Functional Neighbors: Inferring Relationships between Non-Homologous Protein Families Using Family-Specific Packing Motifs

Deepak Bandyopadhyay*
GlaxoSmithKline, 1250 S.
Collegeville Rd, Mail code
UP12-210, Collegeville PA
debug22@gmail.com

Jun (Luke) Huan
Department of Electrical
Engineering and Computer
Science, University of Kansas
jhuan@eecs.ku.edu

Jinze Liu
Department of Biostatistics
University of North Carolina
at Chapel Hill
liuj@cs.unc.edu

Jan Prins, Jack Snoeyink, Wei Wang
Department of Computer Science
University of North Carolina at Chapel Hill
{prins, snoeyink, weiwang}@cs.unc.edu

Alexander Tropsha
Division of Medicinal Chemistry and Natural
Products, School of Pharmacy, University of North
Carolina at Chapel Hill, tropsha@email.unc.edu

Abstract

We describe a new approach for inferring the functional relationships between non-homologous protein families by looking at statistical enrichment of alternative function predictions in classification hierarchies such as Gene Ontology (GO) and Structural Classification of Proteins (SCOP). Protein structures are represented by robust graphs, and the Fast Frequent Subgraph Mining algorithm is applied to protein families to generate sets of family-specific packing motifs, i.e. amino acid residue packing patterns shared by most family members but infrequent in other proteins. The function of a protein is inferred by identifying in it motifs characteristic of a known family. We employ these family-specific motifs to elucidate functional relationships between families in the GO and SCOP hierarchies. Specifically, we postulate that two families are functionally related if one family is statistically enriched by motifs characteristic of another family, i.e. if the number of proteins in a family containing a motif from another family is greater than expected by chance. This function inference method can help annotate proteins of unknown function, establish functional neighbors of existing families, and help specify alternate functions for known proteins.

Introduction

Structural genomics projects generate many new protein structures, including hypothetical proteins from fully- sequenced genomes with unknown function. These developments underlie a need for powerful and reliable function inference methods. In an earlier short paper[4], we described a method for inferring protein function using *family-specific packing motifs*, which are 3D residue packing patterns automatically mined from graph representations of protein families. In contrast to traditional function inference methods that rely on sequence or fold comparison, our approach relies on the local residue patterns (or motifs) as possible determinants of protein function.

In this paper, we describe how family-specific motifs can be used to discover hidden connections between protein families with no apparent fold and sequence similarity, i.e. remote homologs. To this end, we introduce a new measure of functional similarity between families based on statistically significant enrichment of a family with motifs characteristic of another family. We present several case studies demonstrating that our approach could correctly predict functional similarity between remotely homologous protein families.

Related work

We review recent algorithms to discover important local features and link protein families using shared features. Protein function annotation using local structure features is known to be more accurate than using only sequence or structure alignments[1]. The following methods have been proposed for finding local structural motifs in protein families, and known functional sites in protein structures:
— *Depth-first search* starting from simple geometric patterns such as triangles, and progressively finding larger patterns [27, 29, 8, 17].
— *Geometric hashing* can compare two protein struc-

tures [24] or a structure to a database [6].
— *Functional site template* methods represent functional sites as pockets [7], clefts [19], or patches [9], and match them with new protein structures using geometry, conserved residues and electrostatic/chemical properties.
— *String pattern matching* uses string search algorithms on encoded local structure/sequence [27, 31].
— *Graph matching* methods have been developed to compare protein structures modeled as graphs, usually with clique detection techniques [2, 21, 30, 36, 12].
— *Other methods* include inductive programming language [32], fuzzy functional forms [10], computed protonation properties[22] and geometric depth potentials [39].
— *Hybrid methods*, e.g. clique hashing[37].

Several graph representations of protein structure have been developed, with nodes ranging from secondary structure segments [2] to atoms [16]. Our recent work [14] explored Delaunay tessellation as a means to generate a sparse graph representation of a protein structure. Later we introduced *almost-Delaunay* edges to account for imprecision in atomic coordinates by using a parameter ϵ[5]. Almost-Delaunay edge graphs are much sparser than contact distance graphs, since they remove pairs of atoms whose interaction is occluded by other closer atoms. Unlike Delaunay graphs, they are robust in the presence of coordinate perturbations[14], and thus allow us to find frequent patterns from large and diverse protein families with a lot of variation in the coordinates.

The problem of *Frequent Subgraph Mining* is to identify all frequent subgraphs for a set of graphs G, where a subgraph must occur in more than some fraction of G called the *support* (σ) to be considered frequent. Our subgraph mining method[15] builds frequent subgraphs directly using a tree representation, and thus is faster and applicable to larger structures and databases than exhaustive enumeration of subgraphs by depth-first search [18].

There have been recent efforts towards annotation of protein structures (and homology models built from sequences) using functional signatures derived from structural alignments [35], overlapping sphere representations of functional sites [12, 38], and clusters of functionally important residues determined by predicted protonation properties [22] or a geometric depth potential [39, 40], to name just a few. Our method, unlike the first [35], does not depend on a sequence or structure alignment, and can find motifs not conserved in the sequence. It differs from the second [38] in that functionally important residues used in graph patterns are inferred from protein families rather than chosen manually from the literature or bound ligand positions. It distinguishes itself from the other methods mentioned [22, 39] by insisting that the motifs found and used for annotation be unique to each family. Remote similarities found using family-specific motifs are thus more significant than single binding sites matched by the other methods, since the same site can be involved in multiple functions[25].

