

Functional Differentiation of Proteins: Implications for Structural Genomics

Running Title: Functional Differentiation of Proteins

Iddo Friedberg* and Adam Godzik

Program in Bioinformatics and Systems Biology
Burnham Institute for Medical Research
10901 North Torrey Pines Road
La Jolla, CA 92037, CA USA

(Accepted for publication in *Structure* 2007)

(1)(*) Corresponding author.
Tel: +1 858 646 3100, x3516
Fax: +1 858 713 9949
E-mail: idoerg@burnham.org

(2)
Tel: + 1 858 646 3168
Fax: + 1 858 713 9949
E-mail: adam@burnham.org

Summary

Structural genomics is a broad initiative of various centers aiming to provide a complete coverage of protein structure space. Since it is not feasible to experimentally determine the structures of all proteins, it is generally agreed that the only viable strategy to achieve such coverage is to carefully

select specific proteins (“targets”), determine their structure experimentally, and then use comparative modeling techniques to model the rest. Here we suggest that structural genomics centers refine the structure-driven approach in target selection by adopting function-based criteria. We suggest targeting functionally divergent superfamilies within a given structural fold so that each function receives a structural characterization. We have developed a method to do so, and an itemized survey of several functionally rich folds shows that they are only partially functionally characterized. We call upon structural genomics centers to consider this approach and upon computational biologists to further develop function-based targeting methods.

Supplementary material is available at <http://bioinformatics.burnham.org/~iddo/StructSupp/fbts.html>

Introduction

The size and the detailed composition of protein structure space has recently become a subject of much attention, mostly because of the structural genomics' initiative attempt to provide a full coverage of it (Norvell and Machalek, 2000; Stevens et al., 2001). But what do we mean by full coverage? With sequence genomics, complete genomes have been targeted and solved. In contrast, because of prohibitive cost and limiting technology, protein structure space can be studied only in relatively small number of points. The NIH funded Protein Structure Initiative is promising to solve 3,000 protein structures within next five years (<http://www.nigms.nih.gov/Initiatives/PSI/>), and modeling would leverage this number to cover much larger set (Brenner, 2000; Friedberg et al., 2004; Stevens et al., 2001). But which proteins should be targeted (Chandonia and Brenner, 2005; Cort et al., 1999; Jinfeng Liu, 2004; Linial and Yona, 2000; Liu and Rost, 2002)?

Most structural genomics target selection strategies focus on novelty, but with varying interpretations as to what “novelty” means based on different goals. The latest target selection of PSI centers on families is broadly based on PFAM families. However, we would like to draw attention to the fact that the goal of solving or modeling a protein's atomic structure is to understand the atomic-level implementation of its biochemical function. It follows that if a computationally modeled sequence is predicted to have a significant structural difference from the template it is modeled upon, the consequence would be that functional information shall be wrongly applied to the model based on the template. We therefore suggest that the structural genomics community, in addition to any adopted target selection strategy, should also take care to study representatives of families that are predicted to

have significant functional variations within known structural fold groups. This function-driven approach shall provide structural templates for all functional variants in a fold and help characterize the functional space for a given fold, as the ultimate goal is understanding function through structure. The rationale of our approach is illustrated in Figure 1.

Several computational studies have been conducted in the field of functional discrimination. Some have looked at fine grained discrimination, where sequence similarity is high yet functional differences do exist (Gerlt and Babbitt, 2001; Rost, 2002; Shah and Hunter, 1997; Shah and Hunter, 1998; Tian and Skolnick, 2003). At the other end of the sequence identity spectrum Galperin et al. (Galperin et al., 1998) have shown that enzymes supposedly analogous due to undetectable sequence similarity were, in fact homologous. Pawlowski *et al* (Pawlowski et al., 2000) have shown that a distant but significant sequence similarity correlates well with functional similarity. Devos and Valencia have conducted a study illustrating the problems of function transfer, in light of the ambiguousness of protein function, and the unavoidable database inaccuracies (Devos and Valencia, 2000). Todd *et al* have conducted comprehensive characterizations of structural difference between homologs which are functionally different (Todd et al., 2001; Todd et al., 2002). Clearly, sequence and structure similarities are weakly correlated with functional similarity, with many outliers seen on both ends of the similarity spectrum. Other studies (Hegyí and Gerstein, 1999; Nagano et al., 2002 ; Thornton et al., 1999) have examined the functional content of protein folds and also the fold-level structural content of various functions. Those studies have found significant correlation between structure class and enzyme type and cataloged protein folds displaying the highest functional diversity and enzymatic functions that can be assigned to more than a single fold. These studies have highlighted the difficulty of functional assignment based on fold prediction. Several folds were found to have a high functional divergence, but many others correlated with only a single function. However a more sensitive mapping of sequences to structures (Pandit et al., 2004) has revealed that there may be many more functions yet

unaccounted for inhabiting known fold space and that our current view of functional diversity is biased by the ability to associate sequences with folds and by the inherent fold bias of the Protein Data Bank.

To implement our function driven approach to target selection within known fold space, we use PfamA (Bateman et al., 2002) for sequence space and SCOP (Murzin et al., 1995) for structure space. Each of these databases provides a certain approach to clustering proteins, with PfamA focusing primarily on sequence and with SCOP primarily on structure. PfamA is a manually curated database containing multiple sequence alignments of protein families defined by sequence similarity defined Hidden Markov Models. SCOP is a manually curated hierarchical classification of protein structures. In SCOP, structures are classified by a multilevel tree with different granularities, from coarse to fine, termed class, fold, superfamily, and family. We will show here that the SCOP superfamily level, which identifies proteins with significant sequence and structural divergence, is a very good proxy for function, where proteins from different superfamilies are deemed to have different functions. The Pfam families usually fall somewhere between the SCOP family and superfamily levels and partly depends on the history of annotations and our knowledge of the family (Aloy et al., 2002; Pandit et al., 2004). Our goal here is to identify Pfam families that are likely to populate new SCOP superfamilies, and as such one or more of their member proteins should be targeted by structural genomics, versus such Pfam families that are likely to display minimal functional and structural divergence, and thus fall into one of the known superfamilies. Of course there are many Pfam families for which fold prediction is not possible, at least not in an automated way. These families are not within the scope of the method outlined here.

Having demarcated sequence space and structure space, we implement our “same fold but different

function” approach as follows: first we demonstrate the utility of SCOP superfamily classification as a proxy for functional classification. Second we perform an all-*vs.*-all comparison of representatives of all families in the SCOP database using FFAS03 (Jaroszewski et al., 2005), a sensitive profile–profile alignment tool that we apply for detecting distant sequence-based similarities. Third, we use the SCOP-*vs.*-SCOP assignments to calibrate FFAS03-based similarity, including additional alignment features, namely affine gap penalty scores, and to predict functional difference as measured by proteins being in different SCOP superfamilies and thus having different functions. Finally, we compare representatives of all Pfam families to SCOP families, and using the scoring calibrated in the previous step, pick up matches that are predicted to be similar in fold, yet dissimilar in function (i. e. different superfamily).

