J Comput Aided Mol Des (2009) 23:785–797
DOI 10.1007/s10822-009-9277-0

Identification of family-specific residue packing motifs and their use for structure-based protein function prediction: II. Case studies and applications

Deepak Bandyopadhyay · Jun Huan · Jan Prins · Jack Snoeyink · Wei Wang · Alexander Tropsha

Published online: 23 June 2009

Abstract This paper describes several case studies concerning protein function inference from its structure using our novel approach described in the accompanying paper. This approach employs family-specific motifs, i.e. three-dimensional amino acid packing patterns that are statistically prevalent within a protein family. For our case studies we have selected families from the SCOP and EC classifications and analyzed the discriminating power of the motifs in depth. We have devised several benchmarks to compare motifs mined from unweighted topological graph representations of protein structures with those from distance-labeled (weighted) representations, demonstrating the superiority of the latter for function inference in most families. We have tested the robustness of our motif library by inferring the function of new members added to SCOP families, and discriminating between several families that are structurally similar but functionally divergent. Furthermore we have applied our method to predict function for several proteins characterized in structural genomics projects, including orphan structures, and we discuss several selected predictions in depth. Some of our predictions have been corroborated by other computational methods, and some have been validated by independent experimental studies, validating our approach for protein function inference from structure.

Keywords Structural genomics · Protein graphs · Protein function prediction · Family-specific motifs · Frequent subgraph mining · Orphan proteins

Introduction

The functions of proteins can often be inferred from their structure using elements of local packing, known as structural motifs. In a companion paper (Part I [1]) we have described a method for inferring protein function using *family-specific motifs*, i.e. 3D residue interaction patterns automatically extracted from protein families by mining graph representations of the protein structures. We also tested the performance of a graph index implemented to speed up motif searching. Using this method, we have derived a library of motifs characteristic of a large number of families annotated in SCOP [2] and EC databases.

Electronic supplementary material The online version of this article (doi:10.1007/s10822-009-9277-0) contains supplementary material, which is available to authorized users.

D. Bandyopadhyay (✉)
GlaxoSmithKline, 1250 S. Collegeville Rd, Mail Stop,
Collegeville, PA UP12-210, USA
e-mail: deepak.2.bandyopadhyay@gsk.com

J. Huan
Department of Electrical Engineering and Computer Science,
University of Kansas, Lawrence, KS, USA
e-mail: jhuan@eecs.ku.edu

J. Prins · J. Snoeyink · W. Wang
Department of Computer Science, University of North Carolina,
CB#3175 Sitterson Hall, Chapel Hill, NC, USA

J. Prins
e-mail: prins@cs.unc.edu

J. Snoeyink
e-mail: snoeyink@cs.unc.edu

W. Wang
e-mail: weiwang@cs.unc.edu

A. Tropsha (✉)
School of Pharmacy, University of North Carolina,
CB#7360 Beard Hall, Chapel Hill, NC, USA
e-mail: alex_tropsha@unc.edu

In this paper we build upon the motif database described in Part I [1] and describe the application of our method to predict protein function from structure for selected examples of targets from structural genomics projects. We show that our novel approach is able to infer function even for orphan proteins, i.e. those that do not resemble any proteins of known function in either sequence or structure.

Materials and methods

Please refer to the companion paper, Part I [1] for a detailed description of the methods used in this paper. For the characterization of the method, we refer frequently to sections (Family Classification Based on Motifs, Motif Library) and tables in Paper I [1], where we list and categorize the protein families under study.

We present results below for two types of graph representations. First, we use graphs where edges are not labeled with distances, referred to as unweighted edge graphs. Second, we use graphs where edges are labeled with distances between respective vertex residues, which are referred to as weighted edge graphs. Motifs mined from these two graph types are called unweighted and weighted edge motifs.

Results: characterization of function inference

We characterize our method in several ways: by examining the distribution of motifs found in the background for a few families from SCOP and EC selected as case studies; by comparing motifs from different graph representations of proteins; by inferring the function of new members added to families in SCOP 1.67 using motifs from the older SCOP 1.65 families to simulate function inference; by checking that the method finds known function similarities and discriminates families that are structurally similar but functionally dissimilar; and in a comparative study of functions inferred by motifs from overlapping SCOP and EC families.

Distribution of motifs in background

In Part I [1] we have described a method of assigning statistical significance to function inference based on the number of motifs found in a target protein, by choosing a suitable cutoff point that minimizes both false positives and false negatives. False positives are proteins in the background that are mistakenly identified as family members since they contain more motifs than the chosen cutoff point, and there is missing or contradictory evidence for the annotation (e.g. alternative functional classification). False negatives are known family members whose function cannot be inferred since they contain fewer motifs than the chosen cutoff point. The numbers of false positives and negatives may be expressed as sensitivity and specificity, and ROC curves can be drawn for each family. From these curves one may determine a cutoff point that misses as few family members as possible (*sensitivity cutoff*) and one that includes no more than 1% of the background as false positives (*99%-specificity cutoff*).

We have compared the distributions of motifs in the background for different types of families: small families against large families; families with many motifs against those with fewer motifs; EC families and SCOP families with a clearly defined single biochemical function against diverse, multifunctional SCOP families and superfamilies; and weighted edge motifs against unweighted edge motifs. We chose the following families for this case study:

1. *Alcohol dehydrogenase* EC family with 15 members
2. *Oxidoreductase (NADP/NADP+)* EC third-level family with 12 members
3. *Amylase* SCOP family with 39 members in SCOP 1.65, 42 in SCOP 1.67
4. *Antibiotic resistance proteins* SCOP family with four members in SCOP 1.65, 7 in SCOP 1.67
5. *Metallo-dependent hydrolase* SCOP family with 17 members in SCOP 1.65, 21 in SCOP 1.67
6. *Haloacid dehalogenases* SCOP family with nine members in SCOP 1.65 and 19 in SCOP 1.67. The SCOP 1.67 family seems to give better motifs.
7. *CheY-related proteins* SCOP family with 12 members in SCOP 1.65 that did not give good motifs, and 17 in SCOP 1.67 that did. Proteins of the CheY and haloacid dehalogenase families have some local structural similarities in the active sites and are suspected to be functionally related [3, 4].

Figures 1 and 2 (and Fig. 1 in the Supplementary material) show the histogram distributions of the number of motifs in the family and in the background for these case studies. Motifs were mined from both the unweighted and weighted edge graph representations of proteins (cf. discussion in Part I [1]) and from version 1.65 of the SCOP database. New family members added in version 1.67 of SCOP were used for method validation since they could be classified as true positives or false negatives. The scale on the X-axis is from zero to the total number of motifs. ROC curves for these families are plotted in the inset of each figure. The Y axes on the histograms and X axes on the ROC curves are plotted on a logarithmic scale for better visibility of smaller bars and high-specificity parts of the curve, respectively.