Lastly, we study the problem of automatically finding functional relationships between families that are not related by sequence or structure, i.e. finding remote function similarity or functional neighbors. This has been reported for pairs of known families [26], for pairs of individual proteins[20], for a predefined set of structural patterns and a protein[12, 38], and recently, for a computed set of functional sites and the Protein Data Bank [40]. Our method is critically different in that it compares and relates protein families, rather than pairs of proteins or a protein database and a functional site database, using local structure patterns specific to protein families. Comparing the set of features that discriminate a family, rather than individual features that superpose well on a pair of protein structures, makes our conclusion robust and statistically sound.

Materials and Methods

Our method initially finds and calibrates motifs using the FFSM subgraph mining program (`http://www.cs.unc.edu/~huan/FFSM/`). We briefly describe below five steps of the procedure discussed in detail in our previous papers[14, 4], followed by an in-depth discussion of the additional step of enrichment evaluation which is the major new development reported in this paper.

1. Select families of non-redundant proteins from a classification database such as SCOP/EC, or as defined by the user. Also, define *background* dataset to represent all protein structures. We chose 29 EC families and 125 families from SCOP version 1.65, which was current at the time we initiated these studies. Our background dataset used PISCES[34] with sequence identity $\leq$ 90%, resolution $\leq$ 3 Å , and R-factor $\leq$ 1.0, which led to 6625 valid chains.

2. Represent protein structures as graphs, with nodes at each residue, and contact between residues defined using the *almost-Delaunay*[5] edges. It is possible to merge two or more node types to create a reduced set of node labels. We add length-dependent edge labels and distance constraint edges between non-contacting residues to ensure consistent geometry in patterns[13].

3. Mine family-specific motifs using the Fast Frequent Subgraph Mining method[14]. Subgraphs are defined as family-specific motifs if they occur in at least 80% of the family (support), and in at most 5% of the background (background occurrence). If the background check step is omitted, the patterns are merely called frequent subgraphs or spatial motifs.

4. Search for motifs in a new structure, using a graph similarity index to speed up subgraph isomorphism [33].

5. Assign a significance to the function inference, by counting family motifs found in step 4 and examining their distribution in background proteins.
6. Calculate statistical enrichment of motifs in nodes of the SCOP and GO hierarchies, using the hypergeometric distribution with a p-value cutoff of 10^{-6}.

Enrichment evaluation in SCOP and GO

SCOP enrichment evaluation aims to determine if the set of background proteins containing a large fraction of motifs from a SCOP/EC family is enriched with proteins from some other SCOP family, i.e. if there are more proteins of the set from that other family than would be expected by chance.

While checking in the background for occurrence of motifs, proteins containing each motif are extracted into a list, and these lists are used to evaluate enrichment in the SCOP hierarchy. A geometric distribution is used to model the probability that from n proteins sharing the same motif by chance, at least k proteins will belong to a category (i.e. SCOP family) containing f proteins, from a total protein data bank size of g. The $P\text{–}value$ is given by $P = 1 - \sum_{i=0}^{k} \frac{\binom{f}{i}\binom{g-f}{n-i}}{\binom{g}{n}}$. The hyper-geometric distribution [8] is also commonly used: given a collection of representative proteins M, a subset of proteins $T_{motif} \subseteq M$ sharing one common spatial motif, and a subset of proteins $T_{class} \subseteq M$ of a predefined category, the probability of observing a subset of proteins $K \subseteq T_{motif} \wedge T_{class}$ with at least size k is given by

$$P - value = 1 - \sum_{i=0}^{k-1} \frac{\binom{|T_{motif}|}{i}\binom{|M|-|T_{class}|}{|T_{motif}|-i}}{\binom{M}{T_{motif}}}. \qquad (1)$$

For example, it is unlikely that the majority of a group of proteins sharing a motif come from a single SCOP superfamily, so such a category would be statistically significant, with a small P-value close to zero.

We adopt the Bonferroni correction for multiple independent hypotheses[28], $0.001/|C|$ where $|C|$ is the set of categories, as the default threshold for significance of the P-value of individual tests. With $|C| \approx 1300$ SCOP superfamilies, we chose the P-value threshold for significant function similarity between 10^{-6} and 10^{-8}.

Enrichment evaluation helps to verify that motifs from a SCOP family do not occur in too many other families; to find related families and superfamilies on the basis of shared motifs; and to correlate EC classes to SCOP structural families that share their motifs. This analysis enables the characterization of false positive motif matches — proteins not in the same functional family but inferred with high confidence — to determine if they are random or perform a related or unrelated function. Families that are highly enriched with another family's motifs but are not near that family in the SCOP hierarchy are denoted *functional neighbors*; we hypothesize that such families have related functions, or they share some aspects of function.

