Spatial Motif Discovery in Papain-like Cysteine Protease Family

by
Yetian Chen

A thesis submitted to the faculty of the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Master of Science in the Department of Biochemistry and Biophysics, School of Medicine.

Chapel Hill

2008

Approved by:

Advisor: Alexander Tropsha

Reader: Nikolay V. Dokholyan

Reader: Brian Kuhlman

Abstract

YETIAN CHEN: Spatial Motif Discovery in Papain-like Cysteine Protease Family

(Under the direction of Alexander Tropsha)

Spatial motifs, which are amino acid packing patterns, occur frequently within a set of proteins with some common specific functions and features. In this study, we report the application of a novel frequent subgraph mining algorithm to retrieve conserved spatial motifs from protein 3D structures of Papain-like cysteine protease family. Each of the frequent spatial motifs we identified were found highly specific to the PCP family, measured by P-value<10^{-49}. And we showed that the combination of these family specific motifs can discriminate between the PCP family members and the background (a non-redundant subset of PDB) with very good sensitivity and predicative accuracy. These spatial motifs were found to cover either structurally important or functionally important sites, such as the catalytic dyad and the hydrophobic pocket that determines the substrate specificity. A PROSITE-like consensus sequence pattern assembled by mapping these structural motifs to sequence level identifies the PCP sequences in Swiss-Prot database with 100% precision and good recall. These suggest that structurally and functionally specific amino acid packing patterns or motifs can be discovered by computational and statistical geometry analysis of protein structures and used to annotate protein structures and sequences.

To my parents.

ACKNOWLEDGEMENTS

I would like to thank my thesis advisor, Dr. Alexander Tropsha, for his valuable advice and support. I would also like to thank members of my committee, Dr. Nikolay Dokholyan and Dr. Brian Kuhlman for their guidance and support.

I would also want to extend my thanks to the members of the Molecular Modeling lab for their continuous support and help!

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

I. INTRODUCTION

The importance of protein local structural similarity for protein classification and annotation

A central principal of modern structural biology is that a protein's function is determined by its structure. This has been studied experimentally by protein scientists using physical analytical methods such as X-ray crystallography and NMR as well as by computational chemists using a variety of simulation and statistical approaches. Recent improvements in structural biology techniques have made structural characterization more easily achieved. This has led to the emergence of a worldwide program termed Structural Genomics, which aims to solve the structures for all proteins as a means to understand function [1,2,4]. This effort has resulted in a rapid increase in the number of proteins for which the 3D structures is known. The Protein Data Bank (PDB, [3]), a public on-line protein-structure repository, contains over 50,000 entries (April, 2008) and the number is still growing exponentially. Thus, methods to annotate proteins through structure are thus now of growing importance.

There are a number of methods in current use for classifying and annotating proteins through structure. For the most part, they rely on detecting some structural similarity between the target protein and a structure of known function in the PDB. Traditionally, structural comparison is largely based on detecting the global similarity of a pair of protein structures. This has been implemented in many structural classification systems such as SCOP [5], CATH [6] and DALI [7]. The main emphasis of these works has been on the comparison of protein folds by alignment of large portions of protein structures to locate maximal lengths of superimposable main chain [8]. The fold similarity between two protein structures, even in the absence of obvious sequence similarity, can imply common ancestry, which in turn can suggest details as to function. However, it is also possible for such remote homologues to have different functions. This phenomenon now is known as divergent evolutions [9, 10]. For example, TIM-barrel, functioning as a generic scaffold, catalyzes 15 different enzymatic functions. Each of these enzymatic functions has the active sites located at different areas of the common scaffold [11]. The subtle differences at these substructures might not be easily differentiated by the global comparison approaches. Conversely, a type of function can adopt totally different folds. For example, the carbonic anhydrases (EC number 4.2.1.1) are associated with two different folds [12], but they have exactly the same active sites. Obviously, the structural comparison relying on global alignment fails to detect these similarities. Moreover, a large portion of the protein structures solved by the ongoing structural genomics projects represent a new fold [13], making functional insights from fold comparison impossible.

An alternative approach for classifying protein structures is based on detection of more local patterns, without any requirement for similarity in fold. A limited 20 letter amino acid alphabet means that nature is restricted in the choice of atoms for active or binding sites, even proteins perform different functions [14]. Optimal tertiary combinations of a group of amino acids can arise convergently [14], such as proteins with the Ser/His/Asp catalytic triad (subtilisin, trypsin, lipases etc, see [15]). Other examples include various amino acid combinations around metal binding sites, and recently the folding nuclei of proteins [16, 17]. Identification of such sequence order independent patterns can suggest functional or mechanistic details, and can aid the design of novel regulators of function by analogy. It is also provides important insights into the nature of convergent evolution.

The need for developing more versatile and faster local pattern discovery tool

The past few yeas have witnessed the development of highly sophisticated methods for comparison and identification of such substructural features. These include manual identifications of many specific types of site or motifs and more general investigations into small motifs in protein structures. Wallace et al [15] derived tertiary structure templates for the Ser/His/Asp catalytic triad and demonstrated how such templates could be used to search for novel examples of know side-chain patterns. Poter et al [18] manually compiled a library of enzyme active-site templates (called Catalytic Site Atlas) based on information from literature sources. Each of such templates consists of between two and five residues. Laskowski et al [19] then demonstrated how these templates could be used for functional prediction although the rate of false positive matches is unsatisfactory in some cases where the template consists of very few residues. Russell [14] developed an algorithm that can automatically uncover side-chain patterns common to two protein structures, which allows identification of more complicated local patterns. Recently, Wangikar P. et al [20] expanded this idea to allow comparison of a group of proteins via objective and automated graph theoretic approach. In their method, a protein is modeled as a graph in which the vertices are the functional atoms from the side-chains of the amino acid. An edge exists between two vertices if they are within a certain distance. A structural pattern is a complete sub-graph. Common structural patterns are extracted from a group of graphs by subgraph mining technique. They demonstrated that this technique could be used to identify the functional sites in protein families. However, to reduce the computation complexity, both of the two methods exclude the amino acid types that are found to have low probability of being present in functional sites. As a result, both of them will miss patterns that are conserved due to structural stability so that these methods may be of very limited use.

To facilitate the pattern discovery in protein structures, Singh R et al. [23] introduced a computational geometry technique known as Delaunay Tessellation (DT, and its variants) to model protein structures as contact graphs. Compared to distance-based graph representation used by Wangikar P. et al [20], DT-based graph representation substantially reduces the graph density of protein structures, thereby reduces the computation complexity. On the other hand, our colleagues J. Huan and Prof. Wei Wang developed a novel subgraph mining algorithm which affords faster and more accurate identification of common patterns from a group of graphs [29, 30, 31]. To assess the performance of these techniques in recognizing conserved residue packing patterns in protein structural families or functional families and test the feasibility of using these patterns in protein annotation, I chose Papain-like Cysteine Protease family, a typical SCOP family which has been well studied in literature from SCOP version 1.67 to do a case study.

Papain-like cysteine protease family

Papain-like cysteine protease family (PCP, SCOP ID: 54002) comprises a group of papain-related proteases sharing a remarkably conserved two-domain structure: an N-terminal domain and a C-terminal domain. The active site comprising a catalytic dyad Cys-His is located at the V-shape active site cleft extending along the interface between two domains [50]. The left (L-) domain dominated by three α-helices and the right (R-) domain is based on a β-barrel motif. Papain-like cysteine proteases are the most abundant among the cysteine proteases [51].The enzymes, most of which are endopeptidases, consist of papain and related plant proteinases such as chymopapain, caricain, bromelain, actinidin, ficin, and aleurain, and the lysosomal cathepsins B, H, L, S, C and K [51].Most of these proteases have substrate specificity at the S2 site, which forms a hydrophobic pocket and prefer a Phenylalanine at the corresponding site of the substrate peptide [50].

Using the contact graph representation of protein structure and the frequent subgraph mining algorithm, a handful of local patterns or spatial motifs were located for the PCP family. The result showed that each of these frequent spatial motifs is statistically linked to the PCP family, measured by a hyper-geometric distribution. It is also showed that a combination of these motifs can discriminate between the structures of family members and background with over high sensitivity and predictive accuracy. By mapping the residues covered by these motifs to the structure, these residues are shown to perform significantly biological functions such as forming the active sites and the hydrophobic pocket that is important to the substrate specificity. Furthermore, a PROSITE-like sequence pattern derived from the structural motifs can identify the PCP sequences in Swiss-Prot database with precision of 100% and recall of 92.8%, suggesting that the structure-derived sequence patterns characteristic of protein families or classes can be used as queries in mining sequences as well.

II. METHODOLOGY

Graph Representation of Protein Structures

We construct graphs whose vertices represent the amino acids, using the coordinates of Cα atoms and labeling by residue type. Two types of edge may connect residues: a peptide edge that connects two residues that are adjacent in the primary sequence, or a proximity edge that connects two (non-bonded) residues identified as spatial neighbors in the 3D space based on the almost-Delaunay (AD) edges of parameter ε (default 0.1 Å) [32].