First, we discuss motifs characteristic of the EC families shown in Fig. 1:

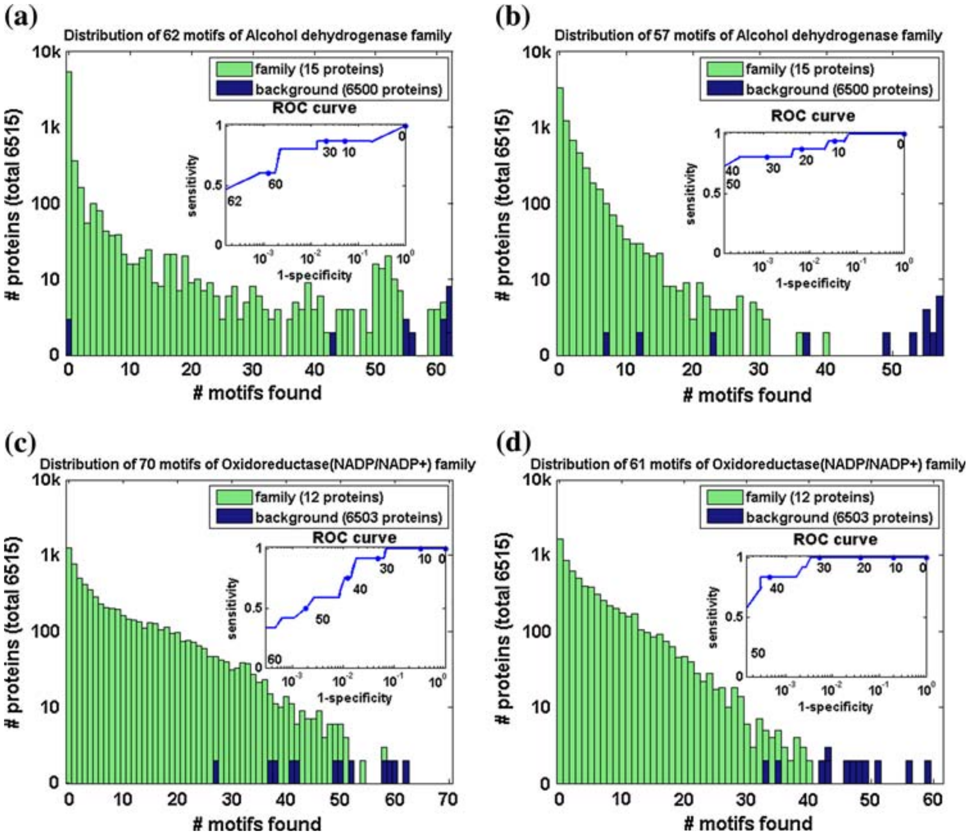


Fig. 1 The distribution in the family (*dark blue*, front) and background (*light green*, back) of unweighted (*left*) and weighted edge (*right column*) motifs for two EC families used in case studies: alcohol dehydrogenase and NADP/NADP+ oxidoreductase. ROC curves are drawn in the inset of each graph, showing sensitivity versus specificity for recognition of the family at different numbers of motifs

1. *Alcohol dehydrogenases* catalyze the extraction of hydrogen from primary, secondary and cyclic alcohols in the presence of NAD(+) or NADP(+) to give aldehydes and ketones. The alcohol dehydrogenase family in our dataset combines two EC numbers differing in the cofactor, i.e. NAD(+)(1.1.1.1) and NADP(+)(1.1.1.2), totaling 15 non-redundant members.

Using protein graph representation with unweighted edges and default subgraph mining parameters, we have identified 62 family-specific motifs. The sensitivity cutoff is set to 43 motifs, while the 99%-specificity cutoff is at 51 motifs. 13 of 15 proteins in the family contain between 43 and 62 motifs, as shown in Fig. 1a, and hence pass the sensitivity cutoff. Two members (1b16A and 1m6hA) share no motifs with the rest of the family. These two proteins belong to a Rossman fold family called “tyrosine-dependent oxidase” in SCOP, while the others belong to the “alcohol dehydrogenase” SCOP family. The functional similarity implied by the shared EC number is lost since the proteins of the alcohol dehydrogenase family outnumber the ones of the tyrosine-dependent oxidoreductase family.

89 proteins in the background pass the sensitivity cutoff, and 61 pass the 99%-specificity cutoff. The five background proteins with largest number of unweighted edge motifs characteristic of alcohol dehydrogenases are 1lluA (all 62), 1h2bA, 1pp2L, 1theA and 1fljA (61 each). 1lluA and 1h2bA are in fact alcohol dehydrogenases according to SCOP, but were not yet added to EC at the time of this case study. The remaining proteins seem to be actual false positives: 1pp2L is a phospholipase, 1theA is a papain-like cysteine protease, and 1fljA is a carbonic anhydrase.

Using protein graph representation with weighted edges and default subgraph mining parameters, 57 motifs were obtained from the same family. Sensitivity and 99%-specificity cutoffs were set at 12 and 17 motifs, respectively. 13 family members contain between 23 and 57 weighted edge motifs, except 1mg5A with 12 motifs and 1hqtA with 7. 161 proteins in the background pass the sensitivity cutoff, while 65 pass the 99%-specificity cutoff. The five background proteins containing the largest number of weighted edge motifs characteristic of the alcohol dehydrogenase family are 1e3jA (40), 1h2bA(37), 1lluA(36), 1lpfA and 1pj5A (31 each). 1e3jA is another SCOP alcohol dehydrogenase unclassified by EC at the time, in addition to 1h2bA and 1lluA identified by unweighted edge motifs. 1lpfA and 1pj5A have FAD/NAD linked reductase domains that are closely related to alcohol dehydrogenases.

2. *Oxidoreductases (NADP)* are a broad functional class at the third level of the EC hierarchy spanning several SCOP families from the NAD(P) binding Rossman and TIM barrel folds: Tyrosine-dependent oxidoreductases (ID: 51751),

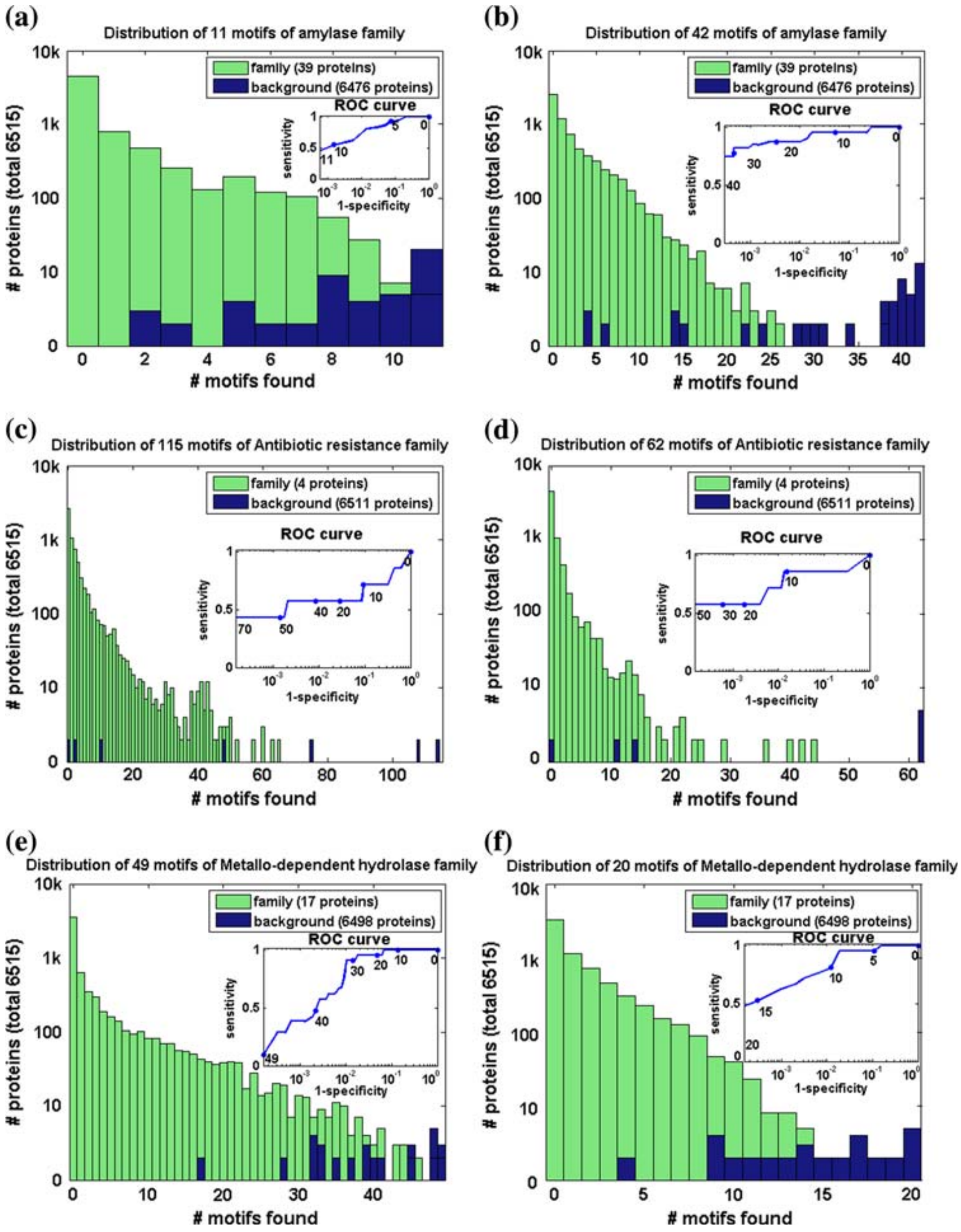


Fig. 2 The distribution in the family (*dark blue*, front) and background (*light green*, back) of unweighted (*left*) and weighted edge (*right column*) motifs for three SCOP families used in our case studies: amylases, antibiotic resistance proteins and metallo-dependent hydrolases. ROC curves are drawn in the inset of each graph, showing sensitivity versus specificity for recognition of the family at different numbers of motifs