The Gene Ontology (GO, [11]) provides a controlled vocabulary for describing protein function. GO terms form directed acyclic graphs(DAGs) connected by relationships such as "is-a" and "part-of". Terms at lower depth in the DAGs describe more general functions; the greater the depth, the more specific is the function.

Enrichment evaluation on the GO hierarchy is similar to that described for SCOP; it aims to determine whether the set of proteins sharing the motifs of a SCOP/EC family is enriched with proteins from a particular functional category (i.e. GO term) to a greater extent than would be expected by chance. Combining GO and SCOP enrichment evaluations, we can test the hypothesis that two SCOP families marked as functional neighbors have related functions.

Results: Enrichment and functional neighbors

To characterize the hits for EC family-specific motifs in SCOP, and to evaluate the functional roles of proteins returned as (potentially false) positives by our annotation method based on local structure patterns, we study the enrichment of all motif hits in the background within the SCOP and GO hierarchies. As an example we have looked at the distribution of GO functions for proteins containing motifs for the serine protease family in SCOP. We extracted all background proteins containing each of the 72 serine protease motifs, and evaluated these lists for GO enrichment. The number of background hits per motif ranged from 60 to 97. The GO categories related to peptidase activity (Figure 1) were consistently enriched in the protein lists for all 72 motifs, with $P\text{–}value < 10^{-15}$.

```
GO:0008233 : peptidase activity (k:55/f:377)
      GO:0004175 : endopeptidase activity (k:55/f:297)
            GO:0004252 : serine-type endopeptidase activity (k:55/f:130)
                  GO:0004263 : chymotrypsin activity (k:52/f:73)
                  GO:0004295 : trypsin activity (k:55/f:85)
      GO:0008236 : serine-type peptidase activity (k:55/f142)
            GO:0004252 : serine-type endopeptidase activity (k:55/f:130)
```

Figure 1. Significantly enriched GO categories for the 62 background hits for one motif of serine proteases. In each GO category, k is the number from the 62 hits and f is the number of background proteins.

The above result suggests that proteins sharing the same motif have similar functions. It follows that any potential false positives from our annotations of serine proteases – which are precisely the background proteins that contain multiple motifs – belong to related families that share some functional similarity with serine proteases. This observation leads to the definition of a *functional neighbor* relation between families that share motifs, with strength proportional to the number of motifs shared. Our definition is broader than EC number similarity (which is for enzymes only), and addresses the EC system's historical inconsistencies and other shortcomings that make it not well-suited for function inference[3].

Functional neighbors of a family provide a set of hypothetical additional functions that could be experimentally validated. They may also be used to classify families of proteins with an unknown function, by deriving a function annotation based on their functional neighbors. Note that the functional neighbor relation as defined is not symmetric; i.e. if family A's motifs are enriched in family B, it does not imply that B's motifs are enriched in A.

Most families share some motifs with their subfamilies, superfamilies or siblings in SCOP, and these are structural neighbors as well as functional neighbors. Some other families share motifs with families in a different branch of the SCOP hierarchy, and thus not obviously related to them; such families are functional neighbors but not structural neighbors, at least in SCOP.

Table 1 shows some SCOP families from our current dataset that share motifs and are functional neighbors, but are not structural neighbors. We calculate these family pairs (motifs of F, enriched in families E_i) by finding all proteins in the background dataset having enough motifs of F that their function can be inferred with 99% specificity, and using them as input to the SCOP enrichment method. We report families E_i that (1) have p-values $< 10^{-7}$ for the enrichment, (2) are at superfamily or family level nodes of SCOP, (3) do not share a parent/child or sibling relationship with F, and (4) for at least a fifth of their members the function F is inferred with 99% specificity.

The choice of 10^{-7} for p-value cutoff highlights strong relationships, and some known functional neighbors with larger p-values are hidden. Also, the restriction to use SCOP families rather than EC families hides the functional neighbor relationship between alcohol dehydrogenases (EC family in our dataset) and FAD/NAD reductases (SCOP). This relationship can be inferred from the many families that have both alcohol dehydrogenases and FAD/NAD reductases or FAD/NAD(P) binding domains as functional neighbors.

In Table 1 we removed 7 families[1] whose functional neighbors were two molybdenum-related protein families: CO-dehydrogenase molybprotein like (SCOP: 54666) and Molybdenum-cofactor binding domain (SCOP: 56004). These two families show local structure similarity to many diverse families, which seems an artifact of either the motifs or the enrichment evaluation; we ignore them for now.

The remaining families in Table 1 show many plausible functional neighbor associations, based on comparing the family names and not assuming biological knowledge.

— Many (oxido)reductase and dehydrogenase families are functional neighbors of each other.

— Multidomain cupredoxins and Cu,Zn superoxide dismutases are mutually functional neighbors, and it would seem that a copper-binding site or some elements of function are shared. The bidirectional similarity with high p-values reinforces the functional neighbor relationship.

— Lipase/lipooxygenase is a functional neighbor of both bacterial and fungal lipases but has a different fold.

— Starch-binding domains and E-set sugar-binding domains bind carbohydrates, similar to β-glycanases and glycosyl hydrolase family 1 that have them as functional neighbors, but with different folds.

— The CheY-related family has the HAD-like family as functional neighbors; these two families are known to share similarity in the Mg^{2+}-ion binding site[26].