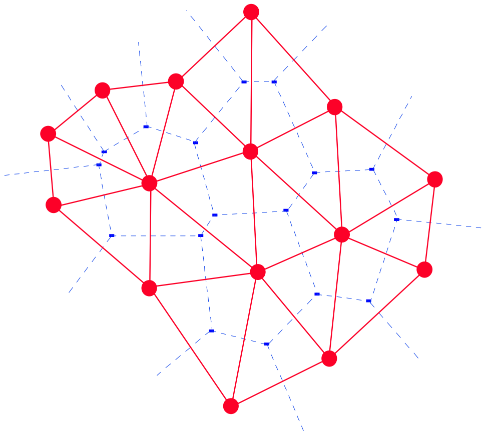


Figure 1. Voronoi/Delaunay tessellation in 2D space (Voronoi polyhedra – blue dashed line, Delaunay simplices - solid line). All the points in a blue polyhedron are closest to the central red point. Modified from [49]

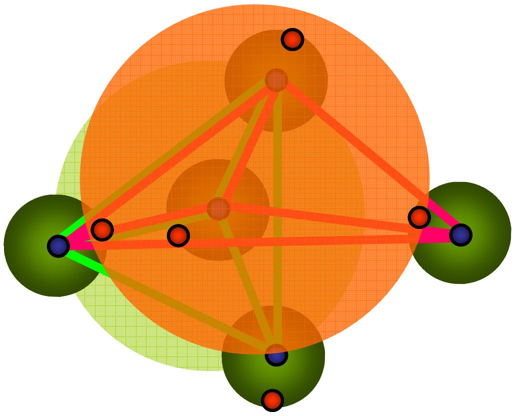


Figure 2. Illustration of AD. A 4-tuple of points is almost-Delaunay with threshold ε, if, by perturbing all points in the set by at most ε, the circumscribing sphere can become empty. A 4-tuple of points is AD(ε) if e is the minimum threshold. Green Delaunay is AD(0); Red is AD(ε).

The definition of Almost-Delaunay edges is derived from the Delaunay tessellation [23], which is dual to the Voronoi diagram. The Voronoi diagram is a mathematical model used to divide space into regions so that all the points in each region are closest to the same point (*Figure 1*).

The Delaunay tessellation (DT) is dual to the Voronoi diagram, i.e., it connects points in neighboring regions of the Voronoi diagram [21]. For any set of points, these structures give a unique set of nearest neighbors. The property of recognizing the nearest neighbors in any high dimensional space has brought wide applications of the Delaunay tessellation into diverse fields [22]. When applied to a collection of points representing amino acid residues of a structure, DT generates an aggregate of space-filling, irregular tetrahedral, or simplices. Each Delaunay simplex defines objectively and uniquely four nearest neighbor residues as vertices of the tetrahedron. Naturally, the entire aggregate could be regarded as a network of contacts between residues thereby forming a connected graph [23, 24, 25]. Researches using DT-represented protein structures to study the protein stability and protein motions have implied strong applicability of DT for modeling protein structures [26, 27]. However, standard Delaunay tessellation still has some drawbacks when applied to protein structure modeling since protein structure coordinates are imprecise [28]. The errors are introduced during the experiment, differences between experimental methods or conditions, and actual motions within the protein. In presence of these noises, the set of nearest neighbors can change, being different for any two measurements of the coordinates in the same structure. To solve this problem, almost-Delaunay edges are defined by relaxing the empty sphere property to say that a pair of points p and q is joined by an almost-Delaunay edge with parameter ε, if by perturbing all points by at most ε, p and q can be made to lie on an empty sphere (*Figure 2*). Equivalently, one looks for a shell of width 2ε, formed by concentric spheres, so that p and q are on the outer sphere, and all points are outside the inner sphere [28]. Various values of the parameter ε correspond to different allowed perturbations or motions: 0.1 – 0.25 Ǻ would model decimal inaccuracies in the PDB coordinates or small vibrations, and 0.5 – 0.75 Ǻ would model perturbations due to coarser motions. The protein graphs constructed with the almost-Delaunay edges are termed AD graphs. Mathematically, the Delaunay Tessellation graph of a protein is a subgraph of AD graph of the same protein. AD based graph representation of protein structure are more robust in recognizing the naturally physical interaction between two residues in protein structure allowing identification of more family specific packing motifs.

Additionally, to add geometric constraints to the graph representation of protein structures, a pair of nearby non-contact residues is connected by a *distance edge*. To reduce complexity, distance edges longer than 12.5 Å are normally eliminated from the graphs. To distinguish edges connecting residues with different inter-atom distances, every edge in a structural graph (contact edge or distance edge) is labeled by the distance of the Cα atoms of the two residues it connects; the distance is further discretized into bins to accommodate noise (Table 1). The width of such bins is commonly referred to as the *distance tolerance*, and popular choices are 1 Å [44], 1.5 Å [45], and 2 Å [46]. In our system, we choose the median number 1.5 Å, which empirically delivers patterns with good geometric conservation. When we perform graph matching, we require that matching nodes have the same label and matching edges have the same type (contact or distance edges) and edge label. By enforcing these matching conditions, we guarantee that the spatial motifs reported by our system have well defined residue identity composition and three dimensional shapes.

Table 1. Edge label by discretizing distance between Cα atoms of two residues into bins (The distance unit is Å)

$d \le 4$	$4 < d \le 5.5$	$5.5 < d \le 7$	$7 < d \le 8.5$	$8.5 < d \le 10$	$10 < d \le 11.5$	$11.5 < d$
1	2	3	4	5	6	7

Mining Frequent Spatial Motifs From Protein Structure Graphs

In our study, a protein structure is represented with an undirected graph, the problem of finding conserved local packing patterns in protein families becomes the identification of recurrent subgraphs from a group of labeled graphs, which is a widely studied subject, termed subgraph mining.

We define a labeled graph G as five element tuple $G=(V, E, \Sigma_V, \Sigma_E, \lambda)$, where V is a set of vertices or nodes and $E \subseteq V \times V$ is a set of undirected edges. Σ_V and Σ_E are disjoint sets of vertex and edge labels, respectively, and λ is a function that assigns labels to vertices and edges: $V \rightarrow \Sigma_V$ and $E \rightarrow \Sigma_E$. We assume that a total ordering is defined on the labels in $\Sigma_V \cup \Sigma_E$.

$G' = (V',E')$ is a *subgraph* of G, denoted by $G' \subseteq G$, if vertices $V' \subseteq V$, and edges $E' \subseteq (E \cap (V' \times V'))$, i.e. E' is a subset of the edges of G that join vertices in V'. A fundamental part of our method is to find an occurrence of a graph H within another graph G. To make this more precise, we say that *H occurs in G* if we can find an *isomorphism* between graph $H = (V_H, E_H, \Sigma_V, \Sigma_E, \lambda_H)$ and some subgraph of $G = (V_G, E_G, \Sigma_V, \Sigma_E, \lambda_G)$. An isomorphism from H to the subgraph of G defined by vertices $V \subseteq V_G$ is a bijection between vertices f: $V_H \rightarrow V$ that preserves edges and edge/node labels.

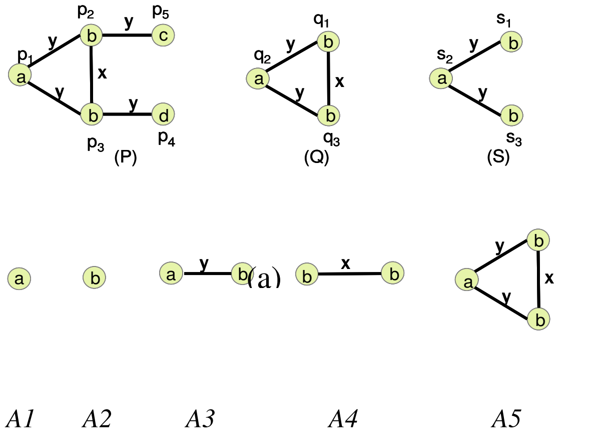


Figure 3. Basic concept of Subgraph mining. (a): Examples of three labeled graphs (referred to as a graph database). The labels of the nodes are specified within the circle and the labels of the edges are specified along the edge. The mapping q1 → p2, q2 → p1, q3→ p3 represents an induced subgraph isomorphism from graph Q to P. (b): All the frequent induced subgraphs with support $\sigma \geq 2/3$ for the graph database. (Modified from [29])

In this study, we restrict ourselves to fully interconnected subgraphs: cliques. A clique is a graph where each node have degree n-1 where n is the size (number of nodes) of c and the degree of a node is defined as the number of edges incident with it. For example, the graph Q in *Figure 3(a)* is a clique since all its nodes have degree 2 while S is not. In protein structure graphs, a clique corresponds to a spatial motif with all inter-residue distances computed. The reason we focus on cliques is that the identified spatial motifs have very strict geometry.

Given a *graph database GD* which is a set of graphs, we define the *support* of a clique as the fraction of graphs in *GD* in which the clique occurs. We choose a threshold $0 < \sigma \leqslant 1$, and define the clique to be frequent if and only if its support is at least σ. Note that while one clique may occur many times within a single graph, for the purposes of support, these counts as only one occurrence. The problem of Frequent Clique Mining is to identify all frequent cliques for a graph database *GD.* Figure 3(b) shows all cliques which appear in at least two graphs in the graph database shown in Figure 3(a). If we use support threshold $\sigma = 2/3$, all five cliques will be reported to users. If we increase σ to 3/3, only the *A1, A2, A3* will be reported.

Since our method reports all the frequent cliques, some cliques are subgraphs of other cliques and these include much redundant information. Thus, we select to report the maximal cliques. That means no one of the reported cliques is a subgraph of other cliques. For example, in Figure 3(b), only *A*5 will be reported because other cliques are subgraphs of *A*5 if support $\sigma = 2/3$ is given.