Other function-driven approaches were proposed for structural genomics. For instance, prioritizing targets by their medical or biological value (Abergel et al., 2003; Claverie et al., 2002; Goulding et al., 2003) or by their similarity to human proteins (Xie and Bourne, 2005). Another approach suggests focusing on the most important 5,000 families from Pfam, dubbed “Pfam 5000” (Chandonia and Brenner, 2005). Our approach differs from the first two by focusing on superfamilies with potentially unknown, novel functions. The Pfam 5,000 strategy aims at finding one representative of each of the large Pfam families, to achieve accurate fold assignment for a large percentage of all proteins. Function based target selection is more specific in that it selects specific PfamA families whose fold can be predicted, but are predicted to be functionally novel. This study defines a truly function driven approach, as opposed to clustering at a given percent identity level or, conversely, selecting only candidates from new folds. It should be emphasized that the method presented here is for functional differentiation, rather than function prediction. In some cases we may not know the actual function of the selected target, but only that it is predicted to have a different function from proteins with the same structure that have had their structure solved.

In this study we find that our function based target selection is accurate and is suitable for selecting targets for structural genomics. Using our method we suggest some 500 new targets whose fold is known but may have a new function. We also present two case studies for specific proteins predicted by our method to be good candidates for having new functions, and show a verification of our findings in the literature. Finally we identify protein folds that are predicted to be functionally rich, yet the number of structures that have been solved in these folds is relatively small. Improving the structural characterization of these folds will improve our understanding of their functional space on the molecular level.

Results

Validating SCOP superfamilies as a proxy for functional similarity

We used the SCOP superfamily classification as the yardstick for function classification for two reasons. First, superfamily classification is expertly curated for structural and functional similarity. Thus, proteins from different superfamilies but within the same fold are expected to have different functions and significant structure divergence. The functional similarity of proteins within the same SCOP fold and superfamily has been established by several studies (Gough and Chothia, 2002; Hegyi and Gerstein, 1999; Hegyi and Gerstein, 2001; Pandit et al., 2004). Second, structural genomics is part of a structural biology field in which the superfamily-based protein classification is commonly accepted. However, here we also establish that SCOP superfamilies are indeed a fitting proxy for function classification given this study's goals. One way to assess the utility of the SCOP superfamily partitioning as a functional discriminator, is to look at a ratio of shared keywords associated with

proteins in the same superfamily, and between superfamilies. To avoid false negatives due to synonymous keywords, a controlled vocabulary, or ontology such as the Enzyme commission classification, the Gene Ontology (Ashburner et al., 2000) or MeSH has to be chosen. A second way to assess SCOP's utility is to use a distance measure between a chosen ontology's nodes (Joslyn et al., 2004; Lord et al., 2003a; Lord et al., 2003b). The second method is more accurate, as it takes into account the information conveyed by each node. For example, sharing the term "catalytic activity" is more likely to happen than sharing the term "shikimate kinase". Using the keyword sharing method, both sharing events receive the same weight. Using the Lord *et al* semantic similarity method, sharing "shikimate kinase" is considered more informative than sharing "catalytic activity" and therefore scores higher.

In order to see whether SCOP superfamily can be used as a proxy for function discrimination, we have done the following: first, we used GOA-PDB (Camon et al., 2004) to assign GO terms to SCOP chains. We then counted the number of identical GO terms between FFAS03 aligned proteins from different SCOP superfamilies, and from within SCOP superfamilies. Finally, we established the median semantic similarity of pairs between different superfamilies, and within a superfamily. The results are shown in Table 1. As can be seen, the percentage of shared GO terms is only 6% when looking at proteins residing in different superfamilies within the same fold. However, intra-superfamily comparison shows a 45% ratio of sharing terms. From these results it is clear that the superfamily level is a good choice as a proxy for functional dissimilarity. Proteins from different superfamilies have a low functional similarity, both by the percent identity criterion, and by the semantic similarity criterion. In contrast, proteins from within the same superfamily (Table 1, rows 2-3), even if from different families (Table 1 row 3) exhibit a high semantic similarity, and a high ratio of shared terms, so that functional discrimination on this level is not feasible.

GOA-PDB assigns GO terms to whole proteins, whereas the work shown here uses SCOP domains, two or more of which may make up a single protein. This may lead to errors in our analysis, as we may compare domains with GO terms that were falsely assigned to them: GO terms that truly belong to the neighboring domain on the same chain. Therefore, we performed the same analysis using only pairs of SCOP domains that come from single domain proteins. As shown in Table 1 rows 1-3 in *italics*, the results from the single domain analysis are comparable to those of the entire SCOP vs. SCOP, and provide the same conclusion with regards to the superfamily level being the best intra-fold functional discriminant.

A Method for SCOP superfamily-based target selection

Having established that SCOP superfamilies are a good marker for functions, we turn to construct a method which differentiates between members of the same and different SCOP superfamilies, using sequence information only. To construct this method we use a sensitive profile-profile based alignment tool, FFAS03. Figure 2 shows the distributions of FFAS03 scores from all-vs.-all comparison of SCOP domains for members of the same and different superfamilies. In all cases, if a FFAS03 score is better than 40, the two proteins being compared come from the same superfamily. For the FFAS03 scores worse than 40, the probability that both proteins come from a different superfamily increases as the FFAS03 score decreases, but the score by itself has a relatively low predictive value as many cases of different/the same superfamily pairs have similar scores.

To develop a superfamily/family classification system based on sequence information only, we used the C4.5 decision tree algorithm (Quinlan, 1993) as implemented in WEKA (Witten and Eibe, 2000). Using the FFAS03 alignment score as the sole attribute gave us poor initial results, as can be seen from the high overlap of the distributions in Figure 2. Interestingly, a related problem has presented itself recently to crystallographers performing molecular replacement (MR). Briefly, MR is a phasing

method using structural models based on a template that is homologous to the protein whose structure is being solved. With the advent of sensitive sequence detection methods, increasingly distant templates could be identified as a possible source of phasing information. A recent study set out to explore how divergent the template can be and still provide useful phasing information (Schwarzenbacher et al., 2004). It was shown that a good predictor for whether a distantly related homolog could be used for successful MR-based model refinement is the number and size of gaps in the alignment, rather than a score of the alignment. We therefore decided to add an affine gap score as an additional classification attribute. This means that after the alignment is performed we generate an attribute based on the number and size of alignment gaps. (See the Experimental Procedures section for the formal details). The results are shown in Figure 3 and in Table 2. Adding gap scores resulted in an area under the ROC curve (the so-called AUC) of 0.75. When using this classification scheme with folds which in SCOP are classified as having more than one superfamily, the AUC increased to 0.84, and for those having more than five superfamilies, to 0.85..

Using the decision tree, we proceeded to identify protein families that may form new superfamilies in known folds. Our first step was to perform an analysis of the results of an all-vs.-all PfamA–SCOP comparison using FFAS03. The first question was how many SCOP folds are covered by PfamA? Our results show that out of 7,166 PfamA v.11.0 families, (after the removal of families with extensive transmembrane and coiled-coil regions), 2,511 families can be matched to 543 SCOP folds above FFAS03's significance threshold. The coverage of is therefore 68%, 543 folds out of 800 in SCOP 1.65.

Predicted scope and functional enrichment

Having established a method for function based target selection, there is still the question of our method's scope. How many new targets does the function-driven target selection approach provide? How many folds can have the potential to be enriched with new functions? Most importantly, how much enrichment in functionality can we expect? "Functional enrichment" is the proportion of structurally uncharacterized functions that could be placed in a SCOP fold. This we expressed as the number of predicted superfamilies divided by the number of existing superfamilies in a fold.