— Succinyl-CoA synthetase of flavodoxin fold is a neighbor of a family of flavoproteins that oxidize succinate.

— Metallohydrolase/oxidoreductase of $\alpha + \beta$ fold is a neighbor of metallodependent hydrolase of TIM-barrel (α/β) fold.

Undoubtedly there are many more valid functional neighbor associations in Table 1, that may be confirmed and elucidated by further computational and biological analysis of common local structures.

Case study: NADPH binding proteins

NADPH (nicotinamide adenine dinucleotide phosphate, reduced form) is a large ligand found in many enzymes. In SCOP [23] there are two superfamilies of NADPH-binding proteins: FAD/NAD(P)-binding domains (SCOPID: 51905) and NAD(P)-binding Rossmann-fold domains (SCOPID: 51735), which share no sequence or fold similarity.

In order to test our method, we applied it to the SCOP superfamily FAD/NAD(P)-binding domain (SCOPID: 51905) to (1) obtain recurring spatial motifs (frequent subgraphs/cliques), (2) search for the occurrences of each identified motif in all representative protein structures, and (3) report those SCOP (super)families in which a particular motif is significantly enriched.

Remote Superfamilies Identified: The superfamilies enriched in spatial motifs are listed in Table 2. As expected, we detect the other SCOP superfamily: NAD(P)-binding Rossmann-fold domains, which has no sequence or fold similarity yet shares several NADPH-binding motifs with the original SCOP family. In Figure 2, we show all signif-

[1] ARM-repeat, β-carbonic anhydrase, carbohydrate phosphatase, CutA divalent ion tolerance, Enolase superfamily, Nucleotidyltransferase, and WD40-repeat

Motifs of family (F)	**Enriched in family(-ies)** (E_i)	**p-value exponent**
ABC transporter ATPase domain (52686)	Elongation factors (50448)	-10
6-phosphogluconate dehydrogenase (48179)	Succinate dehydrogenase/fumarate reductase flavoprotein (46978, 51934, 56426)	-11
	Alcohol dehydrogenase-like (50136, 51736)	-11
	FAD/NAD linked reductase (51943, 55425)	-15
Adenine nucleotide α-hydrolase (52402)	Alcohol dehydrogenase-like (50136, 51736)	-7
	FAD/NAD(P) binding domain (51905)	-7
Adenylyltransferase (52397)	Formate-dehydrogenase/DMSO-reductase (53707)	-8
Amino acid dehydrogenase-like (51883)	Succinate dehydrogenase/fumarate reductase flavoprotein (46978, 51934, 56426)	-10
	Subtilase (52744)	-7
	FAD/NAD linked reductase (51943, 55425)	-13
Alkaline phosphatase (53649)	FAD/NAD linked reductase (51943, 55425)	-7
Bacterial lipase (53570)	Cu,Zn superoxide dismutase-like (49330)	-8
	Lipase/lipooxygenase (PLAT/LH2) domain (49723)	-7
β-Glycanase (51487)	Starch-binding domain (49453)	-7
	E-set domains of sugar-utilizing enzymes (81282)	-8
Carbon-nitrogen hydrolase (56317)	Subtilase (52744)	-10
Carboxylesterase (53487)	Phosphoglycerate kinase (53749)	-7
	FAD/NAD(P) binding domain (51905)	-8
	Colipase-binding domain (49730)	-8
CheY-related (SCOP1.67)(52173)	Haloacid dehalogenase (HAD) like (56784)	-12
CheY-like (sf, SCOP1.67)(52172)	NAD(P)-binding Rossmann fold domain (51735)	-10
Cupredoxin, multidomain (49550)	Cu,Zn superoxide dismutase-like (49330) ($\rightleftarrows$)	-13
	Fe,Mn superoxide dismutase-like (46610, 54720)	-15
	RuBisCo (51650)	-9
	Matrix metalloprotease (55528)	-13
DHS-like NAD/FAD binding domain (52467)	Lactate and malate dehydrogenases (56328)	-10
	Subtilase (52744)	-8
	FAD/NAD(P) binding domain (51905)	-11
dsRNA-binding domain (54768)	Thiolase-related (53902)	-12
EGF/Laminin (57196)	Vertebrate phospholipase A2 (48623)	-7
	Periplasmic binding-like II (53850)	-8
	Snake venom toxin (57303)	-15
	Cystine-knot cytokine (57501)	-11
Endonuclease, His-Me finger (54060)	Formate-dehydrogenase/DMSO-reductase (53707)	-11
	Galactose mutarotase-like (74650)	-8
ETFP subunit (52432)	Alcohol-dehydrogenase-like (50136, 51736)	-8
	FAD/NAD linked reductase (51943, 55425)	-14
Extended AAA-ATPase (81269)	Elongation factors (50448)	-8
FMN-linked oxidoreductase (51396)	Alcohol dehydrogenase-like (50136, 51736)	-9
	Succinate dehydrogenase/fumarate reductase flavoprotein (46978, 51934, 56426)	-8
	FAD/NAD-linked reductase (51943, 55425)	-7
Fungal lipase (53558)	Lipase/lipooxygenase (PLAT/LH2) domain (49723)	-7
Gln-amidotransferase cls I (52318)	Thiamin diphosphate binding (52518)	-10
Glycosyl hydrolase family 1 (51521)	Starch-binding domain (49453)	-7
	E-set domains of sugar-utilizing enzymes (81282)	-8
HAD-like (SCOP1.67)(56784)	DnaQ-like 3'-5' exonuclease (53118)	-7
Inosine Monophosphate dehydrogenase (51413)	Alcohol dehydrogenase-like (50136, 51736)	-11
	FMN-linked oxidoreductase (51396)	-9
	Aldolase, Class I (51570)	-7
Integrin A(I)(53301)	Amino-acid dehydrogenase like (51883)	-8
Metallo-dependent phosphatase (56300)	Cu,Zn superoxide dismutase-like (49330)	-10
Metallohydrolase/ oxidoreductase (56281)	Pectin lyase-like (51126)	-7
	Metallodependent hydrolase (51556)	-7
Metalloprotease "zincin" (55486)	Cu,Zn superoxide dismutase-like (49330)	-7
	Peptide deformylase (56421)	-11
N-type ATP pyrophosphatase (52403)	ALDH-like (53721)	-7
NADH Oxidase/flavin reductase (55468)	Formate-dehydrogenase/DMSO-reductase (53707)	-7
p53-like transcription factor (49417)	Galactose mutarotase-like (74650)	-7
PDZ domain (50157)	Alcohol dehydrogenase-like (50136, 51736)	-8
	Glyceraldehyde-3-phosphate dehydrogenase (51800, 55347)	-8
Phospholipase C/P1 (48537)	Galactose mutarotase-like (74650)	-7
Pyruvate ox/decase (52475)	Thiamin diphosphate binding (52518)	-7
Ribonuclease H (53099)	Thiamin diphosphate binding (52518)	-8
	Subtilase (52744)	-7
Ribulose phosphate binding barrel (51366)	α-Amylase (51012, 51446)	-10
	Aldolase, Class I (51570)	-11
	Cystathionine synthase like (53402)	-8
RuvA C-terminal domain (46928)	DNA polymerase I (56673)	-7
SGNH hydrolase (52266)	FMN-linked oxidoreductase (51396)	-12
	FAD/NAD(P) binding domain (51905)	-17
SIS domain (53697)	FAD/NAD(P) binding domain (51905)	-9
	Thiamin diphosphate binding (52518)	-8
	Subtilisin-like (52743)	-7
Succinyl-CoA synthetase (52210)	Succinate dehydrogenase/fumarate reductase flavoprotein (46978, 51934, 56426)	-8
	FAD/NAD-linked reductase (51943, 55425)	-7
Trp biosynthesis (51381)	Aldolase (sf) (51569)	-7