Algorithm Description

The core software FFSM executable 1.0 (Fast Frequent Subgraph Mining) was developed by Jun Huan, Wei Wang, Jan Prins et al. at Department of Computer Science, University of North Carolina at Chapel Hill and distributed through webpage at http://www.cs.unc.edu/~huan/FFSM.html.

The algorithm details are described in [29,30,31].

Statistical Significance of the Spatial Motif in the Family

To establish the statistical significance, any subgraphs that are frequent in a SCOP family are checked against the background dataset. We determine the statistical significance of a spatial motif by measuring a *P-value*, defined by the following hyper-geometric distribution.

$$P - value = \frac{\binom{|F|}{k}\binom{|M| - |F|}{|T| - k}}{\binom{|M|}{|T|}} \quad \text{(Equation 1)}$$

Here, *M* is a collection of representative proteins, selected from all known structures in PDB; *T* is a subset of proteins ($T \subseteq M$), in which a particular spatial motif occurs; *F*, a subset of proteins ($F \subseteq M$), standing for the structures family select to establish the statistical significance. |*F*|, |*M*|, |*T*| is the cardinality of *F, M, T,* respectively. Thus, P-value is the probability of observing a particular motif containing proteins $K = F \cap T$ with size at least *k*. For example, if a motif occurs in every member of a family *F* and in no proteins outside *F* (i.e. *K* = *F* = *T*) for a large family *F*, we would estimate that this motif is specifically associated with the family; the statistical significance of such case is measured by a *P-value* close to zero. Based on the Bonferroni correction for multiple independent hypotheses [47], 0.001/|*C*|, where |*C*| is the set of categories, is used as the default threshold to measure the significance of the *P*-value of individual test. Since the total number of SCOP families is 2327, a good starting point of *P*-value upper bound is 10^{-7}.

Experiment Setup and dataset cleaning

The protein structure dataset of Papain-like cysteine protease family (PCP, SCOP ID: 54002) was selected from SCOP version 1.67. To avoid the inclusion of many nearly identical structures that would bias the family composition and also invoke the worst case exponential behavior of subgraph mining, a list of PCP protein structures was created so that the pair-wise sequence identities between any two protein chains are no more than 90%. Furthermore, the structures in the list had better than 3 Å resolution and R-factor at most 1.0. This preprocessing resulted in 27 PCP protein structures with pair-wise sequence identity ranging from 13~60% (Table 2). In the preliminary study, I noted that mutations and chemical modifications are extensively present in these crystal structures and these alternations always locate at the functionally important residues. Since our subgraph mining is to search for the motifs with unique residue compositions among these structures, these mutations or chemical modifications will lead to missing of some motifs that are usually functionally important. Hence, I made some corrections to the PDB files: the irregular residue names from chemical modifications were corrected to the normal residue names; the mutated amino acids were changed back to the native ones (Table 2). Here we assumed that these do not make any structural changes to the original proteins. And since the Cα atoms are used as the nodes in current graph representation, this preprocessing would not change the residue positions in the original protein structures. Each of these structures was converted to a labeled graph based on almost-Delaunay edges of parameter ε (default 0.1 Å). Nodes in the graph represent Cα atoms of each residue, labeled by the residue type. The edge between any pair of residues is labeled by the distance of the Cα atoms of the two residues it connects. The distance is further discretized into bins to accommodate noise (see Table 1 for the discretization). The resulting group of graphs was submitted to the Fast Frequent Subgraph Mining (FFSM v1.0) running on the Baobab cluster computing platform (baobab.isis.unc.edu). A number of common subgraphs were reported given a certain occurrence threshold (the required number of occurrences). Then these graphs, termed motifs, were checked against a dataset of 6500 representative proteins from CulledPDB for occurrence. The motifs appeared in more than 5% of these background proteins (nonspecific motifs) were eliminated.

Table 2. The Papain-like Cysteine proteases used for mining motifs

PDBID	Chain	Remark	PDBID	Chain	Remark
1cqd	a		1qdq	a	
1gec	e		2cb5	a	CYS73SER
1gmy	a		1mem	a	
1cs8	a	CYS25OCS	1yal	_	CYS25SCH,CYS117SCH
1the	a		1deu	a	
1jqp	a		8pch	a	
1cv8	_		1m6d	a	
1me4	a		1khq	a	
1k3b	bc		2act	_	
3gcb	_	CYS73ALA	1ppo	_	
1dki	a	CYS47SER	1nqc	a	
1iwd	a		1fh0	a	
1s4v	a		1o0e	a	
1pxv	a	CYS243ALA			

The “chain” column specifies the chains in the PDB files that were used for mining. The Remark column records the mutations or chemical modifications within the protein structures.

Derivation of sequence patterns from spatial motifs

Since our structural motifs discovered by subgraph mining are conserved in terms of residue identity, they may also be used to derive sequence patterns if these residues are also ordered at sequence level in family structures (Fig 4, sequence ordered motif). For example, taking residue identities into account, an elementary packing motif M comprising four residues can be generally defined as the four nearest neighbor residues separated by three sequence distances as in the following PROSITE-like regular expression:

M: = { aa_i aa_j aa_k aa_l; d1, d2, d3}

In this formulation, i, j, k, and l are residue numbers in a protein sequence; the sequence distance d1=j-i, d2=k-j, and d3=l-k. For each equivalent motif in family structures, the residue names of aa_i aa_j aa_k aa_l are identical, but the values of d1, d2, d3 may vary among different structures. In this case, we can define a consensus sequence pattern

m= aa_i (d1) aa_j (d2) aa_k (d3) aa_l,

which is similar to the patterns defined in PROSITE. d1, d2, d3 could be a range to allow the variance.

For example, we assume the motifs in the four protein structures in Figure 4 are {H, C, Q, S;7, 15, 30}, {H, C, Q, S; 5, 16, 27}, {H, C, Q, S; 6, 16, 26} and {H, C, Q, S; 7, 16, 28}, respectively. Then the consensus sequence pattern could be defined as H(5,7)C(15,16)Q(26,30). (5,7) means that the sequence distance between residue H and C could be any number between 5 and 7 (including 5 and 7).

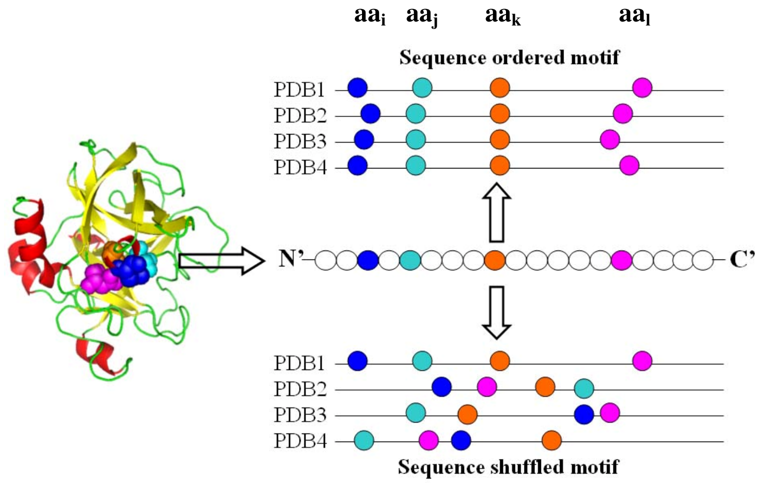


Figure 4. Derivation of sequence pattern from sequence ordered spatial motifs

III. RESULTS

The frequent spatial motifs identified are highly specific to the Papain-like cysteine protease family.

Motifs were identified using different support values, starting from the most strict value of σ =27/27 (meaning that we defined the patterns or motifs to be frequent if they occur in all 27 structures) (Figure 5). It was showed that no common motif was located for more than 23 of these 27 structures. By lowering the support value, one motif was found for σ=23/27, 3 motifs found for σ=22/27, 14 motifs found for σ=21/27, 29 motifs found for σ=20/27, 74 found motifs for σ=19/27 and 166 motifs was identified for σ=18/27. Table 3 documents the number of frequent motifs and the residues covered by these motifs in the 27 structures of Papain-like cysteine protease family with support σ =18/27, 19/27, 20/27, 21/27, 22/27, respectively. To test the uniqueness of PCP family's spatial motifs and evaluate the feasibility of using these motifs for family membership inference, we examined the frequency of these motifs in the background for each given support value σ, as described in the Methodology section. It was showed that for any support value we sampled (σ =18/27, 19/27, 20/27, 21/27 or 22/27), almost all background proteins have fewer motifs than the minimum found in any family member (Fig 6(a), 6(b), 6(c), 6(d) and 6(e)). For instance, when σ = 20/27, at most 29 frequent motifs were found contained in at least 20 members of the family. Except three proteins (1cv8, 1pxv and 1dki), each PCP protein contains at least 14 of such motifs, while in proteins outside the family (6500 proteins from culled PDB list, less than 90% identity), no one contains more than 2 of such 29 motifs (Fig 6(c)). To evaluate the statistical significance of family membership inference using these family-specific motifs, ROC curves were drawn to show specificity vs. sensitivity of the inference using the spatial motifs derived under different support (Fig 7). These ROC curves show that these motifs are very specific signatures of PCP family and can effectively discriminate between the family members and the background. For example, using the 29 motifs derived under the support value σ =20/27, if we set the cutoff value to be 14, which means a protein structure should be predicted as an member of the PCP family when it is found to contain at least 14 of those 29 motifs, the sensitivity of the prediction is over 88% while the accuracy is 100%.