Glyceraldehyde-3-phosphate dehydrogenase (GAPDH) like (51800), Aldo-keto reductases/NADP (51431) and FMN-linked oxidoreductases (51396). The family in our dataset contains 12 non-redundant proteins, and was selected as a case study to test if motifs can be derived from families where most members do not share a single global fold.

70 motifs were found to be characteristic of oxidoreductases by subgraph mining using unweighted edges and permissive parameters (*f*0.7, *b*0.15). The sensitivity cutoff point is set to 37 motifs, which identifies 11 family members (missing 1j96A, an aldo-keto-reductase containing 27 motifs), while the 99%-specificity cutoff point is set to 42 motifs, which identifies eight out of 12 members (except 1j96A, 1jw7A, 1p9lA and 1qsgA). 114 and 58 background proteins pass the sensitivity and 99%-specificity cutoffs, respectively. The seven proteins in the background with the most oxidoreductase-specific motifs are 1j3nA and 1ofdA(58), 1nwhA(54), 1l0lB(52), 1hyhA, 1jscA and 4ubpC(51 each). Of these, 1ofdA is an FMN-linked oxidoreductase in SCOP and a glutamate synthase in EC (EC 1.4.7.1), which seem to have closely related functions to oxidoreductases/NADP. 1nwhA is a GAPDH in SCOP and a aspartate-semialdehyde dehydrogenase in EC (EC1.2.1.11), which seems a related function and also uses NAD(P) as a cofactor. 1hyhA is a lactate dehydrogenase, which may also be considered an enzyme with closely related function. The others seem to be false positives: 1j3nA(thiolase-related, 53902), 1l0lB (MPP-like,

63412), 1jscA (pyruvate oxidase/decarboxylase), and 4ubpC (metallo-dependent hydrolase).

61 motifs were identified for this family using protein graphs with weighted edges and parameters *f*0.8, *b*0.1, *d*2. The sensitivity cutoff point and the 99%-specificity cutoff point are both found at 28 motifs, identifying all 12 family members that contain 33–59 motifs. 65 proteins in the background also pass these cutoffs; the five proteins with the most motifs are 1ea0A(47), 1zfjA(43), 1h2bA(40), 1lvl and 1ofdA(39 each). 1ofdA and 1ea0 (FMN-linked reductases, glutamate synthases), 1zfjA (Inosine monophosphate dehydrogenase), 1h2bA (alcohol dehydrogenase) and 1lvl (FAD/NAD linked reductase) are all proteins with related function that the weighted edge oxidoreductase-specific motifs cannot distinguish from EC 1.3.1 family members.

Next, we move on to the SCOP families in our dataset, and compare the distribution of their motifs in the family and in the background.

3. *Amylase* The SCOP family of α-amylases of the TIM barrel fold catalyzes the endohydrolysis of 1,4-alpha-glucosidic linkages in oligosaccharides and polysaccharides. The family in our dataset has 39 non-redundant members in SCOP 1.65, and three more are added in SCOP1.67.

There are 11 motifs mined from the amylase family by setting *b* to 0.1 and the other parameters at default values. The sensitivity cutoff was set at six of 11 motifs, and the 99%-specificity cutoff at 9. As can be seen from Fig. 2, several members of the family contain fewer motifs than the sensitivity cutoff: 1g5aA, 1ktbA, 1kwgA(5 each), 1d3cA(3) and 1fa2A(2). This is consistent with the fact that most members of the SCOP family are α-amylases (EC 3.2.1.1), which is captured by the motifs; other members with different functions share a part of the mechanism for this function, which is captured by them having fewer motifs.

There are 314 proteins in the background that pass the sensitivity cutoff (i.e. contain six or more motifs), and 36 that pass the 99%-specificity cutoff of nine motifs. There are four proteins in the background with all 11 motifs: 1n1tA (sialidase), 1ofdA(FMN-dependent oxidoreductase), 1g9gA (cellulase) and 1h16A (formate acetyltransferase); the first three are different families of hydrolases acting on carbohydrates, and thus they are functionally related to amylases; the fourth is also involved in carbohydrate metabolism. Among three new members of the amylase family added in SCOP 1.67, 1h3gA (cyclomaltodextrinase) and 1r7aA (sucrose phosphorylase) are annotated in GO as having α-amylase activity, and contain eight motifs each, passing the sensitivity cutoff; 1q6cA (β-amylase) contains only two motifs.

42 weighted edge motifs were identified for amylases. The sensitivity cutoff was placed at 14 motifs and the 99%-specificity cutoff at 16 motifs; 111 and 63 proteins in the background, respectively, were found to contain at least these many motifs. Within the SCOP 1.65 amylase family, two proteins (1d3cA and 1fa2A) contain only four motifs, and the rest contain at least 14. Among the proteins added to the family in SCOP 1.67, 1h3gA(38) and 1r7aA(31) are now inferred with 100% specificity; only one background protein not added in SCOP 1.67 (1ua7A) contains 38 motifs, and it turns out to be an α-amylase based on annotations in the PDB file. This new member could have been inferred only weakly using the unweighted edge motifs; with eight of 11 motifs, it would pass the sensitivity cutoff but not the 99%-specificity cutoff.

4. *Antibiotic resistance proteins* The widespread use of antibiotics has created an evolutionary pressure for bacteria to develop resistance to them, and new strains of bacteria have emerged that are resistant to all commonly used antibiotics such as neomycin and fosfomycin [5]. Enzymatic inactivation by several families of enzymes has been observed to be the predominant mechanism of resistance.

The SCOP family of antibiotic resistance proteins chosen for our dataset (ID: 54598) has only four members in SCOP 1.65, and three new members were added in SCOP 1.67. Subgraph mining with *f*0.8 (three out of four) gave 115 motifs, with family members containing 48–114 motifs. The sensitivity and 99%-specificity cutoffs were respectively set at 15 and 39 motifs, admitting 312 and 60 background proteins. The five background proteins with the most family-specific motifs are 1izdA(65), 1cf3A(63), 1p30A, 1p2zA(60 each) and 1pklA(57). These have the following functional assignments—1izdA: pepsin-like aspartic acid protease from *Aspergillus oryzae*; 1cf3A: glucose oxidase from *Aspergillus niger*; 1p2zA and 1p30: structural (capsid) proteins from Adenovirus; 1pklA: pyruvate kinase from *Leishmania mexicana*. Among these, aspartic protease, glucose oxidase and pyruvate kinase are enzymes that may have evolved into antibiotic resistance proteins under evolutionary pressure [5], and thus partial motif overlap is expected. The first two are also from fungi, which produce antibiotics that antibiotic-resistance proteins often evolve to mimic [6]. Viral capsid proteins from adenovirus are structurally similar to those in bacteriophage PRD1, a virus that attacks antibiotic-resistant bacteria [7]; this link, if confirmed, may help explain the evolutionary origins of antibiotic resistance.