Table 1. Examples of functional neighbor families without overall structural similarity, found by enrichment evaluation in SCOP.

Motif	S	F	p	f
1	5	51905	10^{-10}	0.37
ILVVV		51735	10^{-10}	0.19
2	4	51905	10^{-15}	0.72
GVVV		51735	10^{-12}	0.33
3	4	51905	10^{-14}	0.70
GGGG		50494	10^{-15}	0.58
4	4	51905	10^{-14}	0.70
GGGS		51735	10^{-13}	0.36
		50494	10^{-15}	0.69
		53383	10^{-13}	0.60
5	4	51905	10^{-14}	0.72
GGGT		51735	10^{-15}	0.49
		50494	10^{-14}	0.57
6	4	51905	10^{-11}	0.40
CGGG		50494	10^{-14}	0.56
7	4	51905	10^{-14}	0.60
GGII		51735	10^{-13}	0.33
8	4	51905	10^{-10}	0.46
CGGL		50494	10^{-14}	0.60
9	4	51905	10^{-15}	0.65
GGGL		51735	10^{-13}	0.35
10	4	51905	10^{-15}	0.60
GGGI		51735	10^{-12}	0.30
11	5	51905	10^{-12}	0.40
AGIIV		56235	10^{-10}	0.33

Table 2. Eleven motifs obtained from the SCOP superfamily: FAD/NAD(P)-binding domain proteins. We used σ value 15/43. No new significantly enriched (super)families were discovered using lower support thresholds such as 10. S: the number of residues included in a motif, F: SCOP (super)family ID: 51905: FAD/NAD(P)-binding domain, 51735: NAD(P)-binding Rossmann-fold domains, 50494: Trypsin-like serine proteases, 53383: PLP-dependent transferases, 56235: N-terminal nucleophile aminohydrolases (Ntn hydrolases). p, the p-value of the motif's occurrences in the related SCOP family. f: the support of the motif in the related family.

icant spatial motifs shared between the two NADP binding families in a protein from the FAD/NAD binding domain superfamily. Most residues covered by the motifs are located near the NAD ligand.

In Figure 3 we show a motif that is statistically enriched in both families; it has conserved geometry and is adjacent to NADPH in two proteins that belong to the two families.

We emphasize that we did not include any information from NADPH during our search process, yet were still able to identify the motifs since they represent local residue patterns conserved among proteins in both SCOP superfamilies. This remarkable local commonality and clear interaction between motifs and the ligand show that our method helps reveal hidden biologically significant patterns.

