Structure-based function inference using protein family-specific fingerprints

Special Section on Automated Function Prediction

Deepak Bandyopadhyay*†, Jun Huan*, Jinze Liu*, Jan Prins*, Jack Snoeyink*, Wei Wang* and Alexander Tropsha‡

February 11, 2006

Abstract

We describe a method to assign a protein structure to a functional family using *family-specific fingerprints*. Fingerprints represent amino acid packing patterns that occur in most members of a family but rare in the *background*, a non-redundant subset of PDB; their information is additional to sequence alignments, sequence patterns, structural superposition and active site templates.

Fingerprints were derived for 120 families in SCOP using Frequent Subgraph Mining. For a new structure, all occurrences of these family-specific fingerprints may be found by a fast algorithm for subgraph isomorphism; the structure can then be assigned to a family with a confidence value derived from the number of fingerprints found and their distribution in background proteins.

In validation experiments, we infer the function of new members added to SCOP families, and we discriminate between structurally similar, but functionally divergent, TIM barrel families. We then apply our method to predict function for several structural genomics proteins, including orphan structures. Some predictions have been corroborated by other computational methods, and some validated by subsequent functional characterization.

Keywords: subgraph mining, Delaunay, almost-Delaunay, protein classification, structure-based function inference, structural genomics, orphan structures

*Department of Computer Science, University of North Carolina at Chapel Hill, E-mail:{debug, huan, liuj, prins, snoeyink, weiwang}@cs.unc.edu

†Current address: Johnson & Johnson Pharmaceutical Research and Development, 665 Stockton Drive, Exton, PA

‡Laboratory of Molecular Modeling, School of Pharmacy, University of North Carolina at Chapel Hill, E-mail: tropsha@email.unc.edu

Introduction

Structural genomics projects (Burley, 2000) have generated structures for many proteins of unknown function. Function for these so-called *hypothetical proteins* is traditionally inferred from sequence similarity or overall structure similarity. Structural genomics targets, however, are selected to avoid sequence similarity so as to sample the protein "structure space:" a quarter of structural genomics proteins deposited by May 2005 had less than 30% sequence identity and DALI Z-scores (Holm & Sander, 1996) less than 10 with proteins of known function (Bandyopadhyay, 2005). Other inference tools are needed for these *orphan* protein structures.

Recently, methods have been developed to infer function from local structural similarity, without relying on sequence and overall structure similarity. Aloy *et al.* (2001) found that conserved geometric packing patterns of a few residues are often responsible for protein function, and finding them can lead to more accurate function inference than obtained by structural homology. Laskowski *et al.* (2005b) developed SiteSeer's reverse template method, which also searches for conserved packing patterns within protein structures. Other recent methods find functionally important residues using computed chemical properties (Ko *et al.*, 2005), careful alignments (Pegg *et al.*, 2005), evolutionary information (Wang & Samudrala, 2005), and

computational protein design (Cheng *et al.*, 2005). Still other methods use Gene Ontology (Gene Ontology Consortium, 2004) as a reference to define function, such as ProKnow (Pal & Eisenberg, 2005) and PHUNCTIONER (Pazos & Sternberg, 2004). A recent review (Ofran *et al.*, 2005) covers these and other structure-based function prediction methods.

Graph representations of protein structure allow more flexibility than rigid templates in representing and matching structural motifs. Earlier methods used graph representations to search for known structure patterns (Artymiuk *et al.*, 1994; Stark & Russell, 2003), or determine patterns with limited topology, such as cliques, from groups of proteins (Wangikar *et al.*, 2003; Milik *et al.*, 2003). Using frequent subgraph mining, Huan *et al.* (2004, 2005) defined *family-specific fingerprints* as those packing patterns that are frequent within a family of protein structures but rare within the background. Using serine protease and kinase families, they showed that fingerprints often cover functionally important residues and can distinguish between proteins from similar families.

In this paper, we propose a new method for function inference that uses family-specific fingerprints automatically derived from SCOP families (Murzin *et al.*, 1995). The method searches for fingerprints within a new structure using fast subgraph isomorphism (Ullman, 1976), and assigns a significance score to family membership using the distribution of fingerprints found in members of the family and in the background. Its strength is in distinguishing between proteins with related and similar functions.

Our method does not restrict pattern graph types, or assume the functional sites are known. Each fingerprint is statistically linked to its family, and our consensus approach using multiple fingerprints improves the accuracy and specificity of function inference. Families with different function but similar structure can be distinguished, since the fingerprints tend to identify functionally important parts of a protein. In contrast, methods based on Gene Ontology suggest broader functional categories more than specific functional families (Pazos & Sternberg, 2004; Pal & Eisenberg, 2005).

Results

We derived family-specific fingerprints for proteins in 120 SCOP families using a *background* of 6,749 non-redundant proteins, as described in the Methods section. After this, we examined the family specificity of the fingerprints, then classified new protein structures by identifying cases of functional similarity with and without overall structure similarity, and inferred function for orphan structures from structural genomics targets.

Fingerprint occurrence in family and background: To test the uniqueness of a family's fingerprints, and establish significance of function inference, we examined the frequency of family-specific fingerprints in the background, as described in Materials and Methods. In most families, almost all background proteins have fewer fingerprints than the minimum found in any family member; see Figure 1(a)–(b) for examples of the metallo-dependent hydrolase (SCOP: 51556) and antibiotic resistance (SCOP:54598) families. Some family members have few fingerprints; the majority of those we inspected had either a different function or mechanism from the other members, or errors in the structure file that prevent the identification of fingerprints.

Many background proteins with a majority of the fingerprints for a family turned out to be new family members. For example, four proteins with 30 or more of the 49 metallo-dependent hydrolase fingerprints, 1un7A (48), 1rk6A (40), 1ndyA (33) and 1kcxA (32), were not included in the metallo-dependent hydrolase family in SCOP 1.65, but were in SCOP 1.67. Other

high-scoring proteins had closely related enzymatic functions (e.g. phosphatases, phosphoesterases) but came from different SCOP families, such as metallohydrolase/oxidoreductase of TIM barrel fold (1p9e, 48) and mannose hydrolase of $(\beta\alpha)_7$ fold (1qwn, 44).

Validation on proteins added to SCOP: To test the validity of inferring family membership, we used fingerprints derived from SCOP 1.65 families to classify proteins that were newly added to these families in SCOP 1.67. The detailed results are shown in Tables II– IV in the online supplementary material. Of the 442 new members added to 94 families, the number of proteins that can be inferred using fingerprints from the correct family is 316 (71%) at the sensitivity cutoff, and 284 (64%) at the 99%-specificity cutoff. Most importantly, for 287 (65%) of the new members, among families with fingerprints above 95% specificity, the correct family was the choice with highest specificity. By contrast, for only 234 (53%) of the new members did a member of the correct family have the most significant sequence hit, among all proteins in SCOP 1.65 with at least 40% sequence identity, which is the threshold suggested for inferring function from sequence (Wilson *et al.*, 2000).