The new members added to the antibiotic resistance protein family in SCOP 1.67 contained too few motifs to infer their function: 1nkiA(0), 1npb(10) and 1r9c(2). This hints that motifs mined from three out of four members of a small family are unreliable, as Wangikar et al. [8] have warned.

5. *Metallo-dependent hydrolases (MDH)* This superfamily, originally called amidohydrolases [9], unifies several diverse enzyme families related to urease that share a TIM barrel fold and active site architecture including metal

ion binding site, but do not share detectable sequence similarity. Our interest in this family stems from one of the CASP5/structural genomics targets (PDB:1m65) that we discuss below.

The metallo-dependent hydrolase family in our dataset has 17 members in SCOP 1.65, and 21 in SCOP 1.67. Among the 49 unweighted edge motifs, family members contain between 32 and 49 motifs, with the exception of 1p1m, a "hypothetical" protein from structural genomics with only 17 motifs. The sensitivity cutoff is set to 28 motifs and the 99%-specificity cutoff to 33 motifs; these cutoffs admit 114 and 59 background proteins. The six proteins with the most motifs in the background are 1p9eA (48), 1ed8A(48), 1js8A(46), 1k7hA(45), 1smlA and 1qwnA(44 each). These have the following functional assignments—1p9eA: methyl parathion hydrolase; 1ed8A,1k7hA: alkaline phosphatases; 1js8A: hemocyanin, a Cu^{2+}-binding oxygen transporter protein in molluscs; 1smlA: metallo β-lactamase and 1qwnA: α-mannosidase. Many of these proteins have a function involving metal-ion coordination and phosphate hydrolysis, which are related functions to MDH. The two most interesting cases are 1p9eA and 1qwnA. 1p9eA is assigned to EC class 3.1.8.1 which is named aryldialkylphosphatase, also called phosphotriesterase. Most of the other members of this functional family (e.g. 1psc, 1hzy) are metallo-dependent hydrolases of the TIM barrel fold. Though 1p9e is not classified in SCOP 1.67, DALI shows it as being structurally similar to proteins of the Metallohydrolase/oxidoreductase family that is also included in our dataset (Fig. 3). Similarly, 1qwnA is functionally classified as a mannose (sugar) hydrolase, EC 3.2.1.114; it has a Zn^{2+}-binding site and binds the ligand *N*-acetyl-glucosamine (NAG), similar to many MDHs. SCOP classifies 1qwnA as a new superfamily within the $(\beta\alpha)_7$ fold; it will be shown later that several proteins within this seven-stranded barrel fold have the MDH function.

The four members newly added to the family in SCOP 1.67 contain many family-specific motifs: 1un7A (48), 1rk6A (40), 1ndyA (33) and 1kcxA (32). The first three pass the 99%-specificity cutoff and the last still infers the family function with just under 99% specificity.

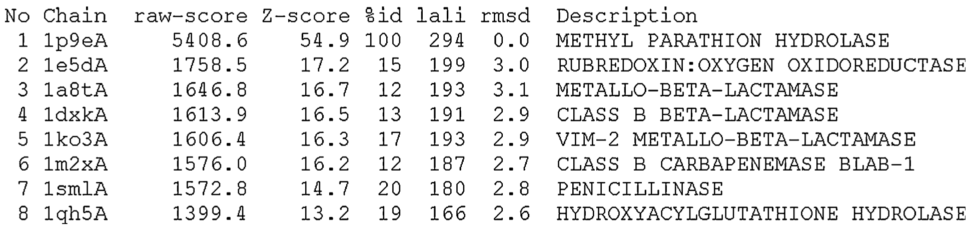
Of the 20 weighted edge motifs of MDH, members of the family in SCOP 1.65 contain between nine and 20 motifs. The sensitivity cutoff is at 11 motifs while the 99% specificity cutoff is at 13 motifs. Two family members contain fewer than 11 motifs: 1j6o(9), a TatD Mg-dependent DNAse from structural genomics, and 1ituA(10), a renal dipeptidase; both form separate families within the MDH superfamily. 125 background proteins also contain at least 11 motifs, while 42 contain at least 13 motifs. The proteins in the background with the most motifs are: 1c96A (16), 1kekA (15), 1xffA, 1oynA, 1a99A and 1bxnA(14 each). These hits have the following functions—1c96A: aconitase/citrate hydro-lyase; 1kek: pyruvate-ferredoxin oxidoreductase; 1xffA: class II Glutamine amidotransferase; 1oynA: CAMP-specific phosphodiesterase; 1a99A: periplasmic binding protein; and 1bxnA: RuBisCo. Apart from the phosphodiesterase, these appear false positives with unrelated function.

Of the four new members added to the family in SCOP 1.67, only one (1un7A) is strongly inferred with 16 motifs; the others, with 9, 9 and 4 motifs, do not pass the sensitivity cutoff.

7. *Structurally distinct superfamilies with similar active site: CheY-like and HAD (haloacid dehalogenase)-like* The large HAD (haloacid dehalogenase) superfamily of hydrolases comprises P-type ATPases, phosphatases, epoxide hydrolases and L2 haloacid dehalogenases [10]. It has been reported that among several families of enzymes structurally similar to the L2 haloacid dehalogenase from *Xanthobacter autotrophicus*, CheY (response regulator protein of bacterial chemotaxis) also has a similar Mg^{2+}-ion binding site [3]. The purpose of including the HAD-like and CheY-like families as case studies is two-fold: to compare their motifs to see if they corroborate the observed functional similarity, and to study two families whose compositions (and hence motifs) have changed drastically between SCOP 1.65 and 1.67. The results may be summarized as follows: the functional similarity does show up in a few members of each family being inferred by the other family's motifs, and the addition of new family members makes little difference to the motifs' strength for HAD, while it is only possible to mine motifs from the CheY families in SCOP 1.67. Further details of this study have been moved to the Supplementary material.

Comparing different graph representations of proteins

Here we compare the motifs mined from different graph representations and evaluate when one is better or which



```
No Chain   raw-score Z-score %id lali rmsd  Description
 1 1p9eA     5408.6    54.9 100  294  0.0  METHYL PARATHION HYDROLASE
 2 1e5dA     1758.5    17.2  15  199  3.0  RUBREDOXIN:OXYGEN OXIDOREDUCTASE
 3 1a8tA     1646.8    16.7  12  193  3.1  METALLO-BETA-LACTAMASE
 4 1dxkA     1613.9    16.5  13  191  2.9  CLASS B BETA-LACTAMASE
 5 1ko3A     1606.4    16.3  17  193  2.9  VIM-2 METALLO-BETA-LACTAMASE
 6 1m2xA     1576.0    16.2  12  187  2.7  CLASS B CARBAPENEMASE BLAB-1
 7 1smlA     1572.8    14.7  20  180  2.8  PENICILLINASE
 8 1qh5A     1399.4    13.2  19  166  2.6  HYDROXYACYLGLUTATHIONE HYDROLASE
```

Fig. 3 Structural similarity of 1p9e, a background protein with all 49 Metallo-dependent hydrolase motifs, to Metallohydrolase/oxidoreductase SCOP family proteins

one should be used more often to find biologically relevant patterns. To this end, we examine and compare the quality of motifs mined from graphs where edges are unlabeled (unweighted) or labeled by Euclidean distances (weighted), over the SCOP and EC families in our dataset.

Families with low sensitivity at cutoff points

The sensitivity of families at their cutoff points, shown in Motif Library tables in Paper I [1], is the fraction of "old" (SCOP version 1.65) and "new" (version 1.67) family members that contain the number of motifs needed for function inference at set thresholds of sensitivity or 99% specificity. The lower the sensitivity, the less robust are the motifs, since only that fraction of new members added to the family in the future are expected to be correctly inferred. Thus, it is of interest to compare the number of families in our dataset where sensitivity at cutoff points is lower than a particular threshold (0.6), when using the unweighted edge vs. weighted edge motifs. We found that eight families had sensitivity below this threshold at their sensitivity cutoff points, while 11 more had it at the 99%-specificity cutoff points as well. Details of this study are in the Supplementary material.

