for detecting remote protein homologies. *Bioinformatics*, **14**, 846–856.
- Linial,M. and Yona,G. (2000) Methodologies for target selection in structural genomics. *Progr. Biophys. Mol. Biol.*, **73**, 297–320.
- Liu,J. and Rost,B. (2002) Target space for structural genomics revisited. *Bioinformatics*, **18**, 922–933.
- Liu,J. and Rost,B. (2003) Domains, motifs and clusters in the protein universe. *Curr. Opin. Chem. Biol.*, **7**, 5–11.
- Lo Conte,L., Ailey,B., Hubbard,T.J., Brenner,S.E., Murzin,A.G. and Chothia,C. (2000) SCOP: a structural classification of proteins database. *Nucleic Acids Res.*, **28**, 257–259.
- Portugaly,E., Kifer,I. and Linial,M. (2002) Selecting targets for structural determination by navigating in a graph of protein families. *Bioinformatics*, **18**, 899–907.

Pl. provide
publisher
details

- Portugaly,E. and Linial,M. (2000) Estimating the probability for a protein to have a new fold: a statistical computational model. *Proc. Natl Acad. Sci. USA*, **97**, 5161–5166.
- Sali,A. (1998) 100,000 protein structures for the biologist. *Nat. Struct. Biol.*, **5**, 1029–1032.
- Sanchez,R., Pieper,U., Melo,F., Eswar,N., Marti-Renom,M.A., Madhusudhan,M.S., Mirkovic,N. and Linial,M. (2002) The metric space of proteins-comparative study of clustering algorithms. *Bioinformatics*, **18**, S14–21.
- Sasson,O., Vaaknin,A., Fleischer,H., Portugaly,E., Bilu,Y., Linial,N. and Linial,M. (2003) ProtoNet: hierarchical classification of the protein space. *Nucleic Acids Res.*, **31**, 348–352.
- Shachar,O. and Linial,M. (2004) A robust method to detect structural and functional remote homologues. *Proteins*, **57**, 531–538.
- Vitkup,D., Melamud,E., Moult,J. and Sander,C. (2001) Completeness in structural genomics. *Nat. Struct. Biol.*, **8**, 559–566.
- Westbrook,J., Feng,Z., Chen,L., Yang,H. and Berman,H.M. (2003) The Protein Data Bank and structural genomics. *Nucleic Acids Res.*, **31**, 489–491.
- Yona,G., Linial,N. and Linial,M. (2000) ProtoMap: automatic classification of protein sequences and hierarchy of protein families. *Nucleic Acids Res.*, **28**, 49–55.
- Zarembinski,T.I., Hung,L.W., Mueller-Dieckmann,H.J., Kim,K.K., Yokota,H., Kim,R. and Kim,S.H. (1998) Structure-based assignment of the biochemical function of a hypothetical protein: a test case of structural genomics. *Proc. Natl Acad. Sci. USA*, **95**, 15189–15193.
- Zavaljevski,N., Stevens,F.J. and Reifman,J. (2002) Support vector machines with selective kernel scaling for protein classification and identification of key amino acid positions. *Bioinformatics*, **18**, 689–696.
- Zhang,C. and Kim,S.H. (2003) Overview of structural genomics: from structure to function. *Curr. Opin. Chem. Biol.*, **7**, 28–32.

Reference is
incomplete
please check