Discriminating between similar structures with different function: To test the discrimination power of fingerprints, we searched for the fingerprints of 20 structurally similar (super)families of the TIM barrel fold that have different functions. As shown in Figure 2, the average member of any of these families has 70–90% of the fingerprints of its own family (orange or red, seen on the diagonal), and 0–40% of the fingerprints of any other family (blue, seen off the diagonal). Exceptions arise from superfamily-subfamily pairs such as enolase C-terminal domains(ENC) and D-glucarate dehydratases(DGL) that share fingerprints since their members overlap, and from families that do not have highly significant fingerprints, such as the ribulose-phosphate-binding barrels(RIB). Thus, fingerprints discriminate between functional families whose members cannot be distinguished easily by overall structure similarity.

Function inference for structural genomics targets: We classified Structural Genomics targets in the PDB as either proteins with known function, proteins with putative function suggested by overall structure similarity, and the *orphan* structures. We applied our method to suggest function assignments for proteins in the last two categories. For example, strong structural similarity to the metallo-dependent phosphatase superfamily (SCOP: 56300) was found in two hypothetical proteins, 1s3l (14% sequence identity, DALI z-score 13.1 with member 1hpu) and 1xm7 (13% identity, z-score 10.6, 1ii7). For these proteins we inferred metallo-dependent phosphatase function with 26 and 125 out of 316 fingerprints, i.e. 100% specificity, corroborating the function inference suggested by structural similarity. More interesting are two case studies for proteins in the last category, i.e. structural orphans.

Functional Inference of YcdX The YcdX protein (PDB: 1m65, CASP5 target T0147) has a rare $(\beta\alpha)_7$ barrel fold called the PHP domain (SCOP: 89551). It had no significant sequence or overall structure similarity with proteins of known function in 2004. We inferred that this protein has a metallo-dependent hydrolase function, with 30 of 49 fingerprints from SCOP superfamily 51556, a TIM barrel family. The fingerprints are shown as subgraphs in Figure 1(c)–(d). The residues included in family-specific fingerprints for this target, depicted in Figure 1(e)–(f), are localized in space and show similar geometric arrangements and chemical properties in family and target.

Our inference was corroborated by (1) active site template and reverse template matches on the ProFunc server Laskowski (2005a,2005b), (2) suggestions by the CASP5 target classifiers (Kinch *et al.*, 2003), and (3) suggestions by the authors of the structure (Teplyakov *et al.*, 2003),

who proposed active site residues for 1m65 that are included in many of our fingerprints, as shown in the supplementary material. The PINTS-weekly service (Stark *et al.*, 2004) found active site patterns from many metallo-dependent hydrolases in this protein. Finally, GenProtEC, the E.Coli genome and proteome database (Serres *et al.*, 2004) has annotated the YcdX gene product as belonging to the SCOP metallo-dependent hydrolase structural domain family, on the basis of the SUPERFAMILY database of HMMs for SCOP families (Gough & Chothia, 2002; Madera *et al.*, 2004).

Functional Inference for Protein Yyce Protein Yyce from Bacillus Subtilis (PDB: 1twu) is unclassified in both SCOP 1.65 and 1.67, and was an orphan structure in 2004, with no significant structural similarity to structures of known function. We found 46 of 62 fingerprints from the antibiotic resistance protein family (SCOP ID: 54598) in 1twu, inferring the antibiotic resistance function with 100% specificity. Figure 1(g)–(h) show the residues covered by fingerprints in 1twu and in 1ecs, an antibiotic resistance protein in SCOP 1.65. Note the geometric and electrostatic similarity between the upper region covered by fingerprints in both 1twu and 1ecs, which suggests that fingerprints cover functionally important residues. When the structural similarity of 1twu was re-evaluated in May 2005 using the current DALI database, it was found to be similar to a protein 1nki that was unclassified in SCOP 1.65 but has been added to the antibiotic resistance protein family in SCOP 1.67. This discovery of homology to a newly classified member of the family corroborates our function inference.

Discussion

Our method of using family-specific fingerprints to infer function for proteins was designed to be robust: the graph construction takes into account natural imprecision in coordinates, and using multiple local motifs as fingerprints accommodates remaining representation errors and flexibility in functional sites. The method is also designed to give information that is not implied by sequence patterns, structural alignments, and templates of known functional sites. Thus, not only may it succeed as a standalone method where other methods may fail, but it may also be profitably used in consensus with other methods.

The successful function inference for new members of SCOP families validates the predictive power of fingerprints; the success rate of 65% for choosing the correct family is high considering that there are functional outliers among existing and new members of SCOP families, and considering that sequence methods could pick the correct family only 53% of the time.

The function discrimination within the TIM barrel fold, and the inference of YcdX as belonging to the sequence-diverse metallo-dependent hydrolase family despite its different fold, indicate that the packing patterns in fingerprints do capture information that is specific to a functional family, rather than shared structural information.

We have seen that the fingerprints detected in YcdX cover its functional regions; this can be attributed to the fact that SCOP families often share a function, and superfamilies often share aspects of function. Our subgraph mining finds fingerprints that characterize the shared local structures exclusive to each family. Our method can also derive fingerprints for intentionally functional classification systems, such as EC (Bairoch, 2000) or GO (Gene Ontology Consortium, 2004); we will report these results in the near future.

We have observed annotations that initially appear to disagree with our inferences, sometimes because the annotation was speculative, and sometimes because the level of classification was too coarse or too fine. An example of both is 1m65, which is in the PHP-domain family in SCOP. We classify it as a metallo-dependent hydro-

lase, and the Gene Ontology Annotation (GOA) database (Camon *et al.*, 2004) annotates it as having *DNA-directed DNA polymerase activity* (GO:0003887), a putative function assignment based on electronic annotation transferred from the sequence database InterPro. The discoverers of the PHP-domain sequence family (Aravind & Koonin, 1998) indicated that the metallo-dependent hydrolases share active site sequence motifs with this family, and hypothesized that bacterial and archaeal DNA polymerases possess intrinsic phosphatase activity. Since several metallo-dependent hydrolases can hydrolyze phosphoester or phosphate bonds, the assigned GO term may still support the function inferred by our method.

The designed robustness of our method suggests that it could be used to predict function from the sequence level using either good quality predicted structures, or sequence patterns derived from fingerprints whose sequence order is preserved within a family. Investigations in this direction are ongoing.

Our method has limitations, arising from representation choices, algorithmic issues, and the nature of the problem itself. In our representation, we use C_α coordinates to calculate graph edges and lengths; this choice captures shared topology, but may miss contacts with long sidechains. Currently we do not allow residue substitutions in patterns, other than unifying V,A,I,L. Merging commonly substituted residue types (e.g. D,E) increases the sensitivity of fingerprints but can decrease their specificity; we may lose fingerprints that are no longer unique to a family. Finally, the distance edge matching criteria may be too restrictive to find patterns with widely varying geometry or containing edges that happen to lie on bin boundaries. We are developing a new distance edge representation to remedy this last problem.

Algorithmically, subgraph mining involves the NP-complete problem of subgraph isomorphism. The FFSM algorithm (Huan *et al.*, 2004) stores graph embeddings, so it does well with small isomorphic subgraphs, but can bog down with the large ones that can arise in families with very similar or identical structures.

It is part of the nature of the problem that classifications that are too fine can produce too many fingerprints due to high local similarity or small sample sizes, i.e. families with 3 or fewer members. Conversely, too coarse a classification can produce no fingerprints that are specific to a family — this happens with 35% of our SCOP families and superfamilies, especially the latter because of their heterogeneity. Because the number, specificity, and sensitivity of fingerprints depends on size and heterogeneity of the family, the support and background occurrence parameters must be varied to find meaningful sets of fingerprints for the maximum number of families.