To answer these questions, we ran the decision tree on all folds in SCOP. The results are shown in Table 3, and in the supplementary material <http://bioinformatics.burnham.org/~iddo/StructSup/fbts.html>. We have found that some 500 PfamA families are potential targets for crystallization based on our prediction of their inclusion in an existing fold, but having a different function. This number is a proportionally large one, making up some 10-15% of the overall number of targets the PSI has set to solve. Moreover, since sequence diversity is growing as we accumulate more genomic data (Friedberg, 2006), this number is expected to grow even more over time, as new versions of both Pfam and SCOP come out, and as the number of new folds in SCOP grow as a result of structural genomics efforts.

We see some 158 folds that can be potentially enriched by at least one superfamily. The most interesting folds are the 27 that are predicted to be enriched by five or more different superfamilies. Included among those are the three folds analyzed above. This brings us to the third question, which is the functional enrichment. Interestingly enough, some of the top enriched folds currently have only a single superfamily. Notable is c.37, P-loop hydrolase. This fold currently has only a single superfamily, but has many families, with many distinct enzymatic functions. It appears that many more still exist in sequence space. The same applies to S-adenosyl L methionine dependent methyltransferases fold (e.8), and to other single superfamily folds in the table. We are now conducting a full study describing the scope of functional enrichment of known structure space based on these preliminary findings.

Delving deeper into function space

For SCOP folds that initially have a high number of superfamilies, there is the possibility of obtaining even more specific results. This is done by training a decision tree for each individual fold.

- 1) First, train and test a decision tree using an all-vs.-all alignment of SCOP members in a given fold.
- 2) Next, use the decision tree to select all PfamA families that are aligned to that SCOP fold, but are predicted to be in a different superfamily using the decision tree previously generated.

We chose three folds with a high number of existing superfamilies: the TIM barrel fold, the immunoglobulin fold, and the flavodoxin-like fold. Figure 4 shows the ROC curves of the decision trees generated for each fold. Because in this case we used specific folds for the training, the decision tree performs much better than for the general fold populace. We found nine PfamA families as candidates for structurally uncharacterized functions within the TIM barrel fold, 18 in the immunoglobulin fold, and 6 in the flavodoxin-like fold. The full findings are listed in Tables 4–6.

Verifying targets using the SCOP–PDB lag, and different SCOP versions

A well-known technical hitch in the structural biology field has enabled us to independently verify our procedure. The SCOP database, being manually curated, lags a few months behind the PDB, which means that there are solved structures in the PDB that have not yet been classified by the SCOP curators. In this study, we have trained and tested our decision tree using SCOP 1.65. Several PfamA families we suggested as possible targets in our study had their structure already determined and published, but not processed by SCOP. Some of those structures have been already classified in a more recent version of SCOP, SCOP 1.67. Those entries are indicated in Tables 4–6. We will discuss in detail two entries in Table 4 from the TIM barrel fold, as those have proven to be the most informative

and interesting.

PF01680: PdxS is part of the PdxS/PdxT heteromeric complex, a heterotrimeric glutamine aminotransferase called pyridoxal 5'-phosphate (PLP) synthase. PLP is the biologically active form of vitamin B6. The PdxS moiety of the heteromer (UniProt: PDXS_BACSU) serves as the glutaminase domain, whereas PdxT (UniProt: PDXT_BACSU) abstracts the ammonia group. In their study, Zhu and colleagues (Zhu et al., 2005) have solved the structure of PdxS (PDB: 1ZNN) and have proposed a model for PdxS/PdxT binding. Because of the uniqueness of the glutamine binding site, and the putative PdxT binding sites in PdxS, this information could not have been revealed by homology modeling. However, the solved structure of PdxS was not in SCOP or PDB when we conducted our study, and PdxS (Pfam family PF01680) was selected as a good target for structural characterization of function.

PF04309: Glycerol operator operon antiterminator related. The protein TM01436 from *T. maritima* (UniProt: Q9X1F0_THEMA) was solved and made public in PDB on May 2004 (PDB: 1VKF). It was not classified in SCOP 1.65, the SCOP version this study uses. However, it was later inserted into SCOP 1.67 and classified in a superfamily of its own containing only this entry. Since this protein was crystallized as part of a structural genomics effort, we still do not know its exact function, although it is hypothesized to be a glycerol-3-phosphate responsive glycerol uptake antiterminator according to a BLAST search against NCBI's Conserved Domain Database (Marchler-Bauer et al., 2005). However, it is interesting to note our prediction that 1VKF belongs to a new superfamily agreed with that of SCOP's curators.

Discussion

The goals of this study are to draw attention to the need for a function-based target selection system, present a viable method for doing so, and provide a preliminary list of targets. The main difficulties with protein function classification are the subjective nature of the definition of a function (Bartlett et al., 2003; Hegyi and Gerstein, 1999; Whisstock and Lesk, 2003), and the representation of functional similarity or identity (Joslyn et al., 2004; Lord et al., 2003b; Shakhnovich, 2005). In this study we developed a two-way classification that answers the following question: Given that the fold of a protein sequence can be reliably predicted, would this protein have a new function not yet characterized within that fold? This question immediately begs the following methodological question: how best to know what is a “new” function? We addressed this problem by comparing the different levels of SCOP-based partitioning of proteins to the Gene Ontology annotations of those proteins. We show that the SCOP superfamily level best approximates the new function requisite. The results of this study show recognizing sequences are likely to have a different, uncharacterized function within a known fold is possible, and could be used in a practical selection of targets for structural genomics.

SCOP superfamily cutoff should not be treated as a rigorous threshold suitable for all cases. For example, the Rossman-like fold (c.2.-.-) is SCOP is clearly multifunctional, having many different enzymes, although it contains only a single superfamily. Furthermore, folds c.3 and c.4 that are described as “Rossman like” contain homologous proteins, the structure / function confusion in this case is too fine grained, reaching a family level. The high throughput method described in this study should be adjusted for this particular case. Conversely the P-loop hydrolase fold (c.37) that has only one superfamily, but has 23 families, many of which are functionally distinct. As shown in the results, our decision tree trained on all SCOP folds with more than one superfamily manages to overcome this

hurdle, and assign putative new functionally distinct sequences to this fold.

Naturally, a global automated approach such as we present here should be supplemented by a careful study of the potential new function candidates of each fold to ensure that they are indeed functionally dissimilar. This could be done by examining the identity of potential active site residues from the candidate sequences against the known active site residues in each of the superfamilies. One such study has recently been done for the PD-(D/E)XK nuclease superfamily, with the result that several new members have been identified (Feder and Bujnicki, 2005). Recently, another approach to function discrimination was proposed that identifies local amino-acid signatures as a functional discriminator (Wang and Samudrala, 2005). In its current form it relies on structural alignments and cannot be used for target selection from sequence databases. However, the authors suggest using sequence-to-structure alignments to extend their methods, and it should be interesting to compare it to the approach described here.