The sensitivity at the sensitivity cutoff point was less than 0.6 for the *weighted edge* motifs of only two SCOP families: α/β knot and Zn-dependent exopeptidase. Additionally, the sensitivity at the 99%-specificity cutoff point was less than 0.6 for the weighted edge motifs of only one adenine nucleotide α-hydrolase family. This analysis suggests that weighted edge motifs afford both higher sensitivity and specificity of function inference that unweighted edge motifs.

Families with very few motifs needed for function inference

A family where a small fraction of the total number of motifs reaches the 99%-specificity cutoff point is said to have strong motifs, since motif matching is more robust. On the flip side, too low a fraction leads to unnecessary long computation time (many more motifs are mined than needed for accurate classification) and poor reliability (false positive matches that could be avoided by setting a higher cutoff).

We compared the number of families having strong unweighted and weighted edge motifs. For unweighted edge motifs, strict (99%-specificity) cutoff points were never less than 10% of the number of motifs, lying between 10 and 25% of the number of motifs for 17 families. In contrast, weighted edge motifs were 99%-specific at *less than 5%* of the total number of motifs for 29 families, at less than 10% for 45 families, and at less than 25% for 94. All these motifs were mined with selective parameters and had high sensitivity at the cutoff point. Further details of this study can be found in the Supplementary Material. This consideration also suggests that weighted edge motifs afford greater number of families with strong motifs.

Families with almost all motifs needed for inference

Using unweighted edge motifs, the following families have their sensitivity (and thus 99%-specificity) cutoff same as the number of motifs: G-proteins (13/13) and PDZ domain (12/12). This means that all the motifs need to be found in a new protein to infer its function. In addition, the following families have their 99%-specificity cutoff points at or near (above 90% of) the total number of motifs: ABC transporter ATPase (43/43), β-Lactamase (20/21), Thioesterase/thiol ester dehydratase/isomerase (10/10), Haloacid dehalogenase/SCOP 1.65 (11/11) and Metalloprotease "Zincin" (27/27). The average sensitivity of these motifs at the 99%-specificity cutoff point was 0.48, i.e. on average, only half of the family members, old and new, contain all the motifs. Requiring a new protein to contain all the motifs for successful annotation leaves no room for structural variations, and immediately precludes the function inference of potential new members similar to the existing members that do not contain all the motifs. Thus, these sets of motifs are not robust.

Using the weighted edge motifs, none of the sets of motifs had either sensitivity or 99%-specificity cutoff points at 80% or more of the number of motifs. The highest 99%-specificity cutoff points were for Adenine nucleotide α-hydrolase (10/13), extended AAA-ATPase domain (10/13) and Antibody constant (C1) domain (9/12). The sensitivity of these motifs at the 99% cutoff point was 0.53, 0.82 and 0.84, respectively; thus, while the first set of motifs seems weak, the others are still usable. These families are omitted from the unweighted edge Motif Library table, since none of them have 10 or more unweighted edge motifs, even when mined with extreme values of parameters (*f*0.7 *b*0.15 *d*3).

The above examples illustrate that a low number of unweighted or weighted edge motifs is usually unstable for function inference. Specificity and sensitivity cutoff points being nearly the same as the total number of motifs is a sign that the motifs are not specific enough. This affects several families with few or many motifs in the unweighted edge representation, and only families with very few motifs in the weighted edge representation.

Strength of motifs at different family sizes

The ratio of sensitivity and specificity cutoff points to the number of motifs (henceforth referred to as *specificity/sensitivity ratio*) serves as a parameter of quality of the motifs, as discussed above. It is correlated with the sensitivity of the motifs to pick up existing and new family members at the cutoff points, but is more informative than the sensitivity; for example, sensitivity is always 1.00 for motifs mined from all members of a family (with *f*1.0), but the motifs are useless if say 5% of the background proteins (330 proteins) also contain all the motifs.

The lower the ratio for the *sensitivity* cutoff point, the more likely the motifs are to find interesting and non-identical members of a family or of other families that have the same function. The lower the ratio for the *99% specificity* cutoff point, the better the chances that one will find these interesting new members without first wading through many false positives with unrelated function.

In Fig. 4, we plot the ratio of sensitivity and 99%-specificity cutoff against family size for unweighted and weighted edge datasets in our motif library from Part I [1]. Motifs for tiny families (3–4 members) and small families (5–7 members) often have the smallest ratios; these families are usually more homogeneous, and also their motifs were typically mined with the most stringent parameters (*f* between 0.8 and 1.0, *b* 0.05 for unweighted edges and 0.01–0.05 for weighted edges).

Overall, unweighted edge motifs have specificity and sensitivity ratios between 0.1 and 1.00, and weighted edge motifs have these ratios between 0.001 and 0.77, as discussed earlier in this section. Also, there is a much larger gap between the sensitivity and 99%-specificity cutoff points for unweighted edges than for weighted edges; on average, the gap between the ratios is 0.2 for unweighted edges and 0.02 for weighted edges.

An interesting fact emerges about tiny families of size 3 on comparing their motifs in unweighted and weighted edge representations. There are four families with 3 members in our dataset that have unweighted edge motifs when mined with *f*1.0; their specificity ratios are 0.25 (EC 1.1.1.82 malate dehydrogenase), 0.38 (Carbon–nitrogen hydrolase), 0.47 (Ubiquinone cyto-chrome-C reductase) and 1.00 (CutA divalent cation tolerance protein). In addition to these four families, another SCOP family with three members (Sec7 domain) has weighted edge motifs. The specificity ratios of these families with their weighted edge motifs are much lower than the corresponding ratios with unweighted edge motifs: 0.01 (Sec7 domain), 0.05 (malate dehydrogenase), 0.09 (CutA and C–N hydrolase) and 0.22 (ubiquinone cytochrome-C reductase). Thus, unweighted edge motifs from tiny families are usually unreliable for function inference, while weighted edge motifs are more reliable.

The comparison shown in Fig. 4 establishes weighted edge motifs as superior to unweighted edge motifs for most families. Most families compared above had both types of motifs, and showed up on both figures. However, there are many families where only weighted edge motifs could be obtained since there were not enough unweighted edge motifs, and a few where only unweighted edge motifs could be obtained since there are too many weighted edge motifs. The quality of motifs varied widely between the two representations, being superior in most cases for weighted edge motifs.

Families with strong weighted edge motifs but weak/no unweighted edge motifs

A list of 16 families (shown in the Supplementary material) had weak or no unweighted edge motifs, and strong weighted edge motifs.

Families with strong unweighted edge but weak/no weighted edge motifs

Only two families from SCOP yielded good motifs in the unweighted edge representation but weak or no motifs in the weighted edge representation: Bacterial luciferase and Metallo-dependent hydrolase. These are both large and diverse SCOP superfamilies. The absence of strong weighted edge motifs indicates that the patterns corresponding to common functional elements are not rigid, or the inter-residue distances between functionally important

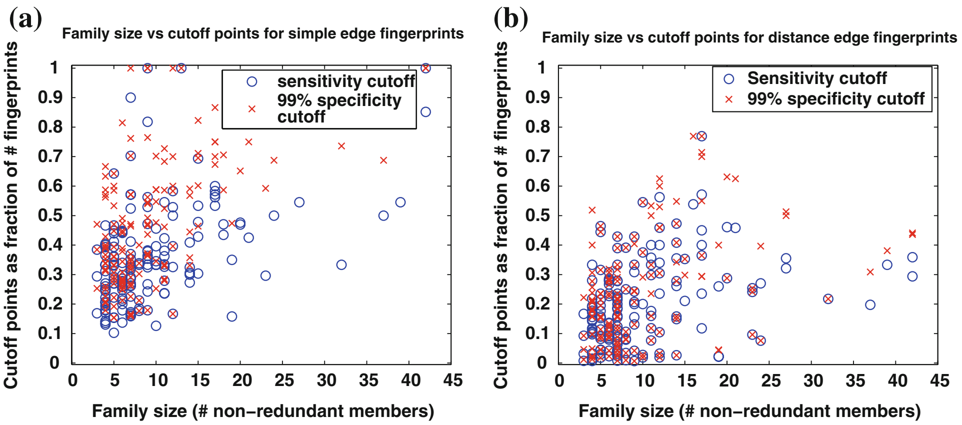


Fig. 4 Comparing the ratio between cutoff points for sensitivity (*blue circles*) and 99% specificity (*red* x's), expressed as fraction of number of motifs, with the family size in our dataset for **a** unweighted and **b** weighted edge motifs

residues are not conserved within the family or needed for the function.