In conclusion, the method identifies fingerprints for functional families with four or more representatives by finding packing patterns characteristic to each family, and uses them to infer function. Structure errors, missing fragments or mutations may lead to failure of fingerprint mining or function inference. Careful manual selection of families and fixing errors in structure files should improve the results further. Since our method infers function for many orphan proteins, the ultimate proof will come from experimental validation of its predictions.

Materials and Methods

Our method initially finds and calibrates fingerprints (steps 1–4) using the FFSM subgraph mining program from (`http://www.cs.unc.edu/~huan/FFSM/`). Then there are two steps (5–6) for each function inference. These are implemented in MATLAB.

1. Family and background selection: We selected 120 families and superfamilies from SCOP version 1.65. Though SCOP 1.67 was released in February 2005, we have retained the fingerprints derived from SCOP 1.65 to allow unbiased function prediction of structural orphans using information known at the

time they were selected, and use new members added in SCOP 1.67 to validate the method. In addition to requiring better than 3 Å resolution and R-factor at most 1.0, we reduced redundancy by using PISCES (Wang & Dunbrack, 2003) to select family members having at most 90% sequence identity. The same criteria when applied to the entire PDB produced a representative set of 6,749 protein chains in May 2005, which we used as the *background* for identifying fingerprints. The lists of families, family members, and background selected are in online supplementary materials at `http://www.cs.unc.edu/~debug/papers/FuncInf`

2. Graph Representation: We represent protein structures as graphs, with nodes at each residue labeled with the amino acid type, with V,A,I,L condensed to a single type since they frequently substitute for one another. Edges represent contact between residues defined by *almost-Delaunay* edges (Bandyopadhyay & Snoeyink, 2004), or distance constraints between non-contacting residues. Edges are labeled with length ranges (0–4, 4–6, 6–8.5, 8.5–10.5, 10.5–12.5 and 12.5–15Å). Fingerprints mined using this graph representation are called *distance edge fingerprints*. We do some experiments (e.g., in the metallo-dependent hydrolase family) using *simple edge fingerprints*, which omit the distance labels.

3. Frequent Subgraph Mining: We mine frequent subgraphs from the graph representation of all proteins in a family using Fast Frequent Subgraph Mining (Huan *et al.*, 2005). We use a support value of 80% to define frequency. Frequent subgraphs are constrained to have high density by having no more than one edge missing from a clique.

4. Fingerprint Identification: Fingerprints are defined as those subgraphs found in at least 80% of the family (support), and at most 5% of the background (background occurrence). The aim for families in our dataset is to have 10–1000 fingerprints; the support and background occurrence are adjusted for small or heterogeneous families until the number of fingerprints is in this range.

5. Search for Fingerprints in Query: We use a graph similarity index to speed up the subgraph isomorphism algorithm of Ullman (1976). For each node of the fingerprints and of a query structure, we create an index vector that stores the labels of neighboring nodes and edges connected to them, and consider a query embedding a node in a fingerprint only if the index vectors match. This reduces billions of potential embeddings of fingerprints to a handful in most cases. Ullman's algorithm then finds all embeddings of the fingerprint in the query that match node and edge labels.

6. Assigning Significance: We assign significance to the function inference by comparing the number of fingerprints found against the distribution of fingerprints in background proteins and in family members. Because these distributions are not normal, we calculate p-values empirically. By picking different numbers of fingerprints at which to infer family membership, we can determine the rates of true and false positives and negatives, calculate specificity and sensitivity, and draw ROC curves as shown in the inset of Figure 1(a)–(b). We choose two cutoff points for each family: a *sensitivity cutoff* to maximize sensitivity with at least 95% specificity, and a higher *99%-specificity cutoff* with no constraints on sensitivity.

Electronic Supplementary Material: Table I describes the SCOP families for which we obtained fingerprints. Tables II– IV give results from the SCOP validation experiment. Other supplementary data, including kinemages showing the graph representations of fingerprints for the proteins in Figure 1(c)–(f), may be viewed at `http://www.cs.unc.edu/~debug/papers/FuncInf`.

Acknowledgments

DB and JS gratefully acknowledge support from NSF grants 9988742 and 0076984; JH, JL, JP and WW from the Microsoft Research eScience RFP award; and AT appreciates the support from the NSF grant ITR/MCB 011289 and a grant from North Carolina – Israel Research Partnership NCI 1999032. The authors thank Ruchir Shah for many useful discussions.

References

Aloy, P., Querol, E., Aviles, F. X., and Sternberg, M. J.. 2001. Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. *J Mol Biol* **311**:395–408.

Aravind, L., and Koonin, E. V.. 1998. Phosphoesterase domains associated with DNA polymerases of diverse origins. *Nucleic Acids Res* **26**:3746–3752.

Artymiuk, P. J., Poirrette, A. R., Grindley, H. M., Rice, D. W., and Willett, P.. 1994. A graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. *Journal of Molecular Biology* **243**:327–44.

Bairoch, A.. 2000. The enzyme database in 2000. *Nucleic Acids Res.* **28** (1):304–305.

Bandyopadhyay, D.. 2005. *A Geometric Framework for Robust Nearest Neighbor Analysis of Protein Structure and Function*. PhD thesis, University of North Carolina, Chapel Hill, NC.

Bandyopadhyay, D., and Snoeyink, J.. 2004. Almost-Delaunay simplices: nearest neighbor relations for imprecise points. In *ACM-SIAM Symposium On Discrete Algorithms* pp. 403–412,.

Burley, S. K.. 2000. An overview of structural genomics. *Nat Struct Biol* **7 Suppl**:932–934.

Camon, E., Magrane, M., Barrell, D., Lee, V., Dimmer, E., Maslen, J., Binns, D., Harte, N., Lopez, R., and Apweiler, R.. 2004. The Gene Ontology Annotation (GOA) database: sharing knowledge in Uniprot with Gene Ontology. *Nucleic Acids Research* **32** (1):D262–D266.

Cheng, G., Qian, B., Samudrala, R., and Baker, D.. 2005. Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design. *Nucl. Acids Res.* **33** (18):5861–5867.

Gene Ontology Consortium. 2004. The Gene Ontology (GO) database and informatics resource. *Nucl. Acids. Res.* **32** (90001):D258–261.

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.

Holm, L., and Sander, C.. 1996. Mapping the protein universe. *Science* **273**:595–602.

Huan, J., Bandyopadhyay, D., Wang, W., Snoeyink, J., Prins, J., and Tropsha, A.. 2005. Comparing graph representations of protein structure for mining family-specific residue-based packing motifs. *Journal of Computational Biology* **12** (6):657–671.

Huan, J., Wang, W., Bandyopadhyay, D., Snoeyink, J., Prins, J., and Tropsha, A.. 2004. Mining protein family specific residue packing patterns from protein structure graphs. In *Eighth Annual International Conference on Research in Computational Molecular Biology (RECOMB)* pp. 308–315,.

Humphrey, William, Dalke, Andrew, and Schulten, Klaus. 1996. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics* **14**:33–38.