Functionally rich folds provide the best performing decision trees. However, we have also generated generic decision trees for all folds and for folds with a low (>1) and high (>5) number of superfamilies. These can be used when the superfamily has a low number of folds, albeit with lower sensitivity and recall. Nevertheless, it is easy to come up with a different proxy for function, such as the Gene Ontology (GO) (Ashburner et al., 2000) annotation or the Enzyme Commission (EC) classification and train a classification scheme based on any of those. At this time results would only be partial at best, since only about 50% of PfamA families are GO annotated and EC numbers apply only to enzymes. As shown in this study, for the purposes of structural genomics, the superfamily level serves well as a practical definition for function-based classification.

Most folds are deemed functionally poor, with few or a single superfamily. However, that may be due to technical reasons: we simply have not looked to experimentally solve structures in the “functionally poor” folds. Another is that certain folds are older and had time to accumulate more functions (e.g. (Coulson and Moulton, 2002; Wong and Frishman, 2006)). The method described here can help enrich the poorer folds that can be enriched, provide an estimate to the number of functions in the functionally rich folds and overall provide us with a better mapping of structure – function relationships. We provide a list of all folds that can be enriched by five superfamilies or more, and a list of 500 targets in the supplementary material.

A few of the targets we found do not have a known function. Those are listed in Tables 4–6 as domains of unknown function (DUF). Functionally uncharacterized, these proteins, if solved, will join the ranks of the 500 proteins currently in PDB listed as “hypothetical” or “unknown.” Obviously, solving the structure of those proteins would not provide us with immediate knowledge about their function. However, having those structures solved ensures that their function, once it becomes known, could be projected and understood at the molecular level.

The method described in this study can be combined with existing target selection and prioritization approaches such as Pfam5000 (Chandonia and Brenner, 2005) or the functional coverage of the human genome or pathogenic organism genomes (Abergel et al., 2003; Claverie et al., 2002; Goulding et al., 2003; Xie and Bourne, 2005) to determine the sampling rate required in those proteins that can be modeled. There is plenty of room for development in this field of function-based target selection in particular and function discrimination—as opposed to function prediction—in general. We are now enhancing our methods discriminative power by adding known functional sites to the features used by the decision tree. Using these tools, we aim to produce a finer map of the predicted functional content

of structure space: how many more functions for each protein fold is expected to hold. We have already begun such a mapping as shown in Table 3 and in the supplementary material, but we plan to refine it using different functional similarity thresholds. Using this path to explore the function–structure relationship can help reveal a better-quality picture of this critical issue in structural biology. Adding this rationale for target selection to those currently employed by structural genomics centers will facilitate a better understanding of the connection between protein structure and functionality.

Acknowledgements

This study was funded by NIH grant 1P01 – 63208. We thank Meytal Landau and the members of the Godzik Lab for stimulating discussions.

Experimental Procedures

Preparation of Database Sequences

SCOP 1.65 and PfamA 11.0 were used in this study. FFAS03 (Jaroszewski et al., 2005), a sensitive profile–profile alignment tool, was used for an all-*vs*-all alignment of protein sequences from both databases. SCOP sequences were clustered at 40% sequence identity using CD-HIT (Li et al., 2001) after we verified that that level of clustering does not lose any superfamily representatives (not shown).

From each PfamA family, a representative sequence most resembling the consensus of the multiple sequence alignment of that family was selected. This was done in the following fashion: a position-specific scoring matrix (PSSM) was constructed for the PfamA sequences. The distance of this PSSM was measured to each of the constructing sequences, and the sequence with the minimal distance was chosen. FFAS03 was used for the sequence–PSSM alignments.

Prior to alignment, the following types of protein sequences were filtered out from PfamA and SCOP 1.65: (1) sequences shorter than 80aa; and (2) sequences containing coiled-coil regions, or transmembrane regions (as predicted using TMHMM (Krogh et al., 2001)). These tend to have nonspecific signals and may associate otherwise dissimilar proteins together.

Following the pairwise alignment of PfamA vs. SCOP, or SCOP vs. SCOP, we removed those alignments with FFAS03 scores lower than 9.5 (p-value of 0.02) and whose members had a length

difference of more than 50% of the shorter sequence's length, or an alignment length of less than 60% of the longer member's length. Thus the alignments were sure to span a substantial length of both members, denoting a domain-length structural similarity.

Alignments

All SCOP sequences were aligned to each other using FFAS03. The alignment FFAS03 scores and affine gap penalties were recorded. To generate the affine gap score attribute for the classification, affine gap scores were extracted from the alignments as follows: 5 for gap insertion, and 2 for gap extension. It should be noted that the gap scores are not those that were used to generate the alignments; rather, they were extracted for the alignment as attributes for classification. Other gap scores were tested and showed no substantial difference in performance over a representative sample of the database (results not shown).

Classification

All-vs.-all SCOP alignments were classified using the C4.5 algorithm for generating classification rules in the form of a decision tree.

To generate a decision tree for SCOP alignments, the following steps were taken:

- 1) Select the corpus of interest: i.e., all-vs.-all alignments of sequences in a given single SCOP fold, or a collection of folds. In any case, the aligned sequences are all from the same fold, but from both different and same superfamilies.

2) Select attributes: i.e., the FFAS03 score and the affine gap scores extracted from the alignment, as described above.

3) Use the WEKA 3.4 implementation of C4.5 decision tree builder to generate a set of hierarchical rules using the corpus of interest as a training set: i.e., 10× stratified cross validation of the training set as a test set. 10x stratified cross-validation means that the data is partitioned randomly into ten sets. Nine are used for training, and one for testing. This procedure is repeated ten times with each set held out in turn. The stratification ensures that each set contains a representative number of testing and training data.

The resulting decision tree is applied to PfamA-vs.-SCOP pairwise alignments to find the PfamA families that belong to a given SCOP fold but are not in any of the known superfamilies. Each PfamA family (as represented by a representative sequence) is also aligned with all other SCOP sequences in the database. This is because a given Pfam sequence may be predicted to be a same-fold-not-same-superfamily within one superfamily, but may be well aligned to another SCOP superfamily, which means that it already has a known function.

Classification Assessment

The following characteristics were used to assess the performance of the classification.

1) Area under ROC: The receiver operator characteristics (ROC) curve describes the capability of the classification scheme to maximize true positives (y-axis) and minimize false positives (x-axis). A straight line of $x = y$ signifies a classification method no better than random, and the area under the curve (AUC) is 0.5. The more a classifier's ROC curve gets pushed up and to the left, the better its

performance. A zero errors classifier would have a ROC curve with an AUC of 1.0. The ROC curve's y-axis is the true positive rate, sometimes called recall, and is calculated as follows: $TP/(TP + FN)$. The x-axis is the false positive rate and is calculated as follows: $FP/(FP + TN)$.

2) Recall: The true positive rate, calculated as follows: $TP/(TP + FN)$

3) Precision, calculated as follows: $TP/(TP + FP)$

where TP = true positives, FN = false negatives, and FP = false positives

Gene Ontology based observations

For the inter- and intra- superfamily observations we looked at terms shared by different superfamilies in the same fold, and between superfamilies. If two chains share more than a single GO, we only counted one shared term. Another way of measuring distance between terms in Gene Ontology is to use semantic similarity as described in Lord et al (2003b). Briefly, each GO term is tagged with its frequency in an annotated protein corpus. The frequency of each term is its own frequency + the sum of the frequency of its children terms. The semantic similarity between two terms is given as:

$$p_{ms}(c1, c2) = \min_{c \in S(c1, c2)} \{p(c)\}$$

$$sim(c1, c2) = -\log(p_{ms}(c1, c2))$$

Where c1, c2 are GO terms, $S(c1, c2)$ is the set of parent terms of c1 and c2. $p(c)$ is the frequency of

term c in the corpus, and $pms(c1, c2)$ is the frequency of the minimal subsuming term of terms $c1, c2$. The data corpus in this case was SCOP clustered at 40% sequence similarity, to remove bias and redundancies. If any two chains shared more than a single GO term, the pairwise similarity between the chains was taken to be the maximal similarity. The final score given in Table 1 is the median score of the maximum for all pairs, and the standard error from the median.