Families with strong unweighted edge but too many weighted edge motifs

Three families yielded good motifs in the unweighted edge representation, but too many motifs to be useful in the weighted edge representation[1]: Enolase (SCOP family, not superfamily) and Aldehyde reductase and Isopropylmalate dehydrogenase (EC). These are all enzyme families with high sequence/structure similarity and hence have few non-redundant members: five for enolase, and four each for the two EC families.

Summary

Weighted edge motifs confer some geometric constraints on the patterns mined, and improve the possibility that the residues included in patterns occurring within a family will superimpose on the corresponding residues in a new member of the family. They have much higher specificity than unweighted edge motifs; many more family datasets were mined with f 1.0, d 0 and b 0.02 or 0.01 in the weighted edge Motif Library in Paper I [1], than in the unweighted edge library. Unweighted edge motifs are more powerful for function inference only for a handful of families where the patterns are flexible, and the connectivity or neighborhood of functionally important residues is conserved but exact geometric arrangements and inter-residue distances are not; weighted edge motifs are not sensitive enough to detect these cases.

With this information in hand, we now evaluate the function inference method, emphasizing assignments made using weighted edge motifs, though considering unweighted edge assignments in special cases where they are more sensitive.

Inference of family members newly added in SCOP 1.67

SCOP families are usually related by evolution, and often by a common function [2], which led us to choose SCOP, a structural classification, to use families for function inference. To test the validity of inferring family membership, we used motifs derived from SCOP 1.65 families to classify proteins newly added to these families in SCOP 1.67 (results discussed only for weighted edge motifs, qualitatively similar for unweighted edge motifs). We count how many new members match enough motifs to pass the sensitivity and 99%-specificity cutoffs.

As a control, we also cross-check the inference by testing new members against the motifs of every other SCOP family in the library. We found that of the 442 new members added to 94 families, 316 (71%) had their function inferred using motifs from the correct family at the sensitivity cutoff, and 284 (64%) at the 99%-specificity cutoff. Most importantly, for 287 (65%) of the new members, among families with motifs above 95% specificity, the correct family was the choice with the 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 with sequence identity at least 40%, the threshold suggested for inferring function from sequence [11]. Detailed results of the SCOP validation experiment are presented in Tables 4–6 in the Supplementary material.

Discrimination of similar structures with different function

Mutual discrimination of TIM barrel families

The Triosephosphate Isomerase (TIM) fold is a 8-stranded $\alpha\beta$-barrel fold that is one of the most versatile folds known, and serves as a generic scaffold for up to 23 distinct functions spanning all five classes of the EC enzyme classification [12]. Different functional families within the TIM barrel fold are so structurally similar that sequence signatures are often unable to distinguish between them, and structural searches with DALI [13] often generate hits to members of other families with higher z-scores than members of the same family. We checked if motifs from superfamilies and families in the TIM barrel fold discriminate amongst the families, i.e. if motifs of each family inferred that function with higher confidence in members of that family than the motifs of other structurally similar families.

We tested the entire non-redundant set of 284 proteins adopting the TIM barrel fold in SCOP 1.65 against motifs derived from 20 (super)families within this fold with both unweighted and weighted edge motifs. Since the results were qualitatively similar, only weighted edge results are discussed here.

We find that the average member of any of these families matches 70–90% of its own family motifs, and 0–40% of any other family's motifs. Exceptions arise from superfamily-subfamily pairs that share motifs since their members overlap, and from families that do not have highly significant motifs. These results are shown graphically in Fig. 2 in the Supplementary material, where Table 7 lists the 20 families, showing the number of non-redundant members and motifs in our Motif Library [1], and a three-letter abbreviation to represent the family.

[1] over a thousand, even with restrictive mining parameters such as f 1.0, b 0.01, d 0.

Structural similarity between active sites without functional similarity

It has been reported [14] that the active sites of influenza virus sialidase (EC 3.2.1.18, a glycosylase/glycosidase that hydrolyzes glycosidic linkages in oligosaccharides, glycoproteins and other related compounds) and Escherichia Coli isocitrate dehydrogenase (ICDH, EC 1.1.1.42, a NADP-linked reductase), share some three-dimensional clusters of residues. This similarity was found by searching the PDB for a graph pattern derived from the literature on the influenza virus sialidase [15] using the ASSAM program [16]. Since the EC numbers are very different, the similarity is structural, not functional. The authors also suggest that the pattern only matches a part of the active site, and its existence does not imply that the sialidase would bind isocitrate.

To verify if motifs capture the functional difference, we derived motifs for both these families from EC, as reported in the Motif Library tables in Part I [1]. Of the 18 weighted edge sialidase motifs, on average 2 and at most 4 appeared in the 4 ICDH proteins; of the 28 weighted edge ICDH motifs, on average 1 and at most four motifs appeared in sialidases. On the basis of these motifs, none of the proteins in these two families inferred the function of the other family. We conclude that influenza virus sialidases and isocitrate dehydrogenases share some similarity, but this similarity is not large enough to imply similarity of function, corroborating the earlier suggestions in the literature [14].

Comparing families from EC and SCOP classification

To check whether the two different classification systems used agree on the definitions of functional families, we inferred the function of proteins from EC families using the SCOP family motifs, and vice versa, and found that the two agreed quite well. We compared a few EC families and their corresponding SCOP families to check how many motifs of one occur in the members of the other, and the percentage of members of one whose family may be inferred by motifs of the other. The families discussed include aldolases, carbohydrate phosphatases, isocitrate/isopropylmalate dehydrogenases and tryptophan synthases. More details of these experiments are given in the Supplementary Material. We conclude that motifs from a structural classification (SCOP) can help infer function equally well as those from a functional classification (EC); that corresponding families from both classifications give motifs that infer function in each other's members; and that structurally similar SCOP families with different EC numbers, or a single SCOP family split over two EC numbers, can be distinguished using motifs. These observations validate our decision to use families from SCOP, a structural classification, for function inference.

Results: function inference on structural genomics targets

In Materials and Methods we have described the collection of a set of proteins characterized in the Structural Genomics projects, and their classification into those with known function, global structural similarity to proteins of known function, and no global structural similarity to proteins of known function (dubbed "structural orphans"). We used our method and the Motif Library described in Part I of this study [1] to suggest functional assignments for proteins in the last two categories, as reported in tables 9-12 in the Supplementary Material. There we also corroborate the observation of Aloy et al. [17] that functional assignments from structural motifs are more accurate than those from the most similar sequence or structure alone. An illustrative example is protein 1r3d in the PDB, annotated correctly as a carboxylesterase using motifs, but incorrectly using DALI [13].

Discussion of selected inferences

We discuss several case studies of function inferences for structural genomics targets of unknown and known function, some of which are mentioned in Tables 9-12 in the Supplementary Material. For selected inferences, we show color-coded pictures of the residues covered by motifs.

Shikimate dehydrogenase (independent biochemical characterization) The EC family of Shikimate 5-dehydrogenases (1.1.1.25) has five members and 114 weighted edge motifs, 33 of which occur in a protein 1npy that is labeled as a hypothetical shikimate 5-dehydrogenase-like protein of unknown function. Prior to 2004 the function of this protein could be inferred by high structural (but not sequence) similarity to known members of EC 1.1.1.25, but the possibility of loss in function due to functional residue mutation made this inference doubtful. The strong inference from motifs (100% specificity, with over 3 times more motifs than contained in other background proteins) reaffirms the function that was inferred from structure. Most excitingly, independent biochemical investigations have identified HI0607 as belonging to a new class of shikimate dehydrogenases [18], which provides preliminary experimental validation of our function inference.