Furthermore, Table 4 documents the composition of 29 motifs (with support σ =20/27), each of which consists of four or five residues. In addition, the *P-value* (equation 1) for each motif is smaller than 10^{-49}, which indicates that each of the motifs we found is highly linked to the PCP family.

The frequent motifs are either functionally or structurally important.

We further investigated the spatial distribution of the residues covered by these motifs, by mapping them in the structure of a PCP protein, 1CQD, shown in Fig 8. Here we studied the 29 motifs derived using support σ = 20/27. We observed that most of these residues are mainly located at the interface of the N-terminal and C-terminal domains, which is where the active site is located. Cys27 and His161 form the catalytic dyad, and several residues (Gln21, Asn181, Ser182, Trp183) near the catalytic center are known interact with the catalytic triad [39] and are detectable as motifs. Included in the covered residues also are those (Gly25, Trp28, Phe30, and Gly68) that form a hydrophobic pocket stabilizing the substrates and hence important to substrate specificity of the enzyme [39]. We also observed a pair of cysteines identified forming a disulfide bond (Cys24-Cys65). According to the literature [39], this disulfide bond may be crucial for the stability or folding of PCP family proteins.

Interestingly, the most frequent motif Q21-C27-H161-S182 (family frequency κ=23) contains the catalytic dyad Cys27-H161. Fig 9(a) and Fig 9(b) compare this motif in two PCP structures 1cqd and 3gcb, between which the pair-wise sequence identity is only 13%. This motif is showed to have very conserved geometry shape in the family. Furthermore, together with other three frequent motifs (C27-H161-S182-W183, Q21-C27-S182-W183 and N181-S182-W183-G191), it forms a six-residue packing network (Fig 10(a)), which captures the feature of the amino acids and hydrogen-bonding network (Fig 10(b)) in the active site of Papain-like cysteine proteases [39].

In summary, these results confirm that the clusters of functionally or structurally important residues have highly conserved geometry and indicate that our method to find frequent spatial motifs can identify these residues, and the geometric features of the packing patterns.

Outliers are found to be inactive zymogen forms or complexed with inhibitors.

In our study, three members of the family (1dki, 1cv8 and 1pxv) were found to contain none of such conserved motifs. We investigated these outliers case by case. 1dki is the crystal structure of the zymogen form of virulence factor SpeB from *streptococcus*. In this structure, a part of the N-terminal prosegment interacts with the active region so that the local structure is perturbed [40]. 1pxv is the PCP-like protease Staphopain complexed with its inhibitor Staphostain. The inhibitor directly contacts with the active region and changes the structure substantially [41]. 1cv8 is the crystal structure of a thiol protease from *Staphylococcus Aureus* V-8 with the E-64 inhibitor, which substantially disturbs the conformation in the active region of the protease [42]. Hence, the reason for being unable to identify the conserved spatial motifs in outliers is that the conserved local structure has been substantially changed due to interaction with other proteins or factors. Based on those observations on Papain-like cysteine protease family, we are convinced that our subgraph mining approach is applicable to find family specific motifs for SCOP protein families. We also noted that outliers were usually non-active protein structures with disturbed local structures (usually functional important). This indicates in future motif mining we should carefully analyze the dataset and expect the need for manual curation.

The PROSITE-like sequence pattern derived from structural motifs can identify the PCP sequences with 100% precision.

We derived PROSITE-like sequence pattern from the frequent spatial motifs using the 29 motifs as described in the Methodology section, and used the pattern to search for Swiss-Prot database. The three outliers were excluded from the consideration. After comparing these motifs in the remaining 24 structures, a consensus sequence pattern “Q-x(5)-C-x(132,170)-H-x(16,24)-N-S-W-x(4)-G-x(2)-G” was assembled using motifs Q21-C27-H161-S182, N181-S182-W183-G191 and S182-W183-G188-G191. According to Swiss-Prot, there is a total of 169 class one Papain-like Cysteine protease

[50]. Mining for the 178,772 sequences in Swissprot with this pattern, we have identified 157 out of the 169 Papain-like Cysteine protease sequences and nothing else. In other words, the *precision* of our method, which is defined in our context as the ratio of the number of true positives in our identified sequences and the total number of identified sequences, is 100% and the *recall* of our method, which is defined as the ratio of the number of true positives in our identified sequences and the total number of papain-like Cysteine protease in Swissprot is 92.8%.

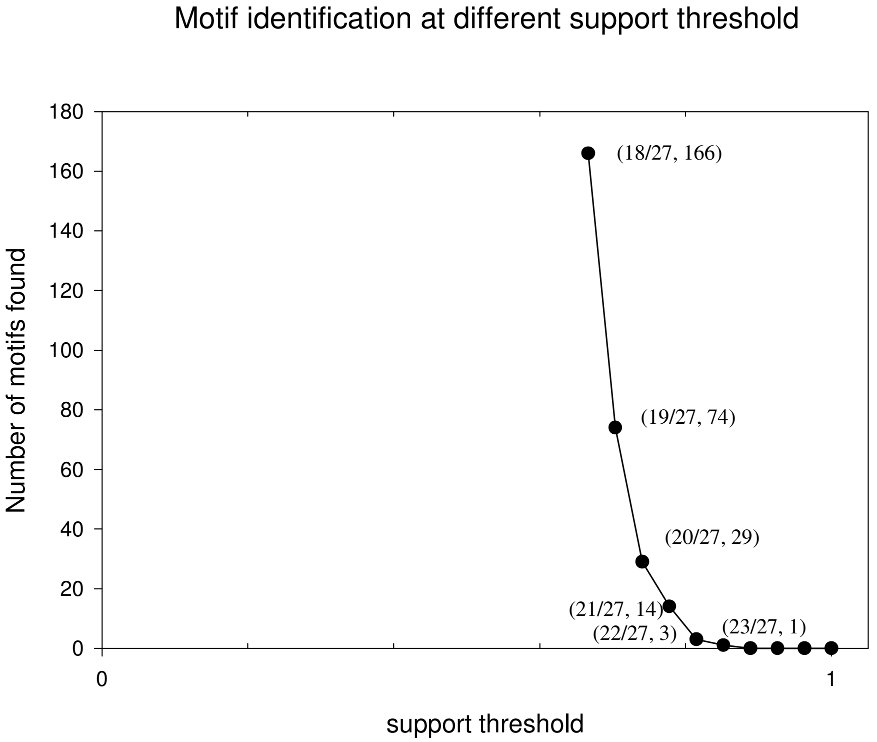


Figure 5. . Motif identification at different support value

Fig 5. Motif identification at different support value. The support values used here are σ = 27/27, 26/27, 25/27, 24/27, 23/27, 22/27, 21/27, 20/27, 19/27, 18/27, respectively. The coordinate vale for each data points is indicated.

Table 3 Numbers of frequent motifs and the residues covered by all these motifs in each of the 27 structures from PCP family

	PCP	# motifs (σ=18/27)	# of residues	# motifs (σ=19/27)	#of residues	# motif (σ=20/27)	#of residues	# motif (σ=21/27)	#of residues	# motif (σ=22/27)	# of residues
1	1cqd	166	28	74	23	29	19	14	12	3	6
2	1iwd	166	26	74	22	29	23	14	12	3	6
3	1s4v	166	26	74	22	29	19	14	12	3	6
4	1yal	166	27	74	23	29	19	14	12	3	6
5	1khq	166	31	74	27	29	18	14	12	3	6
6	1fh0	166	26	74	22	29	19	14	12	3	6
7	1o0e	166	26	74	22	29	19	14	12	3	6
8	1cs8	163	28	72	23	29	18	14	12	3	6
9	1ppo	161	27	73	23	29	20	14	12	3	6
10	1me4	160	25	71	21	29	19	14	12	3	6
11	1qdq	158	24	71	24	29	18	14	12	3	6
12	1mem	157	22	72	22	29	18	14	12	3	6
13	1the	155	20	71	20	28	19	13	12	3	6
14	1gmy	150	19	66	19	28	17	14	12	3	6
15	1m6d	148	26	70	22	27	17	13	11	3	6
16	1gec	111	28	55	24	24	20	13	11	3	6
17	1nqc	99	24	53	20	24	16	13	11	3	6
18	8pch	98	26	43	21	21	19	11	11	2	5
19	2act	98	30	56	26	21	19	10	12	1	4
20	1deu	82	23	44	18	20	12	12	11	3	6
21	1jqp	74	24	36	20	15	17	8	11	3	6
22	1k3b	68	26	38	21	14	14	8	10	1	4
23	2cb5	38	16	25	14	14	17	7	10	3	6
24	3gcb	13	13	7	11	14	9	8	8	3	6
25	1cv8	2	5	1	4	0	0	0	0	0	0
26	1dki	1	4	0	0	0	0	0	0	0	0
27	1pxv	0	0	0	0	0	0	0	0	0	0

Fig 6(a) Distribution of 166 spatial motifs (support = 18/27) in PCP family and in the Background