Tables:

Table 1: Functional kinship within and between SCOP superfamilies

Row 1: Semantic similarity and shared GO terms between pairs of proteins from different superfamilies, in the same fold. A pair was included if its members could be significantly aligned by FFAS03.

Row 2: Semantic similarity and shared GO terms for pairs of proteins that are in the same SCOP superfamily. This includes proteins from different superfamilies that have a significant sequence similarity by FFAS03.

Row 3: Semantic similarity distances and shared GO terms between pairs of proteins in the same SCOP superfamily, but excluding pairs from different SCOP families.

Row 4: Semantic similarity distances and shared GO terms between pairs of proteins in the same SCOP family.

In Rows 1-4, single domain protein comparison numbers are given in *italics*. See text regarding single domain protein pairs analysis.

Fifth row: Semantic similarity and shared GO terms between Pfam families that are predicted to be in existing SCOP superfamilies.

Sixth row: Semantic similarity and shared GO terms between Pfam families that are predicted to be in existing SCOP folds, but be in new superfamilies, unclassified by SCOP

Column 2: protein pairs having identical terms. Column 3: total number of pairs (In parentheses: number of pairs in which both members are GO annotated, if differs from total number of pairs) [In brackets: PfamA families whose function is unknown]. Column 4: the ratio between 2 and 3. As can be clearly seen there are considerably fewer protein pairs sharing identical terms between superfamilies, than in superfamilies. Column 5: median semantic similarity and standard error from the median. Standard error was calculated as:

$$\frac{\sigma'}{\sqrt{n}}$$

n being the number of items, σ' being standard deviation from the median

	1. SCOP level of comparison	2. Protein pairs with identical terms	3. Total pairs	4. Ratio	5. Median GO similarity / Standard error of median
1	Inter-superfamily	29	448	0.06	231 / 10.9

		<i>14</i>	<i>173</i>	<i>0.08</i>	<i>173 / 15.7</i>
2	Intra superfamily, including inter-family pairs	5279 <i>2439</i>	11643 <i>4681</i>	0.45 <i>0.52</i>	980 / 2.58 <i>968 / 3.4</i>
3	Intra-superfamily, no inter-family pairs	2012 968	5698 <i>2312</i>	0.35 <i>0.42</i>	774 / 4.0 <i>980 / 5.43</i>
4	Intra family	4732 <i>2535</i>	7613 <i>3815</i>	0.62 <i>0.66</i>	980/ 3.2 <i>980 / 2.9</i>
5	PFAM, in existing SCOP folds, in existing superfamilies	474	1202 (513) [4]	0.39	851 / 11.6
6	PFAM, in existing SCOP folds, in new superfamilies	44	614 (133) [22]	0.07	371/ 9.4

Table 2

Performance of the C4.5 decision tree for prediction of superfamilies

The first column shows the number of SCOP superfamilies per fold. There are 26 folds with more than five superfamilies and 65 with more than one superfamily. As the number of superfamilies increases, the decision tree constructed performs better in terms of area under the curve, precision, and recall.

NUMBER OF SUPERFAMILIES	AREA UNDER CURVE (AUC)	PRECISION	RECALL	NUMBER OF FOLDS
>5	0.85	0.77	0.85	26
>1	0.84	0.76	0.84	65
>0	0.76	0.65	0.84	700

Table 3: Top Functionally Enriched Folds

SCOP FOLD	PREDICTED NEW SUPERFAMILIES (FUNCTIONS)	EXISTING NUMBER OF SUPERFAMILIES	ENRICHMENT FRACTION
b.1	21	20	1.05
c.37	19	1	19
c.55	15	7	2.14286
a.118	14	17	0.823529
a.4	11	12	0.916667
c.1	11	26	0.423077
f.4	11	4	2.75
c.47	10	2	5
b.82	9	7	1.28571
c.68	9	1	9
d.92	9	2	4.5
e.8	9	1	9
c.66	8	1	8
a.2	7	11	0.636364
d.58	7	48	0.145833
a.63	6	1	6
b.68	6	8	0.75
c.23	6	16	0.375
d.108	6	1	6
a.138	5	1	5
a.24	5	16	0.3125
a.25	5	2	2.5
b.18	5	1	5
c.108	5	1	5
d.15	5	11	0.454545
d.2	5	1	5
f.13	5	1	5

Table 4

Pfam families predicted to be TIM Barrels with structurally uncharacterized functions

We located nine Pfam families that have different functions than those currently in TIM barrels, two of them with unknown functions.

	Pfam ID	Family
1	PF04413	kdotransferase
2	PF01680	SOR/SNZ
3	PF01136	Peptidase family U32
4	PF04476	DUF556
5	PF03644	Glycosyl hydrolase family 85
6	PF03490	Variant-surface-glycoprotein phospholipase C
7	PF04309	Glycerol-3-phosphate responsive antiterminator*
8	PF04273	DUF442*
9	PF05114	DUF692

Table 5

Pfam families predicted to be in the immunoglobulin fold, with structurally uncharacterized functions

Eighteen Pfam families that are predicted to be in the immunoglobulin fold but have different functions than those already solved (two unknown)

	Pfam ID	Family
1	PF06155	DUF971
2	PF02494	HYR domain
3	PF02480	Alphaherpesvirus glycoprotein E
4	PF01688	Alphaherpesvirus glycoprotein I
5	PF06312	Neurexophilin
6	PF06011	Transient receptor potential (TRP) ion channel
7	PF05753	Translocon-associated protein beta (TRAPB)
8	PF05751	FixH
9	PF01835	Alpha-2-macroglobulin family N-terminal region
10	PF03351	DOMON domain
11	PF07427	DUF1511 in PDB not in SCOP, still unknown
12	PF06159	DUF974
13	PF07036	Starch synthase III
14	PF03168	Late embryogenesis abundant protein
15	PF07354	Reovirus sigma C capsid protein

16	PF03896	Translocon-associated protein (TRAP), alpha subunit
17	PF05566	Orthopoxvirus interleukin 18 binding protein
18	PF05506	DUF756

Table 6

Pfam families predicted to be in the flavodoxin-like fold (SCOP c.23), with structurally uncharacterized function

	Pfam ID	Family
1	PF04914	DltD C-terminal region
2	PF06259	DUF 1023
3	PF04204	Homoserine O-succinyltransferase
4	PF03861	ANTAR: bacterial RNA binding domain
5	PF07090	DUF 1355
6	PF00657	GDSL-like lipase/acylhydrolase

FIGURES

Figure legends

Figure 1

Flowchart illustrating the rationale for function-driven target selection

For each protein in a defined sequence space, determine whether the structure can be reliably predicted from the sequence. If the structure can be reliably predicted, but the function cannot be predicted to be the same as that of an already solved structure, then the protein is a target of interest. This complements the “new fold” strategy. The strategy discussed in this study is delineated by the dashed line.