SH3 domains, that mediate intracellular protein-protein interactions by recognizing a proline-rich motif, are a known hard case to annotate by sequence or structure based methods, since most of them do not have appreciable sequence identity to each other; they are essentially a family of remote homologs. It is difficult to find many common subgraphs in this family without merging the labels, which again would make any patterns lose their uniqueness. In spite of this, we were able to obtain 17 motifs from this family by relaxing the conditions to $f = 0.7$ and $b = 0.1$; thus they are classified as a "poor" family.

We tried to annotate 15 structural genomics targets that have the keyword "SH3" in their PDB headers, and some of which are of unknown function: 1j0f, 1oot, 1spk, 1ssh, 1tg0, 1uff, 1ug1, 1ugv, 1uhc, 1ujy, 1va7, 1wfw, 1wi7, 1wie and 1wry. Five of these 15 proteins: 1ssh(14), 1oot, 1uhc(11 each), 1wi7(9) and 1va7(5) were inferred with 99% specificity above the cutoff point of five motifs out of 17, while the remaining 10 proteins matched between 0 and 3 motifs and were not inferred.

CASP Target T0147, E. coli YcdX, PDB 1m65 is classified as a PHP domain (ID: 89551) in SCOP, based on detailed analysis of the domain conservation of two distinct classes of DNA polymerases [19]. Potential homology to the TIM barrel superfamily of metallo-dependent hydrolases was suggested by the authors of the structure based on a conserved metal-binding motif (HXH) and threading [20]. Metal-dependent hydrolases such as cytosine deaminase (1k6w) form a distorted 8-stranded TIM barrel capped by a C-terminal helix similar to that of the target structure, though the target structure is composed of only seven strands. Thus the target structure was classified as an analog, not a homolog of metallo-dependent hydrolases [21]; i.e., its function remains unknown.

Using unweighted edge motifs, 30 of the 49 motifs for metallo-dependent hydrolases are found in 1m65, inferring its function as a metallo-dependent hydrolase with 98.6% specificity and 90% sensitivity. Subsequently, the residues covered by motifs are plotted in VMD [22] using the "surface" representation, and are color-coded as blue (basic amino acids, e.g. His), red (acidic, e.g Asp), magenta (polar, e.g. Thr), and white (hydrophobic, e.g. Leu). These plots are shown in Fig. 5a–d for both 1m65 (the YcdX protein) and in 1nfg (d-hydantoinase, a typical metallo-dependent hydrolase). The figure shows that the volume of residues covered by motifs has roughly the same geometric shape and electrostatic/chemical properties in the two proteins, and thus strengthens the inference that 1m65 has a metallo-dependent hydrolase function, even though it is not in the SCOP metallo-dependent hydrolase family.

Apart from the suggestions by the authors and the CASP5 target classifiers, some other studies also indicate that this inference is valid. For example, the PINTS-weekly service [23] finds active site patterns from many metallo-dependent hydrolases in this protein. The ProFunc meta-server found weak active site and ligand matches and strong binding site matches with metallo-dependent hydrolases. Finally, GenProtEC, the E. coli genome and proteome database [24] 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 [25, 26].

Some other established studies and papers initially seem to oppose this function inference. The most prominent is the GO function annotation of 1m65 on the PDBsum site as having *DNA-directed DNA polymerase activity* (GO: 0003887); however, this function assignment is putative and is made on the basis of electronic annotation transferred from InterPro, a sequence database. Thus, the evidence for the GO annotation is Inferred from Electronic Annotation (IEA), the least reliable evidence code. The discoverers of the PHP domain sequence family [19] indicated shared active site motifs between the metallo-dependent hydrolase family and the PHP-domain family, and hypothesized that bacterial and archaeal DNA polymerases possess intrinsic phosphatase activity that hydrolyzes the pyrophosphate released during nucleotide polymerization. Thus, the assigned GO term does not contradict the function inferred by motifs.

Function inferences using the Metallo-dependent hydrolase motifs for proteins of known function have been discussed already in our earlier case studies. Those included all 4 proteins added to this family in SCOP 1.67 (1ndy, 1kcx, 1un7, 1rk6) as well as proteins structurally similar to other folds but with a phosphotriesterase function shared by many metallo-dependent hydrolases of the TIM barrel fold(1p9e). This group also included additional phosphodiesterases with a distorted TIM-barrel fold that form new superfamilies of the SCOP 7-stranded β-barrel fold along with the PHP domain of 1m65.

The weighted edge motifs of metallo-dependent hydrolases fail to infer the function for 1m65. Reasons for this may be that some edges within this family's unweighted edge motifs vary in length, straddle edge length bin boundaries or change from contact edges to distance constraints across the family, all of which would preclude identification of weighted edge motifs. The weighted edge motifs of MDH have been shown to be weak in their specificity; they also occur in other TIM barrel families such as pyruvate kinase, which makes specific inference of this family's function unreliable. Incorporation of subsequent improvements to the weighted edge matching algorithm to accommodate overlapping edge length bins [27] may alleviate this problem.

Yyce from Bacillus subtilis, PDB 1twu is unclassified in both SCOP 1.65 and 1.67. At the time of SCOP version 1.65 release, 1twu was an orphan structure with no structural similarity to proteins of known function. Function inference using weighted edge motifs mined from SCOP 1.65 families showed 46 of 62 motifs from the Antibiotic Resistance protein family (SCOP ID: 54598), inferring the