Identifying remote relationships from PDB

We extended our analysis from a single SCOP superfamily to all SCOP superfamilies defined on all representative structures in PDB. Instead of running SCOP superfamilies one by one, we tested our motif mining algorithm on all 6625 representative structures in PDB. Each clique appearing in at least 20 protein members is then used to search for its SCOP superfamily enrichments and (possible) remote similarities between two or more SCOP superfamilies.

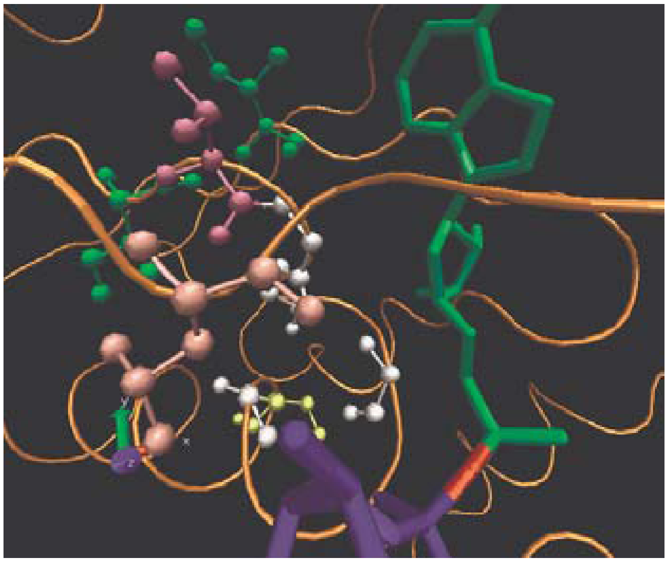

Figure 2. Examples of all motifs which are significantly enriched in SCOP Superfamily FAD/NAD(P)-binding domain in protein 1kew (chain a). Only residues are shown in this figure; their interactions are omitted to for clarity.

In Figure 4 we plotted all identified remote similarities within a subtree of three SCOP classes: all Alpha, all Beta, and Alpha and beta proteins. We also used a very stringent $P\text{–}value$ upper limit: 10^{-12} to present the most significant connections between SCOP superfamilies in these classes.

Conclusions

Our approach affords automated identification of protein family-specific packing motifs that are used to annotate protein structures or even families enriched by such motifs. Consequently, this method helps establish functional similarity and functional neighbor relationships between protein families even if they are unrelated in sequence or structure. These relationships are deduced using statistical enrichment evaluation of family-specific motifs within hierarchical structure classifications such as SCOP and GO. The detected functional similarities could have arisen by several possible mechanisms such as divergent/convergent evolution or domain/site swapping. The enrichment characterized by the P-value (Equation 1) represents a novel function similarity measure for protein families that could help annotate families of proteins with unknown or not well- understood function. We believe that studies described in this paper represent a promising new direction in the area of lo-

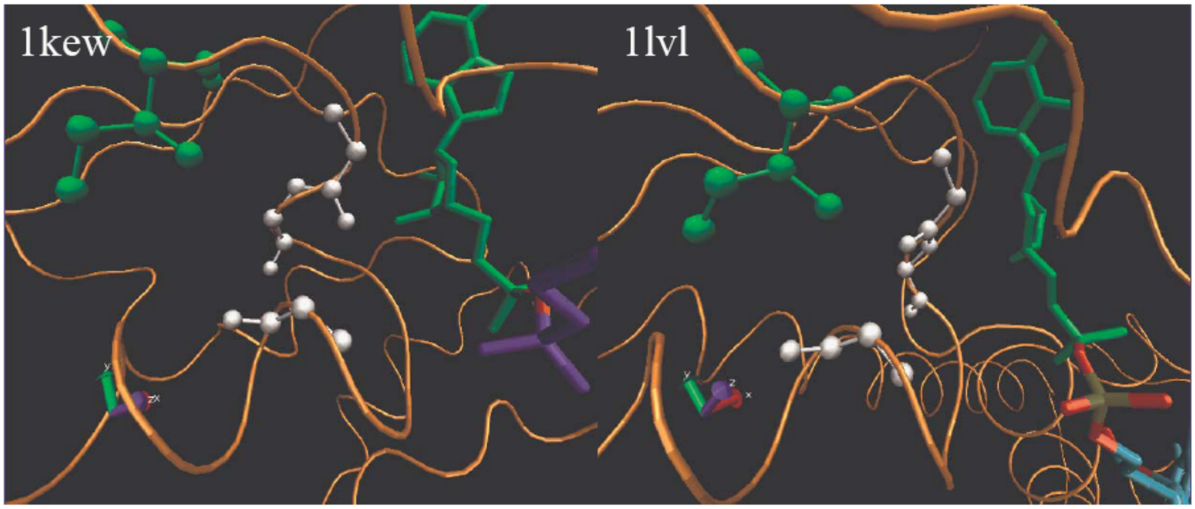

Figure 3. Example of an NADPH binding motif that is significantly enriched in two SCOP superfamilies. The four involved residues are ILE5, GLY7, GLY8, and GLY 13 (PDB: 1kew A), and ILE10, GLY12, GLY13, and GLY17 (PDB: 1lvl). The DALI z-score of the two protein structures is 4.5, pairwise sequence identity is 16%, and the USC local alignment server revealed no sequence similarity in the region of the motif.