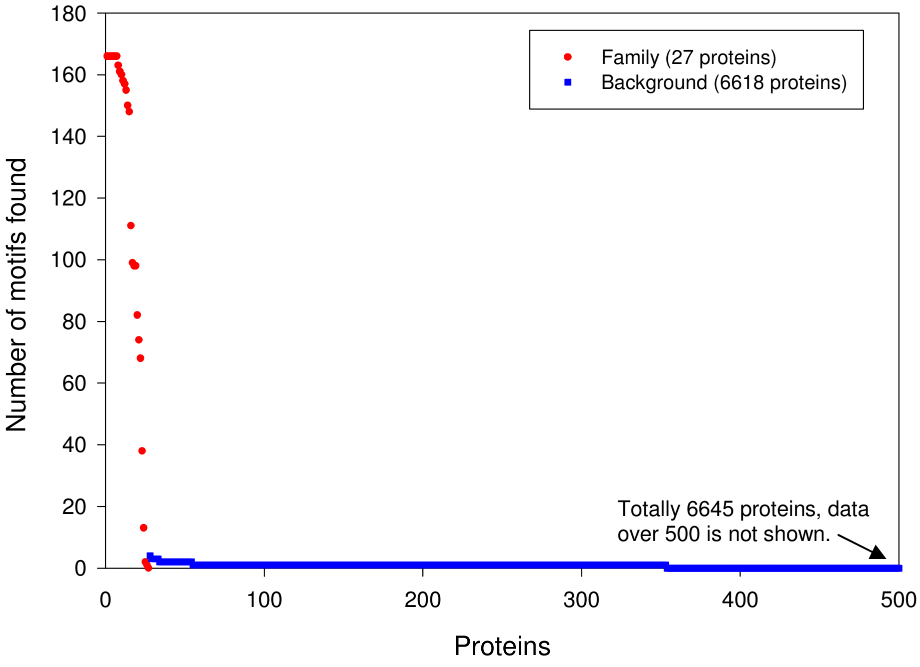


Fig 6(b) Distribution of 74 spatial motifs (support =19/27) in PCP family and in the backgound

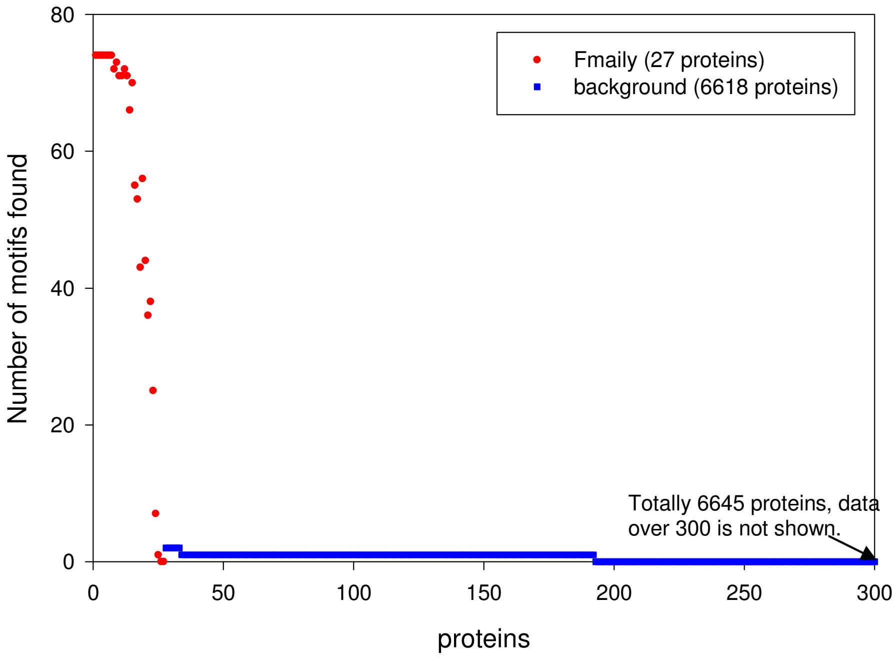


Fig 6(c) Distribution of 29 spatial motifs (support =20/27) in PCP family and in the background

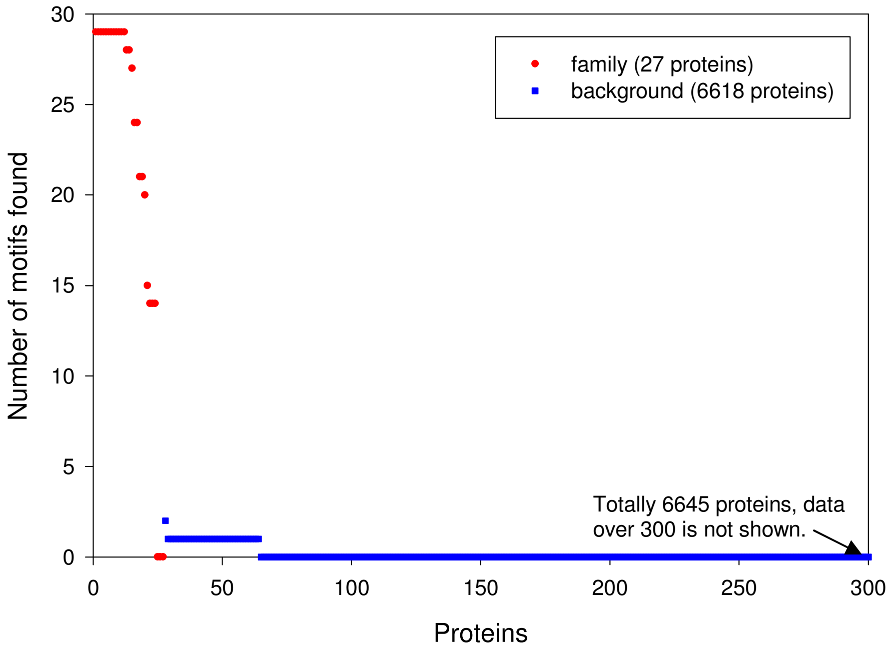


Fig 6(d) Distribution of 14 spatial motifs (support = 21/27) in PCP family and in the background

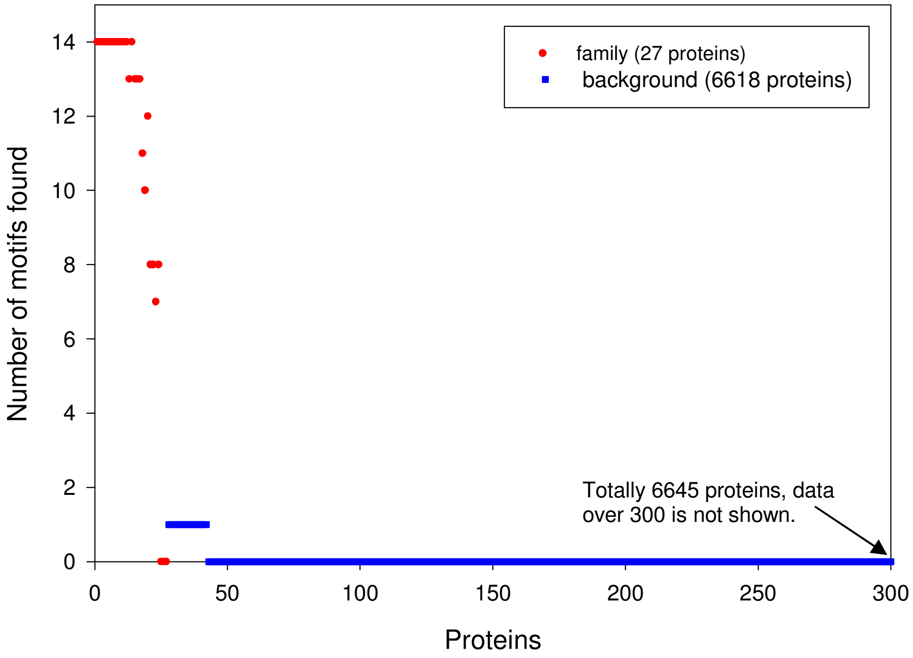


Fig 6(e) Distribution of 3 spatial motifs (support =22/27) in PCP family and in background

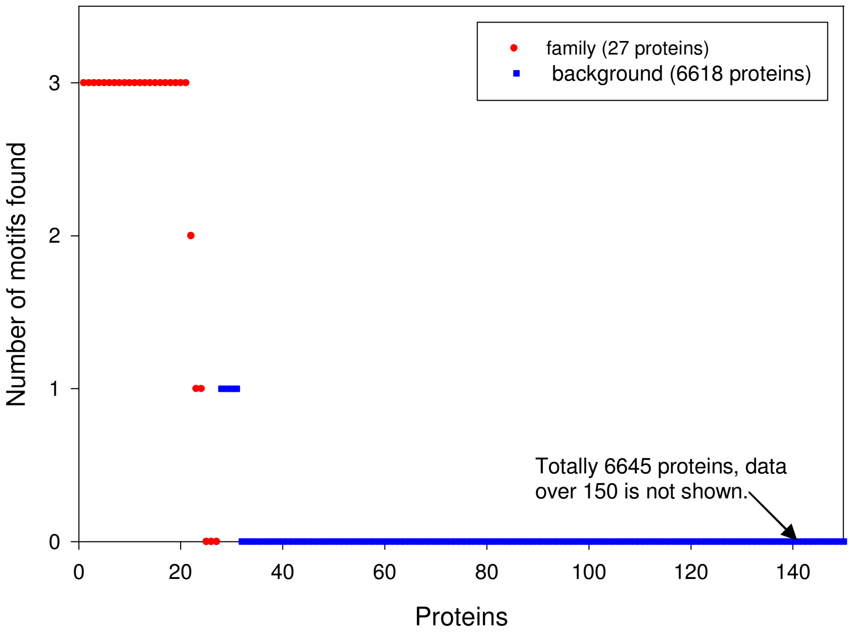


Figure 6. Distribution of PCP's frequent spatial motifs in the background and within the family

Fig 6. Distribution of PCP's frequent spatial motifs in the background and within the family.