Kinch, L. N., Qi, Y., Hubbard, T. J., and Grishin, N. V.. 2003. CASP5 target classification. *Proteins* **53 Suppl 6**:340–351.

Ko, J., Murga, L. F., Wei, Y., and Ondrechen, M. J.. 2005. Prediction of active sites for protein structures from computed chemical properties. *Bioinformatics* **21 Suppl 1**:i258–i265.

Laskowski, R. A., Watson, J. D., and Thornton, J. M.. 2005*a*. ProFunc: a server for predicting protein function from 3D structure. *Nucleic Acids Res* **33**:W89–93.

Laskowski, Roman A., Watson, James D., and Thornton, Janet M.. 2005*b*. Protein function prediction using local 3D templates. *J Mol Biol* **351**:614–626.

Madera, M., Vogel, C., Kummerfeld, S. K., Chothia, C., and Gough, J.. 2004. The SUPERFAMILY database in 2004: additions and improvements. *Nucleic Acids Res* **32**:D235–9.

Milik, M, Szalma, S, and Olszewski, KA. 2003. Common structural cliques: a tool for protein structure and function analysis. *Protein Eng.* **16(8)**:543–52.

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. *Journal of Molecular Biology* **247**:536–40.

Ofran, Y., Punta, M., Schneider, R., and Rost, B.. 2005. Beyond annotation transfer by homology: novel protein-function prediction methods to assist drug discovery. *Drug Discovery Today* **10** (21):1475–1482.

Pal, D., and Eisenberg, D.. 2005. Inference of protein function from protein structure. *Structure (Camb)* **13**:121–130.

Pazos, F., and Sternberg, M. J.. 2004. Automated prediction of protein function and detection of functional sites from structure. *Proc Natl Acad Sci U S A* **101**:14754–14759.

Pegg, S. C., Brown, S., Ojha, S., Huang, C. C., Ferrin, T. E., and Babbitt, P. C.. 2005. Representing structure-function relationships in mechanistically diverse enzyme superfamilies. In *Pac Symp Biocomput* pp. 358–369,.

Serres, M. H., Goswami, S., and Riley, M.. 2004. GenProtEC: an updated and improved analysis of functions of Escherichia coli K-12 proteins. *Nucleic Acids Res* **32**:D300–2.

Stark, A, and Russell, RB. 2003. Annotation in three dimensions. PINTS: patterns in non-homologous tertiary structures. *Nucleic Acids Res* **31** (13):3341–4.

Stark, A., Shkumatov, A., and Russell, R. B.. 2004. Finding functional sites in structural genomics proteins. *Structure (Camb)* **12**:1405–1412.

Teplyakov, A., Obmolova, G., Khil, P. P., Howard, A. J., Camerini-Otero, R. D., and Gilliland, G. L.. 2003. Crystal structure of the Escherichia coli YcdX protein reveals a trinuclear zinc active site. *Proteins* **51**:315–318.

Ullman, J. R.. 1976. An algorithm for subgraph isomorphism. *Journal of the Association for Computing Machinery* **23**:31–42.

Wang, G., and Dunbrack, R. L.. 2003. PISCES: a protein sequence culling server. *Bioinformatics* **19**:1589–1591. `http://www.fccc.edu/research/labs/dunbrack/pisces/culledpdb.html`.

Wang, K., and Samudrala, R.. 2005. FSSA: a novel method for identifying functional signatures from structural alignments. *Bioinformatics* **21** (13):2969–2977.

Wangikar, P. P., Tendulkar, A. V., Ramya, S., Mali, D. N., and Sarawagi, S.. 2003. Functional sites in protein families uncovered via an objective and automated graph theoretic approach. *J Mol Biol* **326** (3):955–78.

Wilson, C. A., Kreychman, J., and Gerstein, M.. 2000. Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J Mol Biol* **297**:233–249.

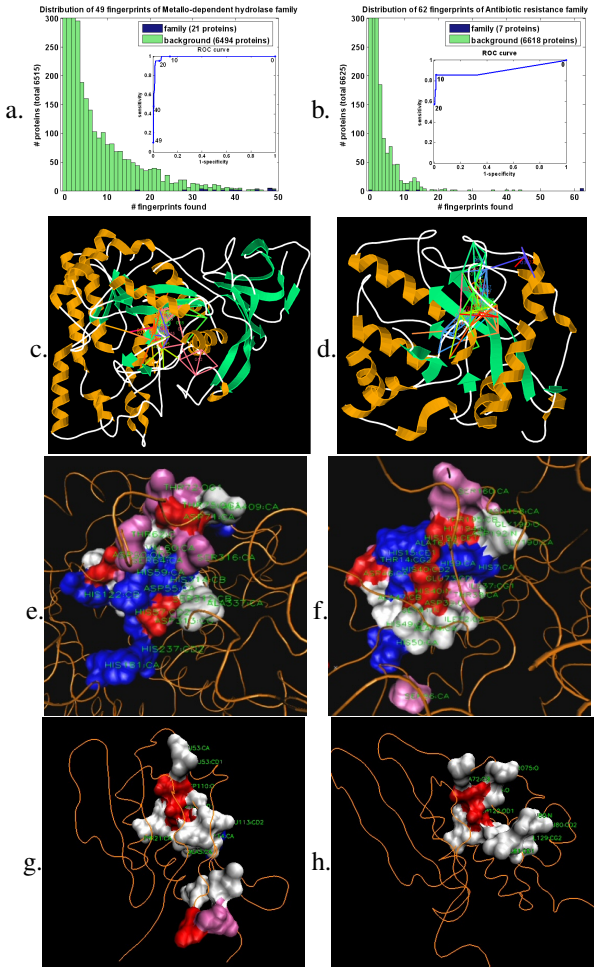

Figure 1: Distribution of (a) Metallo-dependent hydrolase (SCOP: 51556) and (b) Antibiotic resistance (SCOP: 54598) fingerprints in the background (light bars), and within the family (dark). Inset: ROC curve showing specificity vs. sensitivity of function inference at different numbers of fingerprints. (c)–(d) Example of function inference: metallo-dependent hydrolase fingerprints (shown as graphs) in (c) the metallo-dependent hydrolase 1nfg, and (d) 1m65 (Ycdx, unknown function). (e)–(f) the same proteins shown as residues covered by metallo-dependent hydrolase fingerprints, color-coded by chemical properties (g)–(h) Another example of function inference: residues covered by antibiotic resistance fingerprints in (e) the family protein 1ecs, and (f) 1twu (Yyce, unknown function). Figures (c)–(f) are snapshots from VMD (Humphrey *et al.*, 1996).