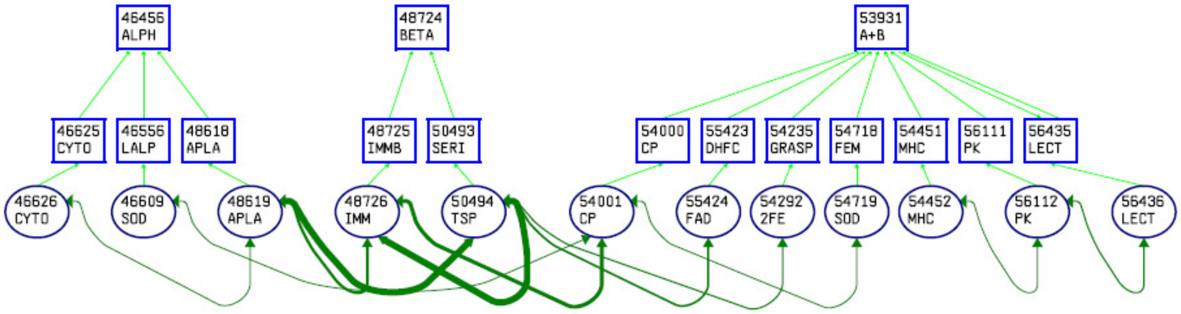

Figure 4. Remote function similarities using motifs mined from whole PDB; the plot is restricted to three classes: α, β, and $\alpha + \beta$ proteins. Low P-value threshold (10^{-12}) is used to only show the most significant links. The thickness of an edge is proportional to the number of motifs shared by the two families(F)/superfamilies(S). SCOP IDs: 55424: F FAD/NAD-linked reductases, dimerisation (C-terminal) domain; 54452: S MHC antigen-recognition domain; 56112: S Protein kinase-like (PK-like); 56436: S C-type lectin-like; 46609: S Fe,Mn superoxide dismutase (SOD), N-terminal domain; 54719: S Fe,Mn superoxide dismutase (SOD), C-terminal domain; 54001 S Cysteine proteinases; 48726: S Immunoglobulin; 48619: S Phospholipase A2, PLA2; 46626: S Cytochrome c; 54292: S 2Fe-2S ferredoxin-like; 50494: S Trypsin-like serine proteases.

cal similarity-based protein function inference.

Acknowledgments

We thank Dr. Ruchir Shah for some useful discussions. These studies were supported by NIH grant GM068665 and NSF grant CCF-0523875.

References

[1] Aloy, P., E. Querol, F. X. Aviles, and M. J. Sternberg: 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.

[2] Artymiuk, P. J., A. R. Poirrette, H. M. Grindley, D. W. Rice, and P. Willett: 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.

[3] Babbitt, P. C.: 2003, 'Definitions of enzyme function for the structural genomics era'. *Curr Opin Chem Biol* **7**, 230–237.

[4] Bandyopadhyay, D., J. Huan, J. Liu, J. Prins, J. Snoeyink, W. Wang, and A. Tropsha: 2006, 'Structure-based function inference using protein family-specific fingerprints'. *Protein Science* **15**(6), 1537–1543.

[5] Bandyopadhyay, D. and J. Snoeyink: 2004, 'Almost-Delaunay Simplices: Nearest Neighbor Relations for Imprecise Points'. In: *ACM-SIAM Symposium On Discrete Algorithms*. pp. 403–412.

[6] Barker, J. and J. Thornton: 2003, 'An algorithm for constraint-based structural template matching: application to 3D templates with statistical analysis.'. *Bioinformatics* **19**(13), 1644–9.

[7] Binkowski, T. A., P. Freeman, and J. Liang: 2004, 'pvSOAR: Detecting similar surface patterns of Pocket and Void Surfaces of Amino Acid Residues on proteins'. *Nucleic Acid Research* **32**, W555–W558.

[8] Bradley, P., P. S. Kim, and B. Berger: 2002, 'TRILOGY: Discovery of sequence-structure patterns across diverse proteins'. *Proceedings of the National Academy of Sciences* **99**(13), 8500–8505.

[9] Ferre, F., G. Ausiello, A. Zanzoni, and M. Helmer-Citterich: 2004, 'SURFACE: a database of protein surface regions for functional annotation'. *Nucl. Acids. Res.* **32**(90001), D240–244.

[10] Fetrow, J. S. and J. Skolnick: 1998, 'Method for prediction of protein function from sequence using the sequence-to-structure-to-function paradigm with application to Glutaredoxins/Thioredoxins and T1 Ribonucleases'. *J. of Mol. Biol.* **281**, 949–968.

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

[12] Hambly, K., J. Danzer, S. Muskal, and D. Debe: 2006, 'Interrogating the druggable genome with structural informatics'. *Mol. Divers.* **10**, 273–281.