(a). support σ =18/27, totally 166 motifs identified; in the background, one protein contains 4 motifs, 5 proteins contains 3 motifs, 21 proteins contain 2 motifs, and 299 proteins contain 1 motif;

(b). support σ =19/27, totally 74 motifs identified; in the background, 6 proteins contains 2 motifs, 159 protein contains 1 motif, 21 proteins contain 2 motifs, and 299 proteins contain 1 motif; support σ =19/27;

(c). support σ =20/27, totally 29 motifs; in the background, 1 protein contains 2 motifs, 36 proteins contain 1 motif.

(d). support σ =21/27, totally 14 motifs; in the background, 15 protein contains 1 motif;

(e). support σ =22/27, totally 3 motifs; in the background, 4 proteins are found to contain 1 of such motifs.

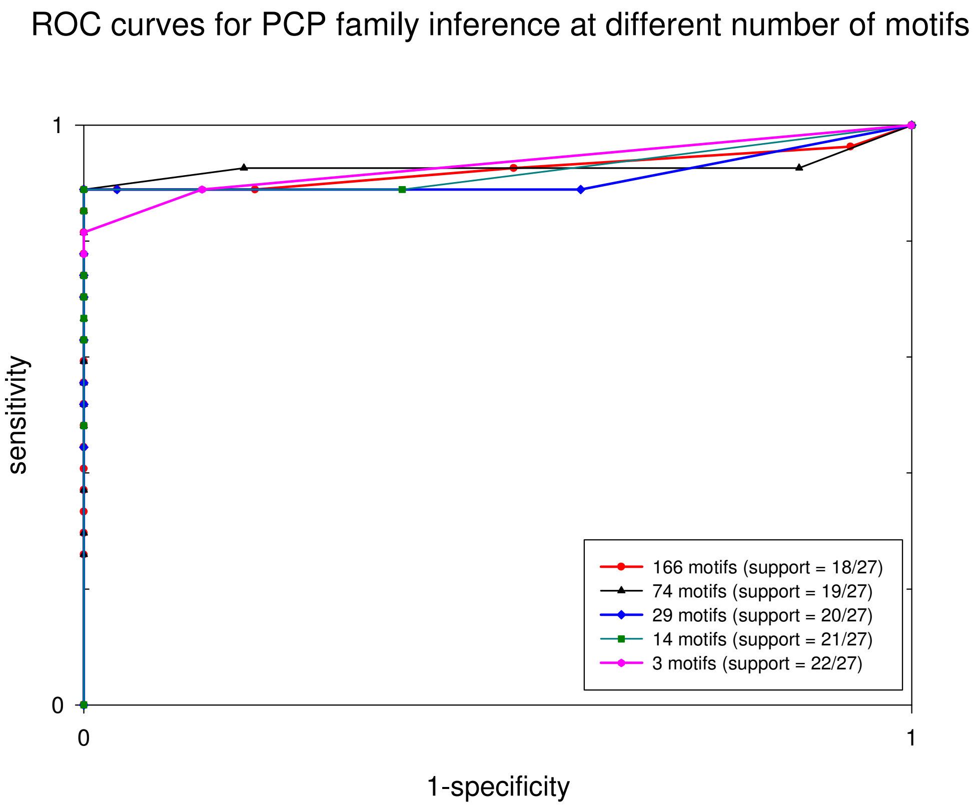


Figure 7. ROC curves for PCP family inference at different number of motifs

Fig 7. ROC curves for PCP family inference at different number of motifs. Sensitivity is calculated as the True Positive Rate (TPR), 1-specificity is calculated as False Positive Rate (FPR).

Table 4 Frequent spatial motifs identified in Papain-like cysteine protease family (ID:54002) N:27, σ:20/27

Motif	Composition	κ	δ	Motif	Composition	κ	Δ	Motif	Composition	K	δ
1	**HC**QS	23	3	**11**	WCSQ	21	0	**21**	WHCQS	20	0
2	FSQC	22	3	**12**	WSFC	21	2	**22**	WFCSQ	20	0
3	FQCG	22	10	**13**	WWGS	21	1	**23**	WFCQG	20	0
4	W**HC**S	21	0	**14**	WHCQ	21	0	**24**	WFCG	20	0
5	WCQG	21	0	**15**	SGQN	20	3	**25**	HCSS	20	2
6	WGNS	21	3	**16**	WFQG	20	0	**26**	WHGC	20	2
7	WGSG	21	3	**17**	SGCC	20	1	**27**	HCSG	20	9
8	WFCS	21	2	**18**	FQCG	20	2	**28**	WGFQ	20	7
9	WFCQ	21	0	**19**	WFSQ	20	7	**29**	WWGG	20	4
10	HCQG	21	6	**20**	CCGG	20	4				

N: total number of structures included in the data set. σ: the support threshold used to obtain frequent spatial motifs. Composition: the sequence of one-letter residue codes for the motif. κ: the actual support value of a motif in the family. δ: the background frequency of the motif. The P-values for all the motif are smaller than 10^{-49}. The motifs were sorted by their actual support values in descending order. The bold **HC**s indicate the catalytic dyad.

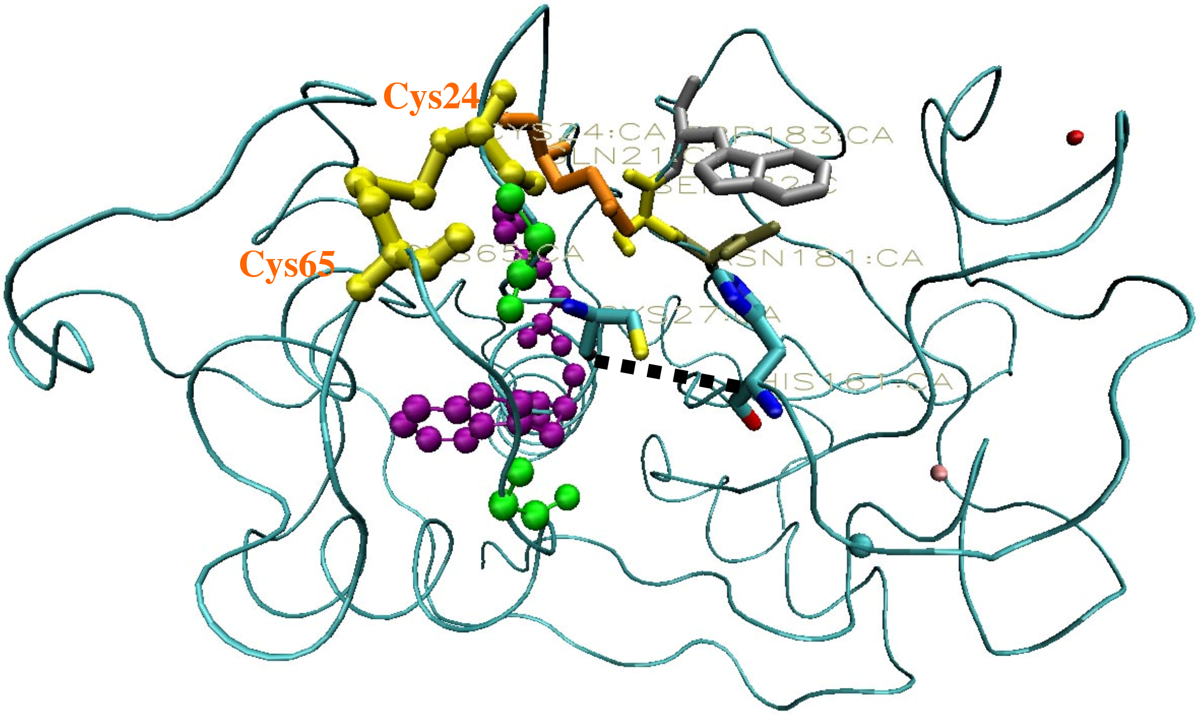
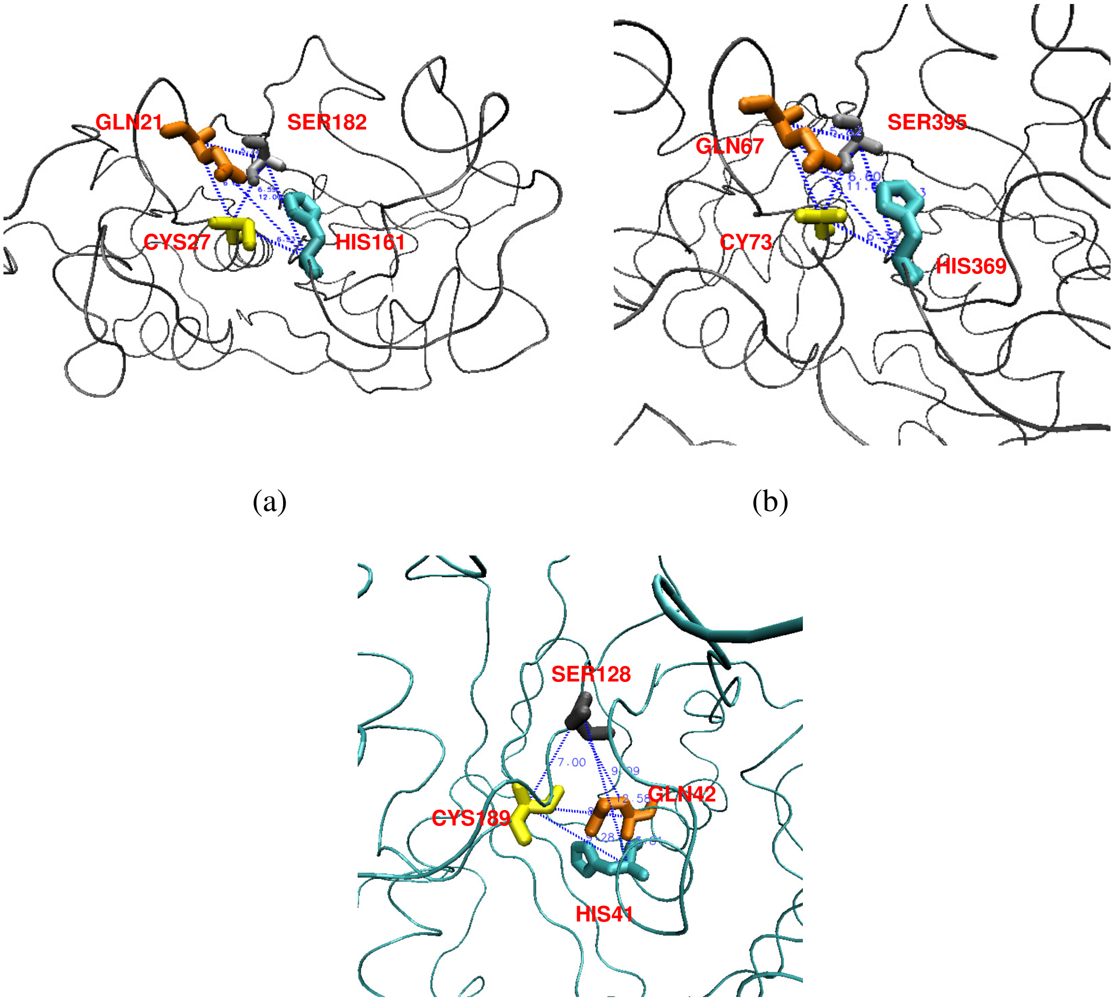