(a)

abrv	SCOP (sup)fam	#mem	#fp	$\#fp_{max}^{TIMfold}$
ALD	Aldolase (Class I)	17	74	41 (1gteA=FMN)
AKN	Aldo-keto reductase	9	72	4 (1hvxA=AMY)
AMY	Amylase	39	42	23 (1mucA=DGL)
BLU	Bacterial luciferase	6	86	20 (1ofdA=FMN)
BGL	Beta-glycanase	27	28	19 (1cgt=AMY)
B12	Cobalamin (B12) dependent enzyme	5	66	8 (1a3wA=PYK)
DGL	D-glucarate dehydratase-like	9	29	6 (1gkrA=MDH)
DHP	Dihydropteroate synthase-like	4	41	5 (1eepA=IMP)
ENC	Enolase C-terminal domain like	14	111	60 (1ofdA=FMN)
FMN	FMN-linked oxidoreductase	14	55	31 (1me8A=IMP)
GH1	Glycosyl hydrolase family 1	9	887	12 (1bqcA=BGL)
HIS	Histidine biosynthesis enzyme	5	290	47 (1ofdA=FMN)
MDH	Metallo-dependent hydrolase	17	49	37 (1lwjA=AMY)
PYK	Pyruvate kinase	5	819	25 (1ea0A=FMN)
RIB	Ribulose-phosphate binding barrel	17	20	19 (1jr1A=IMP)
RBC	RuBisCo	7	315	14 (1ea0A=FMN)
TIM	Triosephosphate isomerase	12	366	9 (1ea0A=FMN)
TRP	Tryptophan biosynthesis enzyme	6	22	15 (1pvnA=IMP)
XIM	Xylose isomerase	7	758	10 (1n8iA=MAL)
XIL	Xylose-isomerase-like	11	17	5 (1e0tA=PYK)

(b)

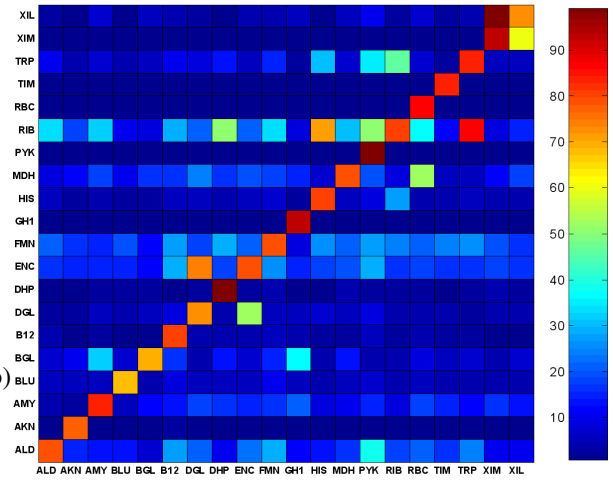

Figure 2: Discriminating the TIM barrels using fingerprints. (a) The 20 families selected, with columns listing a 3-letter abbreviation for each family, number of members and fingerprints, and maximum number of fingerprints found in a non-family protein of the TIM fold. Families mentioned in the last column for which fingerprints were not identified: IMP (inosine monophosphate dehydrogenase) and MAL (malate synthase). (b) Pseudo-color matrix plot showing the percentage of fingerprints of the TIM barrel family in each row found in an average member of the family in each column. High values on the diagonal (red) and low off-diagonal (blue) indicate high discrimination. Exceptions to this trend are documented in the text.

Fingerprints for SCOP families using distance edge representation									
SCOP	**node**	**family**	**#**	**non-std.**	**#**	**sensitivity**		**99% specificity**	
ID	**type**	**name**	**prot**	**param**	**fp**	**cut pt**	**sens.**	**cut pt**	**sens.**
48179	sf	6-phosphogluconate dehydrogenase C-terminal domain-like	12	$b0.1$	16	10	0.85	10	0.85
53384	fa	AAT-like	14	$b0.1d2$	36	10	1.00	10	1.00
52686	fa	ABC transporter ATPase domain-like	12	$b0.1$	17	4	0.87	9	0.67
55753	sf	Actin depolymerizing	9	$f0.7b0.15d2$	94	29	0.80	38	0.80
52402	sf	Adenine nucleotide α-hydrolase-like	16	$b0.1d2$	13	7	0.74	10	0.53
52397	fa	Adenylyltransferase	9	$b0.1$	216	47	0.91	47	0.91
51883	fa	Aminoacid dehydrogenase-like, C-terminal domain	18	$f0.7b0.1d2$	376	145	1.00	153	0.94
51570	fa	Aldolase, Class I	17	$b0.1d2$	74	27	1.00	27	1.00
51431	fa	Aldo-keto reductases (NADP)	9	$f0.9b0.02$	72	3	1.00	3	1.00
53649	sf	Alkaline phosphatase-like	7		75	12	1.00	12	1.00
75217	sf	α/β knot	6	$b0.02$	22	3	0.42	5	0.42
51446	fa	Amylase, catalytic domain	39	$b0.1$	42	14	0.95	16	0.87
54598	fa	Antibiotic resistance proteins	4	$f1.0$	62	11	0.86	13	0.71
		Antibiotic resistance (*SCOP 1.67*)	7		18	5	0.86	5	0.86
48942	fa	Antibody constant (C1) domain	259	$b0.1d0$	12	2	1.00	9	0.84
48727	fa	Antibody variable (V) domain	310	$b0.1$	92	7	0.98	15	0.98
48371	sf	ARM repeat	18	$f0.7b0.1d2$	294	86	0.83	124	0.78
53570	fa	Bacterial lipase	5	$f1.0b0.02d0$	132	4	1.00	6	1.00
51679	sf	Bacterial luciferase-like	6	$d0$	86	8	0.75	11	0.75
53057	fa	β-carbonic anhydrase	4	$f1.0b0.02d0$	62	2	1.00	3	1.00
51487	fa	β-glycanases	27	$f0.7b0.15d2$	152	54	0.92	78	0.75
				$f0.7b0.1d2$	28	9	0.85	14	0.81
56602	fa	β-Lactamase/D-ala carboxypeptidase	20	$f0.7b0.1d2$	73	21	0.95	21	0.95
56655	sf	Carbohydrate phosphatase	12	$b0.1$	69	21	0.92	21	0.92
49384	sf	Carbohydrate-binding domain	7	$b0.02$	17	4	1.00	4	1.00
56317	sf	Carbon-nitrogen hydrolase	3	$f1.0d0$	97	9	1.00	9	1.00
53487	fa	Carboxylesterase	5	$d0$	59	9	1.00	9	1.00
51990	fa	Cellulase	4	$f1.0b0.02d0$	295	4	1.00	6	1.00
52172	sf	CheY-like (*SCOP 1.67*)	20	$f0.7b0.15d2$	141	65	0.85	89	0.65
52173	fa	CheY-related (*SCOP 1.67*)	17	$f0.7b0.02d2$	17	2	0.82	5	0.65
49348	sf	Clathrin adaptor appendage domain	5		77	15	1.00	15	1.00
52243	fa	Cobalamin (vitamin B12)-binding	4	$f1.0$	35	6	1.00	9	1.00
51703	sf	Cobalamin (vitamin B12)-dep. enzyme	5	$b0.02d0$	66	6	1.00	6	1.00
49330	fa	Cu,Zn superoxide dismutase-like	10	$b0.02d0$	183	5	0.91	5	0.91
49550	fa	Cupredoxin, multidomain	11	$b0.02$	222	17	1.00	17	1.00
75434	fa	CutA divalent ion tolerance protein	3	$f1.0d0$	58	5	1.00	5	1.00
75434	fa	CutA divalent ion tolerance (augmented)	6	$f0.9b0.1d0$	132	12	0.88	12	0.88
53402	fa	Cystathionine synthase-like	15	$f0.7d2$	17	6	0.72	6	0.72
50353	sf	Cytokine	9	$b0.1$	45	9	1.00	11	0.92
51609	fa	D-glucarate dehydratase-like	9		29	6	0.92	6	0.92
52467	sf	DHS-like NAD/FAD-binding domain	14	$b0.1$	31	11	0.89	17	0.79
51717	sf	Dihydropteroate synthetase-like	4	$f1.0b0.02d0$	41	4	1.00	4	1.00
53118	fa	DnaQ-like 3'-5' exonuclease	11		15	5	0.93	5	0.93
54768	sf	dsRNA-binding domain-like	5		39	10	1.00	10	1.00
				$b0.02d2$	43	7	1.00	7	1.00
57196	sf	EGF/Laminin	21	$b0.1d0$	24	11	1.00	15	0.88
50090	sf	Electron transport accessory protein	5	$b0.1$	28	13	1.00	13	1.00
54060	sf	Endonuclease, His-Me finger	5	$b0.1d2$	57	18	1.00	18	1.00
55608	sf	Endonuclease, homing	4	$f1.0b0.1$	25	7	1.00	10	1.00
51604	sf	Enolase C-terminal domain-like	14	$b0.1d0$	111	29	0.94	38	0.82
52432	fa	ETFP subunit	6	$f0.9d0$	85	20	1.00	20	1.00
50514	fa	Eukaryotic serine protease	56	$b0.02d0$	155	5	0.97	5	0.97
81269	fa	Extended AAA-ATPase domain	17	$b0.1d2$	13	10	0.82	10	0.82
54602	fa	Extradiol dioxygenase	6	$f0.9$	150	18	1.00	18	1.00
46610	fa	Fe,Mn superoxide dismutase (SOD), N-terminal domain	14	$b0.02d0$	113	3	1.00	3	1.00
51396	fa	FMN-linked oxidoreductase	14	$b0.1$	55	26	1.00	26	1.00
50354	fa	Fibroblast growth factor (FGF)	6	$f0.9d0$	143	6	1.00	6	1.00
53558	fa	Fungal lipase	8	$b0.02$	85	9	1.00	9	1.00
52592	fa	G protein	42	$b0.1d2$	39	14	0.98	17	0.96
51187	fa	Germin/Seed storage 7S protein	10	$f0.7b0.1d2$	24	9	0.90	9	0.90
52318	fa	Glutamine amidotransferase class I	10	$b0.1$	55	11	1.00	16	0.83