[13] Huan, J., D. Bandyopadhyay, J. Snoeyink, J. Prins, A. Tropsha, and W. Wang: 2006, 'Distance-based Identification of Spatial Motifs in Proteins Using Constrained Frequent Subgraph Mining'. In: *IEEE Computational Systems Bioinformatics Conference (CSB)*.

[14] Huan, J., D. Bandyopadhyay, W. Wang, J. Snoeyink, J. Prins, and A. Tropsha: 2005, 'Comparing Graph Representations of Protein Structure for Mining Family-Specific Residue-Based Packing Motifs'. *Journal of Computational Biology* **12**(6), 657–671.

[15] Huan, J., W. Wang, and J. Prins: 2003, 'Efficient Mining of Frequent Subgraph in the Presence of Isomorphism'. In: *ICDM '03: Proceedings of the Third IEEE International Conference on Data Mining*. pp. 549–552.

[16] Jacobs, D., A. Rader, L. Kuhn, and M. Thorpe: 2001, 'Protein flexibility predictions using graph theory'. *Proteins* **44**(2), 150–165.

[17] Jambon, M., O. Andrieu, C. Combet, G. DelÃl'age, F. Delfaud, and C. Geourjon: 2005, 'The SuMo server: 3D search for protein functional sites'. *Bioinformatics* **21**, 3929–3930.

[18] Krissinel, E. B. and K. Henrick: 2004, 'Common subgraph isomorphism detection by backtracking search'. *Software: Practice and Experience* **34**(6), 591–607.

[19] Laskowski, R. A., N. M. Luscombe, M. B. Swindells, and J. M. Thornton: 1996, 'Protein clefts in molecular recognition and function'. *Protein Sci* **5**(12), 2438–2452.

[20] Lisewski, A. and O. Lichtarge: 2006, 'Rapid detection of similarity in protein structure and function through contact metric distances'. *Nucleic Acids Res.* **34**, e152.

[21] Milik, M., S. Szalma, and K. Olszewski: 2003, 'Common Structural Cliques: a tool for protein structure and function analysis.'. *Protein Eng.* **16(8)**, 543–52.

[22] Murga, L., Y. Wei, and M. Ondrechen: 2007, 'Computed protonation properties: unique capabilities for protein functional site prediction'. *Genome Inform* **19**, 107–118.

[23] Murzin, A., S. Brenner, T. Hubbard, and C. Chothia: 1995, 'SCOP: a structural classification of proteins database for the investigation of sequences and structures'. *Journal of Molecular Biology* **247**, 536–40.

[24] Nussinov, R. and H. J. Wolfson: 1991, 'Efficient detection of three-dimensional structural motifs in biological macromolecules by computer vision techniques'. *PNAS* **88**, 10495–99.

[25] Petsko, G. and D. Ringe: 2004, *Protein Structure and Function*. New Science Press Ltd.

[26] Ridder, I. S. and B. W. Dijkstra: 1999, 'Identification of the Mg2+-binding site in the P-type ATPase and phosphatase members of the HAD (haloacid dehalogenase) superfamily by structural similarity to the response regulator protein CheY'. *Biochem J* **339 (Pt 2)**, 223–226.

[27] Russell, R. B.: 1998, 'Detection of protein three-dimensional side-chain patterns: new examples of convergent evolution'. *Journal of Molecular Biology* **279**, 1211–1227.

[28] Shaffer, J. P.: 1995, 'Multiple Hypothesis Testing'. *Ann. Rev. Psych* (46), 561–584.

[29] Stark, A. and R. Russell: 2003, 'Annotation in three dimensions. PINTS: Patterns in Non-homologous Tertiary Structures.'. *Nucleic Acids Res* **31**(13), 3341–4.

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

[31] Taylor, W. R. and I. Jonassen: 2004, 'A Structural Pattern Based Method for Protein Fold Recognition'. *Proteins* **56**(2), 222–234.

[32] Turcotte, M., S. Muggleton, and M. Sternberg: 2001, 'Automated discovery of structural signatures of protein fold and function.'. *J Mol Biol.* **306(3)**, 591–605.

[33] Ullman, J. R.: 1976, 'An Algorithm for Subgraph Isomorphism'. *Journal of the Association for Computing Machinery* **23**, 31–42.

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

[35] Wang, K. and R. Samudrala: 2006, 'Automated functional classification of experimental and predicted protein structures'. *BMC Bioinformatics* **7**, 278.

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

[37] Weskamp, N., D. Kuhn, E. Hullermeier, and G. Klebe: 2004, 'Efficient similarity search in protein structure databases by k-clique hashing'. *Bioinformatics* **20**, 1522–1526.

[38] Xie, L.: 2004, 'Methods for comparing functional sites in proteins'. *WIPO patent* (number WO/2005/045424). `http://www.wipo.int/pctdb/en/wo.jsp?WO=2005045424`.

[39] Xie, L. and P. Bourne: 2007, 'A robust and efficient algorithm for the shape description of protein structures and its application in predicting ligand binding sites'. *BMC Bioinformatics* **8 Suppl 4**, S9.

[40] Xie, L. and P. Bourne: 2008, 'Detecting evolutionary relationships across existing fold space, using sequence order-independent profile-profile alignments'. *Proc. Natl. Acad. Sci. U.S.A.* **105**, 5441–5446.