Figure 8. Spatial distribution of residues found in 29 motifs within protein 1CQD

Fig 8. Spatial distribution of residues found in 29 motifs within protein 1CQD. The residues in catalytic dyad CYS27-HIS161 are connected by a black dotted line and important surrounding residues GLN21, ASN181, SER182, and TRP183 are labeled, these residues are drew in Licorice representation. The residues forming hydrophobic pocket are drew in CPK method colored in purple. The disulfide formed between Cys24 and Cys65 is in yellow.



(a)

(b)

(c)

Figure 9. Illustration of motif HCQS

Fig 9. Illustration of motif Q21-C27-H161-S182 in PCP members 1CQD (a), 3GCB (b) and in a background protein 1LJ8 (c). The four amino acids are shown in Licorice representation, and labeled by the residue name. Edges are drawn in blue dotted line to connect the Ca atoms of each pair of residues, the lengths (in angstrom) of the edge is indicated.

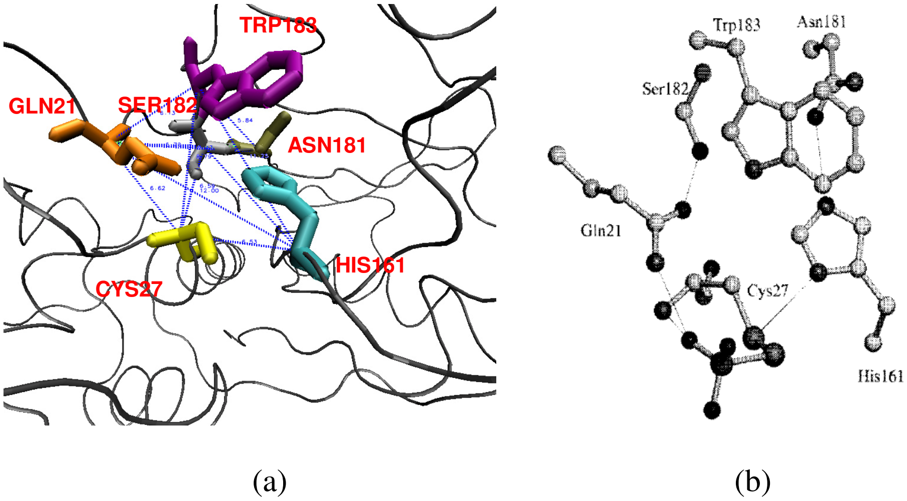
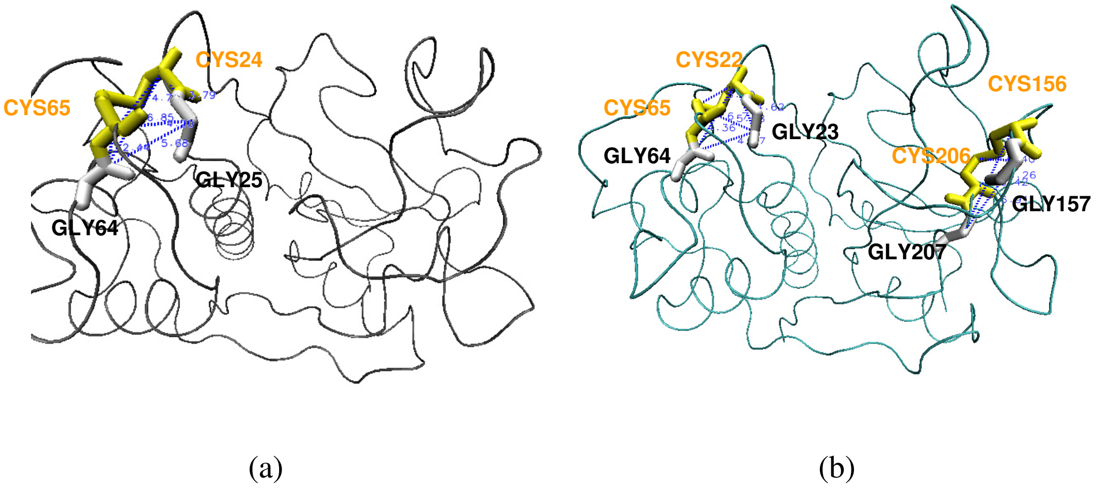


Figure 10. . Comparison of a motif cluster with the hydrogen-bonding network at the active site in a PCP protein 1CQD.

Fig 10. Comparison of the motif cluster (a) with the hydrogen-bonding network (b) at the active site in a PCP protein 1CQD. (a) A cluster of motifs at the active site. The six amino acids are shown Licorice representation, and labeled by the residue name. Edges are drawn in blue dotted line to connect the Ca atoms of residues, the lengths (in angstrom) of these edges are indicated. (b) The amino acids and hydrogen-bonding network of the active site in 1CQD (modified from [39]).



(a) (b)

Figure 11. Example of a motif (GCCG) with multiple occurrences within one structure

Fig 11. Example of a motif (GCCG) with multiple occurrences within one structure. (a) motif GLY64-CYS65-CYS24-CLY25 in PCP protein 1CQD; (b) motif GLY64-CYS65-CYS22-GLY23 and GLY207-CYS206-CYS156-GLY157 in PCP structure 2ACT. Residues forming the motifs are shown in shown Licorice representation, and labeled by the residue name.

IV. CONCLUSION AND DISCUSSION

In this study, we report the application of a novel frequent subgraph mining algorithm to retrieve conserved spatial motifs from protein 3D structures of Papain-like cysteine protease family. Each of the frequent spatial motifs we identified were found highly specific to the PCP family, measured by P-value<10^{-49}. And we also showed that the combination of these family specific motifs can discriminate the PCP family members and the background with very good sensitivity and predicative accuracy. By mapping the residues covered by these motifs to the structure, these residues are shown either structurally important or functionally important, such as forming the active sites and the hydrophobic pocket that is important to the substrate specificity. A PROSITE-like sequence pattern assembled by using these structural motifs can identify the PCP sequences in Swiss-Prot database with 100% precision and good recall. These suggest that structurally and functionally specific amino acid packing patterns or motifs can be discovered by computational and statistical geometry analysis of protein structures and used to annotate novel protein structures. Certainly, we need to further validate this by applying the method to more protein structural or functional families.

Since our method uses 2D graphs to represent the three dimensional protein structures, it certainly would miss some features of 3D objects. One of the features that our method is not able to identify is chirality. Fig 9 gives such an example. Fig 9(a) and Fig 9(b) compare the conserved four-residue motif HCQS in two different PCP structures.

These two tetrahedrons could be superimposed onto each other perfectly. Fig9(c) shows the occurrence of this motif in a background protein 1LJ8. This instance of the motif, which is an enantiomer of those in PCP structures, is still accepted as a match by our method. Thus, our method couldn't discriminate between the enanntiomers.

Furthermore, difficulties and problems are found through the case study on PCP family. Since the frequent subgraph mining algorithm relies on sampling the local structural features, the method is quite sensitive to the small perturbations in the local structure which result from interacting with other proteins (such as inhibitor) and crystal packing, etc. Since the PDB database contains many structures of inactive forms that always have locally structural change in the conserved region, it poses much trouble to systematic analysis of the PDB database with our method. Additionally, the mutations and chemical modifications that extensively exist in the PDB structures also bring to the automatic analysis a big challenge.