Continued on next page

SCOP ID	node type	family name	# prot	non-std. param	# fp	sensitivity cut pt	sensitivity sens.	99% specificity cut pt	99% specificity sens.
56236	fa	Glutamine amidotransferase class II	7	$d0$	299	37	0.88	37	0.88
				$b0.02d0$	83	6	0.88	6	0.88
55931	sf	Glutamine synthetase/guanido kinase	8	$f1.0d0$	16	4	1.00	4	1.00
				$f0.9b0.01d0$	941	9	0.89	9	0.89
51011	sf	Glycosyl hydrolase domain	37	$b0.1d0$	81	16	0.88	25	0.86
51521	fa	Glycosyl hydrolase family 1	9	$f0.9b0.01d0$	887	6	1.00	6	1.00
56784	sf	Haloacid dehalogenase (HAD) like	9	$d2$	45	7	0.95	11	0.84
56784	sf	HAD-like (*SCOP 1.67*)	19	$f0.7b0.1$	50	13	0.84	20	0.84
53531	fa	Haloperoxidase	4	$f1.0b0.01d0$	372	5	1.00	5	1.00
51367	fa	Histidine biosynthesis enzymes	5	$d0$	290	26	1.00	26	1.00
51413	fa	Inosine monophosphate dehydrogenase (IMPDH)	6	$f0.9b0.01d0$	391	9	1.00	9	1.00
53301	fa	Integrin A (or I) domain	6	$f0.9b0.02$	214	17	1.00	22	0.89
53659	sf	Isocitrate/Isopropylmalate dehydrogenase-like	8	$d0$	208	17	1.00	17	1.00
49944	fa	Laminin G-like module	4	$f1.0b0.1$	119	23	1.00	33	1.00
63379	fa	MarR-like transcriptional regulator	5	$b0.1d2$	36	16	0.71	16	0.71
51556	sf	Metallo-dependent hydrolase	17	$b0.1$	49	28	0.95	33	0.76
56300	sf	Metallo-dependent phosphatase	10	$b0.02$	316	6	0.91	6	0.91
56281	sf	Metallo-hydrolase/oxidoreductase	7		30	6	0.89	6	0.89
55486	sf	Metalloprotease ("zincin") catalytic domain	42	$b0.1d2$	34	10	0.94	15	0.85
53218	sf	Molybdenum cofactor biosynthesis	5	$b0.02d0$	73	6	0.83	6	0.83
52641	fa	Motor protein	11	$b0.02$	25	4	0.87	4	0.87
52403	fa	N-type ATP pyrophosphatase	6	$b0.02$	118	10	0.88	10	0.88
55468	cf	NADH oxidase/flavin reductase	6	$f0.9b0.1$	60	16	1.00	16	1.00
54431	fa	NTF2-like	5	$b0.1$	17	6	0.86	7	0.71
48509	fa	Nuclear receptor ligand-binding	23	$b0.1$	67	17	1.00	17	1.00
				$b0.1d2$	261	63	1.00	63	1.00
				$b0.02d2$	83	8	1.00	8	1.00
81301	sf	Nucleotidyltransferase	6	$b0.1d0$	110	29	1.00	29	1.00
55811	sf	Nudix hydrolase	5	$d2$	211	24	1.00	24	1.00
49417	sf	p53-like transcription factors	11	$f0.7b0.1d2$	57	17	0.91	17	0.91
54002	fa	Papain-like cysteine protease	19	$b0.02$	178	4	0.86	8	0.86
				$f0.9b0.02$	327	7	0.89	14	0.89
49354	sf	PapD-like	7	$b0.1$	123	27	0.88	27	0.88
50157	fa	PDZ domain	9	$b0.1$	132	32	1.00	32	1.00
63550	fa	Penta-EF-hand proteins	4	$f1.0b0.1$	107	11	1.00	17	1.00
50646	fa	Pepsin-like acid protease	24	$b0.02$	145	11	0.88	11	0.88
53822	sf	Periplasmic binding protein-like I	8		23	9	1.00	9	1.00
51998	sf	PFL-like glycyl radical enzymes	4	$f1.0d0$	21	4	1.00	5	1.00
48537	sf	Phospholipase C/P1 nuclease	4	$f1.0d0$	32	5	1.00	5	1.00
88723	sf	PIN domain-like	6	$f0.9b0.1$	78	24	0.86	24	0.86
55771	fa	Profilin (actin-binding protein)	7	$b0.02d0$	76	4	1.00	4	1.00
50495	fa	Prokaryotic serine protease	10	$b0.1d0$	11	6	0.86	6	0.86
54815	fa	Prokaryotic type KH domain (type II)	5	$b0.02$	76	12	1.00	12	1.00
88854	fa	Protein kinase, catalytic subunit	32	$b0.1$	92	20	0.98	20	0.98
53167	sf	Purine and uridine phosphorylases	7	$b0.1d0$	129	29	0.67	41	0.67
53182	sf	Pyrrolidone carboxyl peptidase	4	$f1.0b0.01d0$	558	4	1.00	7	1.00
51622	fa	Pyruvate kinase	5	$f1.0b0.02d0$	819	15	1.00	15	1.00
52475	fa	Pyruvate oxidase and decarboxylase, middle domain	5	$b0.02d0$	177	8	1.00	8	1.00
52670	fa	RecA protein-like (ATPase-domain)	11	$b0.1$	15	6	0.91	8	0.73
53099	fa	Ribonuclease H	7		155	24	1.00	24	1.00
51366	sf	Ribulose-phosphate binding barrel	17	$b0.1$	20	8	0.88	14	0.76
50371	fa	Ricin B-like	7	$f0.9b0.02d0$	412	10	1.00	10	1.00
				$f0.9d0$	521	23	1.00	23	1.00
51650	fa	RuBisCo, C-terminal domain	7	$f0.9b0.02d0$	315	10	1.00	10	1.00
46928	cf	RuvA C-terminal domain-like	7	$b0.1d2$	59	23	0.71	23	0.71
48426	fa	Sec7 domain	3	$f1.0b0.01d0$	679	5	1.00	8	1.00
56575	fa	Serpins	12	$b0.02d2$	689	22	1.00	22	1.00
				$b0.1$	287	30	1.00	30	1.00
52266	sf	SGNH hydrolase	5	$b0.1d0$	42	9	1.00	9	1.00
50045	fa	SH3-domain	17	$f0.7b0.1d2$	17	4	0.88	5	0.82
53697	sf	SIS domain	6		453	61	1.00	61	1.00