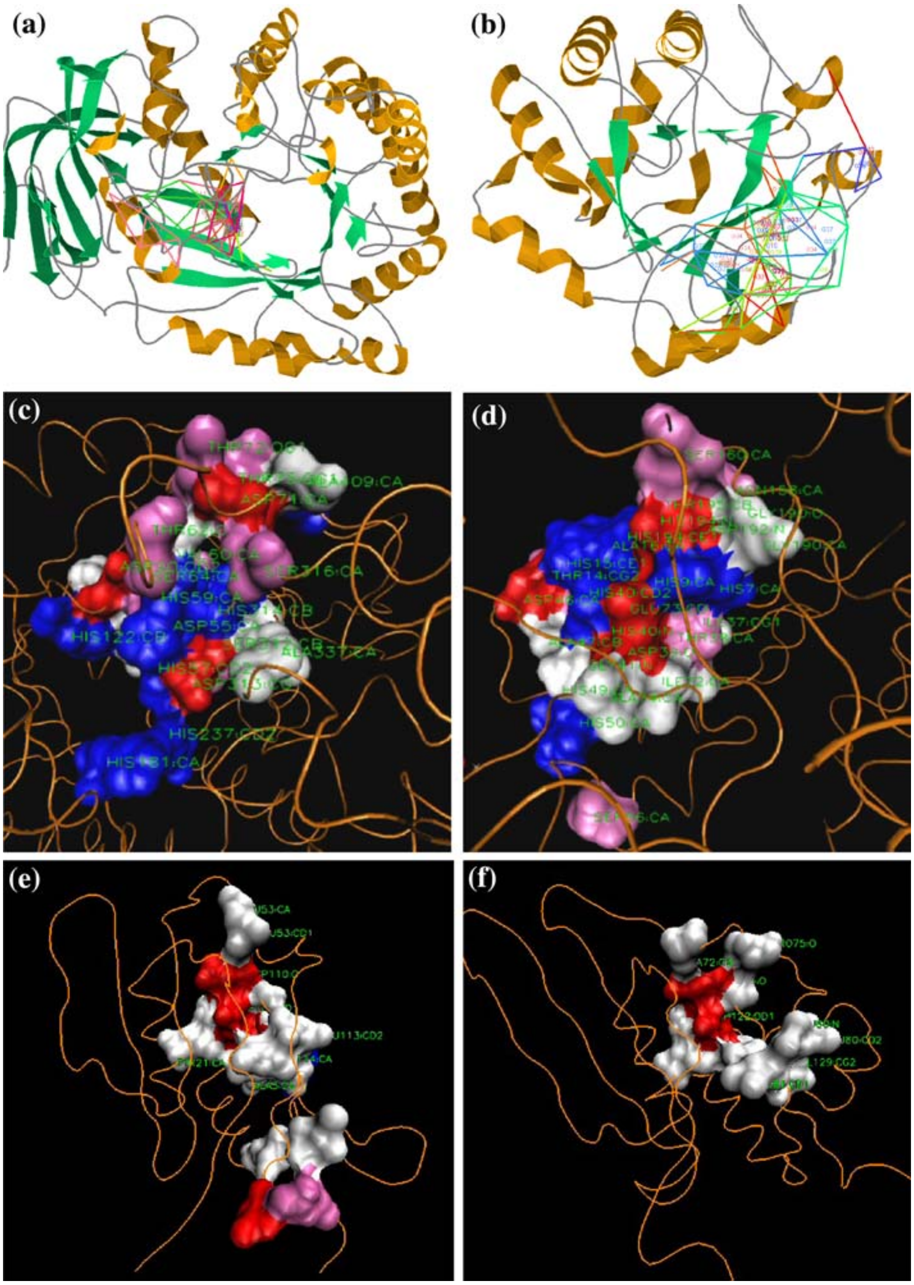


Fig. 5 Examples of function inference: residues covered by metallo-dependent hydrolase (MDH) motifs in 1nfg, an MDH (**a**, **c**), and in 1m65, the YcdX protein with unknown function (**b**, **d**). **a** and **b** show the actual subgraphs for the motifs of MDH found in 1nfg and 1m65, plotted as edges between corresponding C_αs and viewed superimposed on the protein structure using KiNG [29]. **c** and **d** show the same proteins displayed in VMD [22] with the residues covered by motifs plotted as residue surfaces, and color-coded based on electrostatic and chemical properties: *white* hydrophobic (VAILMGPFW); *magenta* polar (CSTYNQ); *red* acidic (DE); and *blue* basic (RHK). A second example of function inference is shown in (**e**, **f**): residues covered by antibiotic resistance family (SCOP: 54598) motifs in **e** 1ecs, an antibiotic resistance protein in SCOP 1.65, and **f** 1twu, the Yyce protein with unknown function that has structural similarity only to newly added proteins in SCOP 1.67. **c**)–**f** Adapted from [28]

antibiotic resistance function with specificity 100%. Figure 5e and f shows the residues covered by motifs in 1ecs, an antibiotic resistance protein in SCOP 1.65 and in 1twu. Note the geometric and electrostatic similarity between the upper region covered by motifs in 1ecs and the one in 1twu.

The Antibiotic Resistance protein family had only 4 non-redundant members in SCOP 1.65, but the motifs derived from both unweighted and weighted edges had good specificity and sensitivity for function inference. Thus, the confidence in the inference is very high. Re-examining the structural similarity of protein 1twu to all known proteins using the version of the DALI FSSP database that was current in May 2005, we found it was similar to a protein 1nki that was unclassified in SCOP 1.65 but was added to the Antibiotic Resistance protein family in SCOP 1.67. Thus, this case study provides another firm evidence that our approach could provide accurate function inference when the confidence in the prediction is high.

Discussion

Our method of using family-specific motifs to infer protein function was designed to be maximally robust: the graph construction based on almost-Delaunay tessellation of protein structure takes into account natural imprecision in coordinates, and using multiple subgraph motifs accommodates representation errors, (limited numbers of) missing or substituted residues, and structural flexibility. The method is designed to find information that is not implied by sequence patterns, structural alignments, and known functional site templates. Thus it may succeed where other methods fail, or be profitably used in combination with other methods in consensus prediction.

The successful function inference for new members of SCOP families confirms the predictive power of motifs; the success rate of 65% for choosing the correct family for proteins added to the SCOP classification in version 1.67 vs. 1.65 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 for only 53% of the added proteins.

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 [28], indicate that family-specific motifs do capture function-related rather than shared structural information. We have seen that the motifs 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.

The designed robustness of our method suggests that it could be potentially used to predict function from sequence, using either accurately predicted structures, or sequence patterns derived from structural motifs with preserved sequence order within a family. We have obtained preliminary results of function prediction using predicted structures and models as well as sequence motifs, and further investigations in these directions are ongoing.

In conclusion, the method described in this and accompanying [1] papers identifies packing patterns characteristic of functional families having four or more proteins with known 3D structure, and uses them to infer function of new members of these families. Structural errors, missing fragments or mutations may lead to failure of motif mining or function inference. Careful manual selection of families and fixing errors in structure files should improve the results further. Since our method is capable of inferring function for many orphan proteins, the ultimate proof will come from experimental validation of its predictions.

Acknowledgments The authors gratefully acknowledge support from NIH grant GM068665 and NSF grant CCF-0523875.

References

1. Bandyopadhyay D, Huan J, Prins J et al (2008) Identification of family-specific residue packing motifs and their use for structure-based protein function prediction: I. Method development. J Comput Aided Mol Des. doi:10.1007/s10822-009-9273-4
2. Murzin AG, Brenner SE, Hubbard T et al (1995) J Mol Biol 247:536
3. Ridder IS, Dijkstra BW (1999) Biochem J 339(2):223
4. Meng EC, Polacco BJ, Babbitt PC (2004) Proteins 55:962
5. Burk DL, Ghuman N, Wybenga-Groot LE et al (2003) Protein Sci 12:426
6. Fong DH, Berghuis AM (2002) EMBO J 21:2323
7. Benson SD, Bamford JK, Bamford DH et al (1999) Cell 98:825
8. Wangikar PP, Tendulkar AV, Ramya S et al (2003) J Mol Biol 326:955
9. Holm L, Sander C (1997b) Proteins 28:72
10. Koonin EV, Tatusov RL (1994) J Mol Biol 244:125
11. Wilson CA, Kreychman J, Gerstein M (2000) J Mol Biol 297:233
12. Nagano N, Orengo C, Thornton J (2002) J Mol Biol 321:741
13. Holm L, Sander C (1996) Science 273:595
14. Poirrette AR, Artymiuk PJ, Grindley HM et al (1994) Protein Sci 3:1128
15. von Itzstein M, Wu W, Kok G et al (1993) Nature 363:418
16. Artymiuk PJ, Poirrette AR, Grindley HM et al (1994) J Mol Biol 243:327
17. Aloy P, Querol E, Aviles FX et al (2001) J Mol Biol 311:395
18. Singh S, Korolev S, Koroleva O et al (2005) J Biol Chem 280:17101
19. Aravind L, Koonin EV (1998) Nucleic Acids Res 26:3746
20. Teplyakov A, Obmolova G, Khil PP et al (2003) Proteins 51:315
21. Kinch LN, Qi Y, Hubbard TJ et al (2003) Proteins 53(Suppl 6):340
22. Humphrey W, Dalke A, Schulten K (1996) J Mol Graph 14:33
23. Stark A, Shkumatov A, Russell RB (2004) Structure (Camb) 12:1405
24. Serres MH, Goswami S, Riley M (2004) Nucleic Acids Res 32:D300
25. Gough J, Chothia C (2002) Nucleic Acids Res 30:268
26. Madera M, Vogel C, Kummerfeld SK et al (2004) Nucleic Acids Res 32:D235
27. Huan J, Bandyopadhyay D, Snoeyink J et al (2006) In: IEEE computational systems bioinformatics conference (CSB). Stanford, CA, USA
28. Bandyopadhyay D, Huan J, Liu J et al (2006) Protein Sci 15:1537
29. Davis IW (2001) Kinemage, next generation. http://www.kinemage.biochem.duke.edu/software/king.php