Figure 2

Distribution of FFAS03 scores for pairwise alignments in of sequences from SCOP .

Histogram of the distribution of FFAS03 scores from two populations of protein sequences which were pairwise aligned, all-vs. all within each population. The white bars histogram shows the distribution of scores from alignments of proteins taken from different superfamilies, but in the same fold. The black bars histogram shows the distribution of scores from pairwise alignments of proteins in the same superfamilies.

Figure 3

ROC curves for all superfamilies

The performance of C4.5 on recognition of same and different superfamilies using all vs. all on sequences taken from all folds (diamonds); sequences taken from folds with more than one superfamily (crosses); and sequences taken from folds with more than five superfamilies (squares). AUC, precision, recall, and population size are all given in Table 2.

Figure 4

ROC curves for selected folds

TIM barrel fold (SCOP c.1.-.-); diamonds, immunoglobulin fold (b.1.-.-); squares, flavodoxin-like (c.23.-.-) crosses.

References

- Abergel, C., Coutard, B., Byrne, D., Chenivresse, S., Claude, J. B., Deregnacourt, C., Fricaux, T., Giancesini-Boutreux, C., Jeudy, S., Lebrun, R., *et al.* (2003). Structural genomics of highly conserved microbial genes of unknown function in search of new antibacterial targets. *J Struct Funct Genomics* 4, 141-157.
- Aloy, P., Oliva, B., Querol, E., Aviles, F. X., and Russell, R. B. (2002). Structural similarity to link sequence space: new potential superfamilies and implications for structural genomics. *Protein Sci* 11, 1101-1116.
- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., *et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* 25, 25-29.
- Bartlett, G. J., Todd, A. E., and Thornton, J. M. (2003). Inferring protein function from structure. *Methods Biochem Anal* 44, 387-407.
- Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etmiller, L., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M., and Sonnhammer, E. L. (2002). The Pfam protein families database. *Nucleic Acids Res* 30, 276-280.
- Brenner, S. E. (2000). Target selection for structural genomics. *Nat Struct Biol* 7 *Suppl*, 967-969.
- Camon, E., Barrell, D., Lee, V., Dimmer, E., and Apweiler, R. (2004). The Gene Ontology Annotation (GOA) Database--an integrated resource of GO annotations to the UniProt Knowledgebase. *In Silico Biol* 4, 5-6.
- Chandonia, J. M., and Brenner, S. E. (2005). Implications of structural genomics target selection strategies: Pfam5000, whole genome, and random approaches. *Proteins* 58, 166-179.
- Claverie, J. M., Monchois, V., Audic, S., Poirot, O., and Abergel, C. (2002). In Search of new anti-bacterial target genes: a comparative/structural genomics approach. *Comb Chem High Throughput Screen* 5, 511-522.
- Cort, J., Koonin, E., Bash, P., and Kennedy, M. (1999). A phylogenetic approach to target selection for structural genomics: solution structure of YciH. *Nucl Acids Res* 27, 4018-4027.
- Coulson, A. F., and Moulton, J. (2002). A unfold, mesofold, and superfold model of protein fold use. *Proteins* 46, 61-71.
- Devos, D., and Valencia, A. (2000). Practical limits of function prediction. *Proteins* 41, 98-107.
- Feder, M., and Bujnicki, J. M. (2005). Identification of a new family of putative PD-(D/E)XK nucleases with unusual phylogenomic distribution and a new type of the active site. *BMC Genomics* 6, 21.
- Friedberg, I. (2006). Automated protein function prediction--the genomic challenge. *Brief Bioinform* 7, 225-242.
- Friedberg, I., Jaroszewski, L., Ye, Y., and Godzik, A. (2004). The interplay of fold recognition and experimental structure determination in structural genomics. *Curr Opin Struct Biol* 14, 307-312.
- Galperin, M. Y., Walker, D. R., and Koonin, E. V. (1998). Analogous enzymes: independent inventions in enzyme evolution. *Genome Res* 8, 779-790.
- Gerlt, J. A., and Babbitt, P. C. (2001). Divergent evolution of enzymatic function: mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annu Rev Biochem* 70, 209-246.

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.

Goulding, C. W., Perry, L. J., Anderson, D., Sawaya, M. R., Cascio, D., Apostol, M. I., Chan, S., Parseghian, A., Wang, S. S., Wu, Y., *et al.* (2003). Structural genomics of *Mycobacterium tuberculosis*: a preliminary report of progress at UCLA. *Biophys Chem* 105, 361-370.

Hegyí, H., and Gerstein, M. (1999). The relationship between protein structure and function: a comprehensive survey with application to the yeast genome. *J Mol Biol* 288, 147-164.

Hegyí, H., and Gerstein, M. (2001). Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Res* 11, 1632-1640.

Jaroszewski, L., Rychlewski, L., Li, Z., Li, W., and Godzik, A. (2005). FFAS03: a server for profile-profile sequence alignments. *Nucleic Acids Res* 33, W284-288.

Jinfeng Liu, H. H., Thomas B. Acton, Gaetano T. Montelione, Burkhard Rost, (2004). Automatic target selection for structural genomics on eukaryotes. *Proteins: Structure, Function, and Bioinformatics* 56, 188-200.

Joslyn, C. A., Mniszewski, S. M., Fulmer, A., and Heaton, G. (2004). The gene ontology categorizer. *Bioinformatics* 20 *Suppl 1*, I169-I177.

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J Mol Biol* 305, 567-580.

Li, W., Jaroszewski, L., and Godzik, A. (2001). Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics* 17, 282-283.

Linial, M., and Yona, G. (2000). Methodologies for target selection in structural genomics. *Progress in Biophysics and Molecular Biology* 73, 297-320.

Liu, J., and Rost, B. (2002). Target space for structural genomics revisited. *Bioinformatics* 18, 922-933.

Lord, P. W., Stevens, R. D., Brass, A., and Goble, C. A. (2003a). Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics* 19, 1275-1283.

Lord, P. W., Stevens, R. D., Brass, A., and Goble, C. A. (2003b). Semantic similarity measures as tools for exploring the gene ontology. *Pac Symp Biocomput*, 601-612.

Marchler-Bauer, A., Anderson, J. B., Cherukuri, P. F., DeWeese-Scott, C., Geer, L. Y., Gwadz, M., He, S., Hurwitz, D. I., Jackson, J. D., Ke, Z., *et al.* (2005). CDD: a Conserved Domain Database for protein classification. *Nucl Acids Res* 33, D192-196.

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.

Nagano, N., Orengo, C. A., and Thornton, J. M. (2002). One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. *J Mol Biol* 321, 741-765.

Norvell, J. C., and Machalek, A. Z. (2000). Structural genomics programs at the US National Institute of General Medical Sciences. *Nat Struct Biol* 7 *Suppl*, 931.

Pandit, S. B., Bhadra, R., Gowri, V. S., Balaji, S., Anand, B., and Srinivasan, N. (2004). SUPFAM: a database of sequence superfamilies of protein domains. *BMC Bioinformatics* 5, 28.