Continued on next page

SCOP ID	node type	family name	# prot	non-std. param	# fp	sensitivity cut pt	sensitivity sens.	99% specificity cut pt	99% specificity sens.
50386	sf	STI-like	7	$b0.02$	91	8	1.00	8	1.00
					265	30	1.00	30	1.00
52744	fa	Subtilases	7	$f1.0b0.01d0$	53	2	1.00	3	1.00
52210	sf	Succinyl-CoA synthetase domains	5	$f1.0b0.1d0$	45	11	1.00	16	1.00
55620	sf	Tetrahydrobiopterin biosynthesis-like	4	$f1.0b0.1$	27	9	1.00	14	1.00
48453	fa	Tetratricopeptide repeat (TPR)	9	$b0.1$	14	6	1.00	6	1.00
54637	sf	Thioesterase/thiol ester dehydrase-isomerase	7	$b0.1d2$	35	15	1.00	15	1.00
52834	fa	Thioltransferase	11		42	4	0.76	4	0.76
51352	fa	Triosephosphate isomerase (TIM)	12	$f0.9b0.02d0$	366	7	1.00	7	1.00
50495	fa	Trypsin-like serine proteases	66	$d0$	45	3	0.97	5	0.94
50514	fa	(prokaryotic and eukaryotic)							
51381	fa	Tryptophan biosynthesis enzyme	6	$f0.9d0$	22	5	1.00	7	0.83
54496	fa	Ubiquitin conjugating enzyme	14	$d0$	19	3	0.94	3	0.94
48468	fa	VHS domain	4	$f1.0$	84	9	1.00	15	1.00
50603	fa	Viral cysteine protease, trypsin fold	7	$d0$	17	4	1.00	4	1.00
50979	fa	WD40-repeat	8	$b0.02$	1763	60	1.00	60	1.00
				$d0$	366	29	1.00	29	1.00
51658	sf	Xylose isomerase-like	11	$d0$	17	3	0.91	4	0.82
51665	fa	Xylose isomerase	7	$f0.9b0.01d0$	758	6	1.00	6	1.00
55299	fa	YjgF/L-PSP	4	$f1.0b0.02d0$	131	15	1.00	24	1.00
57668	fa	Zinc finger, classic C2H2	5		17	4	1.00	4	1.00
53187	sf	Zn-dependent exopeptidases	11	$b0.1$	14	5	0.58	7	0.50

Table I: **Left:** The SCOP families used to define distance edge fingerprints, shown as SCOP ID, node type (fa=family, sf=superfamily), family description and number of 90% non-redundant structures in the family. Exception: Simple edge fingerprints are shown for Metallo-dependent Hydrolases, as discussed in the paper. **Middle:** Fingerprints were obtained with default mining parameters – distance threshold 8.5 $\mathring{\text{A}}$, $AD(0.1)$ graph representation with 17 amino acid labels (V,A,I,L merged); by default family support(f) was 0.8, background occurrence(b) was 0.05 and density(d) was at most 1 edge missing from a clique. Non-default values of these parameters are shown, along with the number of fingerprints obtained after mining. **Right:** Using the ROC curve from the distribution of fingerprints in the background, cutoff points were determined for each family based on sensitivity and on 99% specificity; they are listed along with the coverage (sensitivity) for the family at that number of fingerprints.

SCOP ID	SCOP (super)family	# prot	# fp	new family members inferred #	sens	spec	other >spec	other fns. inferred spec	sens	# inf from sequence
53384	AAT-like	14	36	2	2	2	0	1	5	1
52686	ABC transporter ATPase domain-like	12	17	3	3	1	2	6	10	2
55753	Actin depolymerizing protein	9	94	1	1	1	0	2	3	1
52402	Adenine nucleotide α-hydrolase	16	13	3	2	0	1	1	3	0
51431	Aldo-keto reductases (NADP)	9	72	8	8	8	0	2	9	3
53649	Alkaline phosphatase-like	7	75	2	1	1	0	2	6	0
51446	Amylase, catalytic domain	39	42	3	2	2	0	3	8	1
54598	Antibiotic resistance protein	4	62	3	2	1	1	1	1	2
48727	Antibody variable (V) domain	310	92	63	63	63	1	1	3	60
48371	ARM repeat	18	294	6	2	1	1	1	2	1
51487	β-glycanase	27	28	9	6	5	0	2	5	2
56602	β-Lactamase/D-ala carboxypeptidase	20	73	7	5	5	1	4	10	5
53487	Carboxylesterase	5	59	1	1	1	0	2	8	0
75434	CutA divalent ion tolerance protein	3	58	4	4	4	0	1	2	2
49330	Cu,Zn superoxide dismutase-like	10	183	1	1	1	0	1	1	1
50353	Cytokine	9	45	4	4	3	0	1	1	2
51609	D-glucarate dehydratase-like	9	29	3	3	3	0	6	12	1
52467	DHS-like NAD/FAD-binding domain	14	31	5	3	2	3	27	38	2
51717	Dihydropteroate synthetase-like	4	41	1	1	1	0	10	17	0
53118	DnaQ-like 3'-5' exonuclease	11	15	3	2	2	1	10	23	1
57196	EGF/Laminin	21	24	3	3	3	0	3	4	1
51604	Enolase C-terminal domain-like	14	111	3	2	1	1	5	8	1
50514	Eukaryotic serine protease	56	155	5	4	4	0	3	5	2
81269	Extended AAA-ATPase domain	17	13	4	2	2	2	18	27	1
46610	Fe,Mn superoxide dismutase (SOD)	14	113	2	2	2	0	2	3	2
50354	Fibroblast growth factor (FGF)	6	143	3	3	3	0	1	1	1
51396	FMN-linked oxidoreductase	14	55	8	4	4	1	21	35	3
53558	Fungal lipase	8	85	1	1	1	0	1	2	0
52592	G protein	42	39	8	8	8	0	10	17	8
52318	Glutamine amidotransferase Class I	10	55	2	2	0	1	1	2	0
55931	Glutamine synthetase/guanido kinase	8	16	1	1	1	0	1	3	1
51011	Glycosyl hydrolase domain	37	81	6	2	2	0	3	10	0
56784	HAD-like	9	45	10	9	7	1	8	13	1
49944	Laminin G-like module	4	119	1	1	0	1	1	1	0
51556	Metallo-dependent hydrolases	17	49	4	4	3	1	9	25	1
56281	Metallo-hydrolase/oxidoreductase	7	30	2	1	1	0	0	1	0
55486	Metalloprotease ("zincin")	42	34	5	5	3	1	3	4	4
52641	Motor protein	11	25	4	3	3	0	1	6	1
49550	Multidomain cupredoxin	11	222	1	1	1	0	2	4	1
52403	N-type ATP pyrophosphatase	6	118	2	1	1	0	3	4	0
48509	Nuclear receptor ligand-binding domain	23	83	7	7	7	0	3	5	3
55811	Nudix	5	211	7	5	5	0	1	3	0
49417	p53-like transcription factor	11	57	3	1	1	0	0	0	1
54002	Papain-like	19	178	9	7	7	0	1	3	7
50646	Pepsin-like acid protease	24	145	3	3	3	0	3	5	2
48537	Phospholipase C/P1 nuclease	4	32	1	1	1	0	1	2	1
50495	Prokaryotic serine protease	10	11	4	3	3	2	2	4	0
88854	Protein kinases, catalytic subunit	32	92	17	17	17	1	5	10	9
53167	Purine and uridine phosphorylase	7	129	5	1	1	0	0	2	3
52475	Pyruvate oxidase and decarboxylase	5	177	3	3	3	1	33	45	0
52670	RecA protein-like (ATPase-domain)	11	15	2	1	1	1	1	3	1
50371	Ricin B-like	7	412	3	3	3	0	2	6	2
56575	Serpins	12	287	6	1	1	0	0	0	3
53697	SIS domain	6	453	2	1	1	0	0	2	0
52744	Subtilase	7	53	4	4	4	1	21	35	3
51381	Tryptophan biosynthesis enzyme	6	22	3	2	0	1	2	10	1
53187	Zn-dependent exopeptidase	11	14	13	3	1	2	3	4	4