Pawlowski, K., Jaroszewski, L., Rychlewski, L., and Godzik, A. (2000). Sensitive sequence comparison as protein function predictor. *Pac Symp Biocomput*, 42-53.

Quinlan, J. R. (1993). C4.5: Programs for Machine Learning, Morgan Kaufmann).

- Rost, B. (2002). Enzyme function less conserved than anticipated. *J Mol Biol* 318, 595-608.
- Schwarzenbacher, R., Godzik, A., Grzechnik, S. K., and Jaroszewski, L. (2004). The importance of alignment accuracy for molecular replacement. *Acta Crystallogr D Biol Crystallogr* 60, 1229-1236.
- Shah, I., and Hunter, L. (1997). Predicting enzyme function from sequence: a systematic appraisal. *Proc Int Conf Intell Syst Mol Biol* 5, 276-283.
- Shah, I., and Hunter, L. (1998). Identification of divergent functions in homologous proteins by induction over conserved modules. *Proc Int Conf Intell Syst Mol Biol* 6, 157-164.
- Shakhnovich, B. E. (2005). Improving the precision of the structure-function relationship by considering phylogenetic context. *PLoS Comput Biol* 1, e9.
- Stevens, R. C., Yokoyama, S., and Wilson, I. A. (2001). Global Efforts in Structural Genomics. *Science* 294, 89-92.
- Thornton, J. M., Orengo, C. A., Todd, A. E., and Pearl, F. M. (1999). Protein folds, functions and evolution. *J Mol Biol* 293, 333-342.
- Tian, W., and Skolnick, J. (2003). How well is enzyme function conserved as a function of pairwise sequence identity? *J Mol Biol* 333, 863-882.
- Todd, A. E., Orengo, C. A., and Thornton, J. M. (2001). Evolution of function in protein superfamilies, from a structural perspective. *J Mol Biol* 307, 1113-1143.
- Todd, A. E., Orengo, C. A., and Thornton, J. M. (2002). Sequence and structural differences between enzyme and nonenzyme homologs. *Structure* 10, 1435-1451.
- Wang, K., and Samudrala, R. (2005). FSSA: a novel method for identifying functional signatures from structural alignments. *Bioinformatics* 21, 2969-2977.
- Whisstock, J. C., and Lesk, A. M. (2003). Prediction of protein function from protein sequence and structure. *Q Rev Biophys* 36, 307-340.
- Witten, I., and Eibe, F. (2000). *Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations*, First edn (San Diego, CA USA, Morgan Kaufmann).
- Wong, P., and Frishman, D. (2006). Fold designability, distribution, and disease. *PLoS Comput Biol* 2, e40.
- Xie, L., and Bourne, P. E. (2005). Functional Coverage of the Human Genome by Existing Structures, Structural Genomics Targets, and Homology Models. *PLoS Comput Biol* 1, e31.
- Zhu, J., Burgner, J. W., Harms, E., Belitsky, B. R., and Smith, J. L. (2005). A new arrangement of (beta/alpha)₈ barrels in the synthase subunit of PLP synthase. *J Biol Chem* 280, 27914-27923.

Figure 1

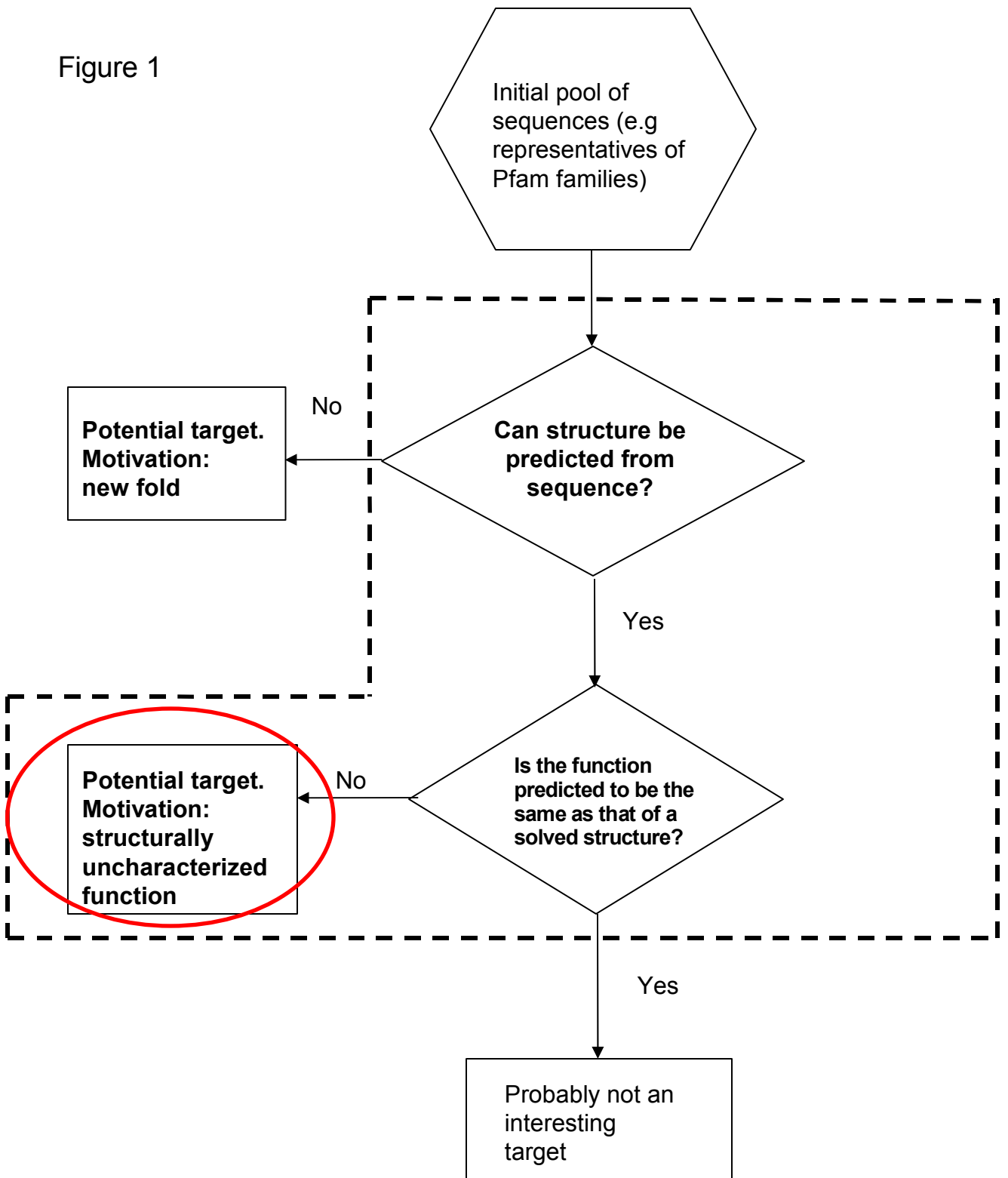


Figure 2

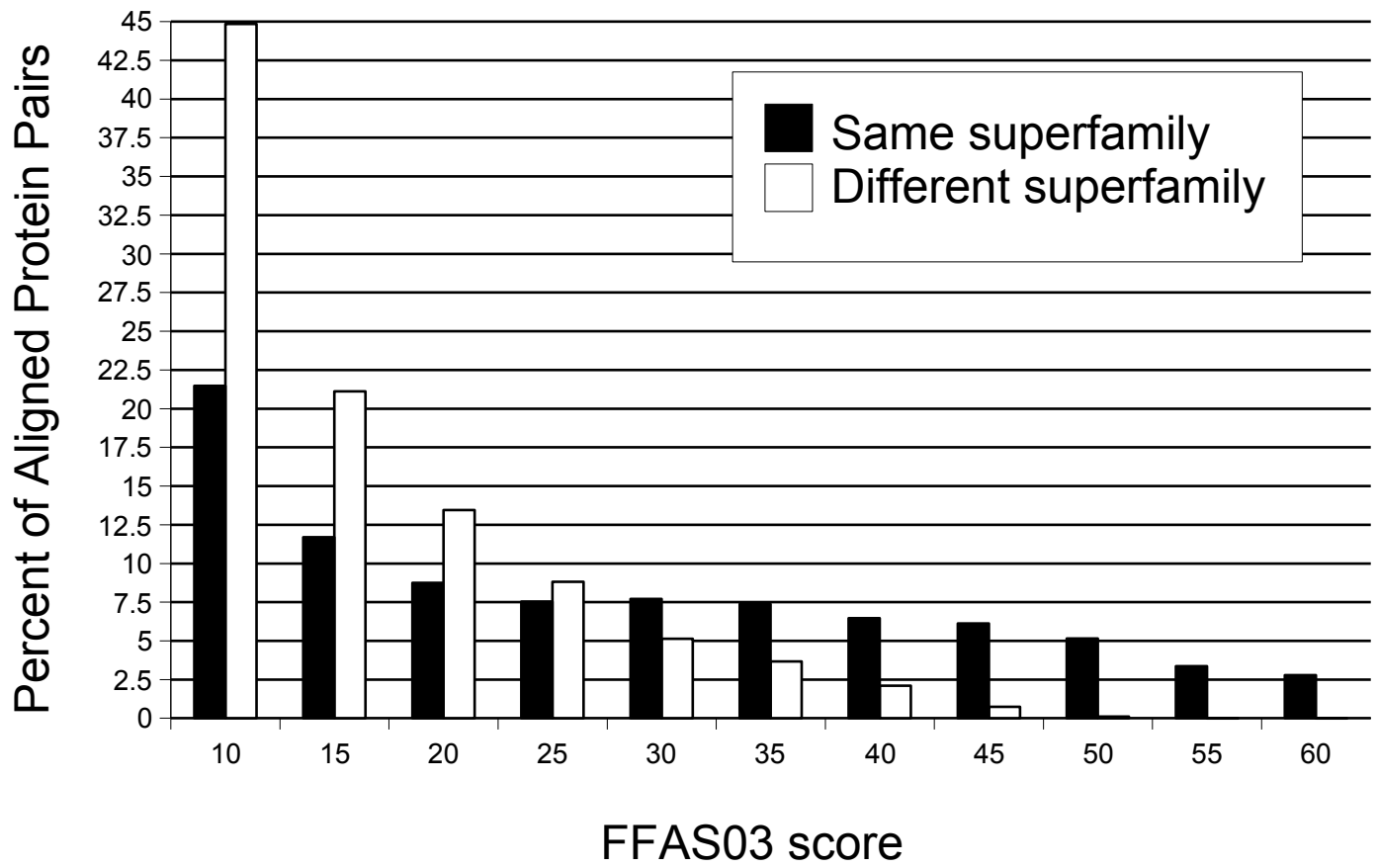


Figure 3

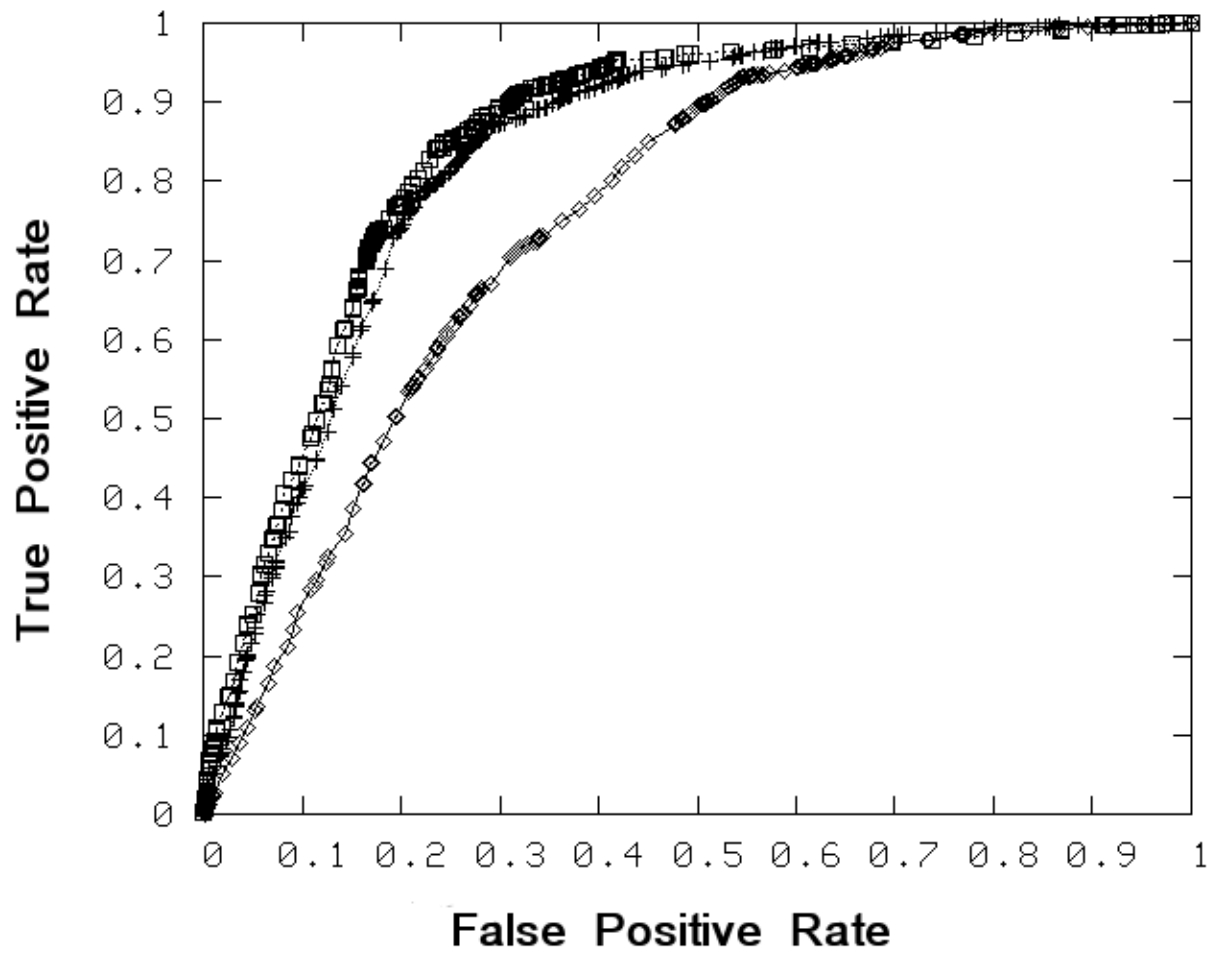


Figure 4