Table II: Function inference of new members in SCOP 1.67 using distance edge fingerprints (only metallo-dependent hydrolase shown with simple edge fingerprints). The left part of the table describes the family used for mining fingerprints; the middle part lists the number of new members added, the number inferred at the sensitivity and 99% specificity cutoffs and the number inferred more strongly by another family's fingerprints; the right part list the number of other functions inferred for the new members at 99%-specificity and sensitivity cutoffs. The rightmost column gives the number of proteins for which the desired family can be inferred above 40% sequence identity, which is the threshold Wilson *et al.* (2000) use to infer function from sequence alone. Family members from SCOP 1.65 are sometimes missed when their sequence representative in the 90%-non-redundant dataset is not classified in SCOP 1.65, and appear to be new members. For such members, we ignore sequence hits above 90% identity to family members when evaluating the inference from sequence, since such hits would map to the same non-redundant protein.

SCOP ID	SCOP (super)family	# prot	# fp	# new members	# fingerprints in new members	sens. cutoff	# inf from sequence
48179	6-phosphogluconate dehydrogenase C-terminal domain-like	12	16	1	5	9	0
51570	Aldolase, class I	17	74	6	20 19 17 15 11 7	27	2
75217	α/β knot	6	22	6	2 2 2 1 1 0	3	0
51883	Aminoacid dehydrogenase-like, C-terminal domain	18	376	3	111 100 64	145	2
51679	Bacterial luciferase-like	6	86	2	6 5	8	0
56655	Carbohydrate phosphatase	12	69	1	16	21	1
49384	Carbohydrate-binding domain	7	17	1	3	4	1
49348	Clathrin adaptor appendage domain	5	77	1	4	15	0
53402	Cystathionine synthase-like	15	17	3	3 3 0	6	1
56236	Glutamine amidotransferase Class II	7	83	1	9	37	0
54060	His-Me finger endonucleases	5	57	1	11	18	0
55608	Homing endonucleases	4	25	1	6	7	0
53659	Isocitrate/Isopropylmalate dehydrogenase-like	8	208	3	16 15 7	17	0
63379	MarR-like transcriptional regulators	5	36	2	5 0	16	0
53218	Molybdenum cofactor biosynthesis proteins	5	73	1	1	6	0
81301	Nucleotidyltransferase	6	110	5	24 17 17 16 6	29	0
49354	PapD-like	7	123	1	9	27	0
53822	Periplasmic binding protein-like I	8	23	6	6 3 2 2 1 0	9	1
88723	PIN domain-like	6	78	1	19	24	1
53099	Ribonuclease H	7	155	1	8	24	1
51366	Ribulose-phoshate binding barrel	17	20	4	5 3 1 1	8	1
46928	RuvA C-terminal domain-like	7	59	2	6 5	23	0
55620	Tetrahydrobiopterin biosynthesis enzymes	4	27	3	5 1 0	9	0
48453	Tetratricopeptide repeat (TPR)	9	14	1	2	6	1
54637	Thioesterase/thiol ester dehydrase-isomerase	7	35	8	7 6 5 5 4 3 3 3	15	0

Table III: The families from Table II for which none of the new members were inferred; the number of fingerprints in each new member and the sensitivity cutoff points are also shown. The rightmost column gives the number of proteins whose function can reliably be inferred from sequence alone; this is usually low for these families.

SCOP ID	SCOP (super)family	# prot	# fp	new family members inferred #	sens	spec	other >spec	other fns. inferred spec	sens	# inf from sequence
50157	PDZ domain	9	132	5	5	5	0	0	0	2
52266	SGNH hydrolase	5	42	1	1	1	0	0	0	0
54496	Ubiquitin conjugating enzyme	14	19	1	1	1	0	0	0	1
48468	VHS domain	4	84	1	1	1	0	0	0	1
57668	Zinc finger, classic C2H2	5	17	1	1	1	0	0	0	0
48942	*Antibody constant (C1) domains*	259	12	*57*	*57*	*49*	*0*	*0*	2	57
56317	*Carbon-nitrogen hydrolase*	3	97	*1*	*1*	*1*	*0*	*0*	2	0
53301	*Integrin A (or I) domain*	6	214	*3*	*3*	*2*	*0*	*0*	2	1
56300	*Metallo-dependent phosphatase*	10	316	*1*	*1*	*1*	*0*	*0*	3	0
50045	*SH3-domain*	17	17	*4*	*4*	*4*	*0*	*0*	1	3
55299	*YjgF/L-PSP*	4	131	*3*	*3*	*3*	*0*	*0*	2	2

Table IV: The families from Table II for which all new members added in SCOP 1.67 were inferred using fingerprints of the SCOP 1.65 family, and there were either no other functions inferred (top half of tables) or there were no other functions inferred with 99% specificity or with higher specificity than the family fingerprints (bottom half, italics). The rightmost column gives the number of proteins whose function can reliably be inferred from sequence alone.