In summary, the results from the case study on PCP family suggest that that structurally and functionally specific amino acid packing patterns or motifs can be uncovered by computational and statistical geometry analysis of protein structures. These protein family specific motifs can be used to recognize the family members. In future studies, we will apply the approach to protein families defined by various protein structural and functional categories such as SCOP, Enzyme classification, DNA-binding proteins etc. The accumulation of significant motifs characteristic of known protein functional and structural families will aid the annotation of protein structures resulting from structural genomics projects, as well as facilitate our understanding the role of protein local structures in protein function and protein evolution.

V. REFERENCE

1. M. R. Chance, A. R. Bresnick, S. K. Burley, J. S. Jiang, C. D. Lima, A. Sali, S. C. Almo, J. B. Bonanno, J. A. Buglino, S. Boulton, H. Chen, N. Eswar, G. He, R. Huang, V. Ilyin, L. McMahan, U. Pieper, S. Ray, V. M, and L. K. Wang. Structural genomics: a pipeline for providing structures for the biologist. Protein Sci, 11:723-38, 2002.

2. T. C. Terwilliger, G. Waldo, T. S. Peat, J. M. Newman, K. Chu, and J. Berendzen. Class-directed structure determination: foundation for a protein structure initiative. Protein Sci, 7:1851-1856, 1998.

3. H. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. Bhat, H. Weissig, I. Shindyalov, and P. Bourne. The protein data bank. Nucleic Acids Research, 28:235-242, 2000.

4. J. S. Richardson. Class-directed structure determination: foundation for a protein structure initiative. Adv Protein Chem, 34:167-339, 1981.

5. A. Murzin, S. Brenner, T. Hubbard, and C. Chothia. SCOP: a structural classi_cation of proteins database for the investigation of sequences and structures. Journal of Molecular Biology, 247:536-40, 1995.

6. C. Orengo, A. Michie, S. Jones, D. Jones, M. Swindells, and J. Thornton. CATH - a hierarchic classi_cation of protein domain structures. Structure, 5(8):1093-1108, 1997.

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

8. P. J. Artymiuk, A. R. Poirrette, H. M. Grindley, D. W. Rice and P. Willett. A Graph-theoretic approach to the identification of three-dimensional patterns of amino acid side-chains in protein structures. J. Mol. Biol. 243: 327-344, 1994.

9. A. G. Murzin. Can homologous proteins evolve different enzymatic activities? Trends Biochem. Sci. 18, 403-405, 1993.

10. A.G. Murzin. Structural classi®cation of proteins: new superfamilies. Curr. Opin. Struct. Biol. 6, 386-394, 1996.

11. N. Nagano, C. Orengo, and J. Thornton. One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. Journal of Molecular Biology, 321:741-765, 2002.

12. H. Hegyi and M. Gerstein. The Relationship between Protein Structure and Function: a Comprehensive Survey with Application to the Yeast Genome. J. Mol. Biol. 288, 147-164, 1999

13. M. B. Swindells., A. Martin, D.T. Jones & C.A. Orengo. Contemporary approaches to protein structure classi®cation. Bioessays, 20(11): 884-891, 1998.

14. R. B. Russell. Detection of Protein Three-dimensional Side-chain Patterns: New Examples of Convergent Evolution. J. Mol. Biol. 279, 1211-1227, 1998.

15. A. C. Wallace., R.A. Laskowski and J. M. Thornton. Derivation of 3D coordinate templates for searching structural databases: application to Ser-His-Asp catalytic triads in the serine proteinases and lipases. Protein Sci. 5, 1001-1013, 1996.

16. C. Branden and J. Tooze. Introduction to Protein Structure. Garland Publishing, 1991.

17. N. V. Dokholyan, S. V. Buldyrev, H. E. Stanley, and E. I. Shakhnovich. Identifying the protein folding nucleus using molecular dynamics. J. Mol. Biol., 296:1183-88, 2000.

18. C. T. Porter, G. J. Bartlett, and J. M. Thornton. The Catalytic Site Atlas: a resource of catalytic sites and residues identified in enzymes using structural data. Nucl. Acids. Res. 32: D129-D133, 2004.

19. R. A. Laskowski, J. D.Watson, and J. M. Thornton. Protein function prediction using local 3D templates. J. Mol. Biol. 351: 614–626, 2005.

20. Wangikar PP, Tendulkar AV, Ramya S, Mali DN, Sarawagi S. Functional sites in protein families uncovered via an objective and automated graph theoretic approach. J Mol Biol. 326(3):955-78, 2003.

21. B. Delaunay. Sur la sph`ere vide. A la memoire de Georges Voronoi. Izv. Akad. Nauk SSSR, Otdelenie Matematicheskihi Estestvennyh Nauk, 7:793–800, 1934.

22. F. Aurenhammer. Voronoi diagrams: A survey of a fundamental geometric data structure. ACM Comput. Surv.,b23(3):345–405, 1991.

23. R. Singh, A. Tropsha, and I. Vaisman. Delaunay tessellation of proteins. J. Comput. Biol., 3:213–222, 1996.

24. Zheng W, Cho S J, Vaisman I I, Tropsha A. A new approach to protein fold recognition based on Delaunay tessellation of protein structure. Pac Symp Biocomput 1997, 486-497.

25. Tropsha A, Carter C W, Jr., Cammer S, Vaisman I I. Simplicial neighborhood analysis of protein packing (SNAPP): a computational geometry approach to studying

proteins. Methods Enzymol 2003, 374: 509-544.

26. Carter C W, Jr., LeFebvre B C, Cammer S A, Tropsha A, Edgell M H. Four-body potentials reveal protein-specific correlations to stability changes caused by hydrophobic core mutations. J Mol Biol 2001**,** 311: 625-638.

27. Sherman DB, Zhang S, Pitner JB, Tropsha A. Evaluation of the relative stability of liganded versus ligand-free protein conformations using Simplicial Neighborhood Analysis of Protein Packing (SNAPP) method. Proteins. 2004 Sep 1;56(4):828-38.

28. Bandyopadhyay, D. and Snoeyink, J. Almost-Delaunay simplices: Nearest neighbor relations for imprecise points. In ACM-SIAM Symposium on Discrete Algorithms, 2004, 403–412.

29. Huan, J., Wang W., Bandyopadhyay, D., Snoeyink, J., Prins, J., Tropsha, A. Mining protein family specific residue packing patterns from protein structure graphs. In: Proc. 8th Annu. Intl. Conf. Res. Comput. Mol. Biol. (RECOMB), 2004, 308-315.

30. Huan, J., Wang W., Prins J. Efficient Mining of Geometric Patterns Using Graph-based Techniques, SIGKDD, 2005.

31. Huan, J., Wang, W., Prins, J., and Yang, J. SPIN: Mining Maximal Frequent Subgraphs in Graph Databases. Proc. 10th ACM SIGKDD Intl. Conf. Knowledge Discovery Data Mining, 2004.

32. Huan, J. Bandyopadhyay, D., Liu, J., Prins, J., Snoeyink, J., Tropsha, A., Wang, W. Identification of statistically significant residue packing motifs in protein families using a contact graph representation of protein structures. Bioinformatics. 2005, 00-09.

33. Lo Conte L, S. E. Brenner, T. J. Hubbard, C. Chothia, A. Murzin. SCOP database in 2002: refinements accommodate structural genomics. Nucleic Acids Res. 30(1):264-7, 2002.

34. L.A. Mirny, E. I. Shakhnovich. Universally conserved positions in protein folds: reading evolutionary signals about stability, folding kinetics and function. J. Mol. Biol, 291. 177-196, 1999

35. N. M. Luscombe , J. M. Thornton. Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity.J Mol Biol. 320(5):991-1009, 2002.

36. N. M. Luscombe, S. E. Austin, H. M. Berman, J. M. Thornton. An overview of the structures of protein-DNA complexes. Genome Biol. 2000;1(1):REVIEWS001. Epub 2000 Jun 9. Review.

37. S. Jones, D. T. Daley, N. M. Luscombe, H. M .Berman, J. M. Thornton. Protein-RNA

interactions: a structural analysis. Nucleic Acids Res. 29(4):943-54, 2001.

38. S. Jones, J. A. Barker, I. Nobeli, J. M. Thornton.Using structural motif templates to identify proteins with DNA binding function. Nucleic Acids Res. 31(11):2811-23, 2003.
39. K. H. Choi, R. A. Laursen, and K. N. Allen. The 2.1 angstrom structure of a cysteine protease with proline specificity from ginger rhizome, zingiber o_cinale. Biochemistry, 7, 38(36):11624-33, 1999.

40. T. F. Kagawa et al. *PNAS* 97(5), 2235-2240, 2000.

41. R. Filipek, et al. *JBC* 278(42), 40959-40966, 2003.

42. Hofmann, B., Schomburg, D., H.J, Hecht. *Acta Crystallogr.,Sect.A (Supplement)* 49 *pp.* 102, 1993

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

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

45. P. Bradley, P. S. Kim & B. Berger Trilogy: Discovery of sequence-structure patterns across diverse proteins. Proceedings of the National Academy of Sciences. 99, 8500–8505, 2002.

46. S. Schmitt, D. Kuhn & G. Klebe A new method to detect related function among proteins independent of sequence and fold homology. J. Mol. Biol. 323(2), 387–406, 2002.

47. J. P. Shaffer Multiple hypothesis testing. Ann. Rev. Psych. 561–584, 1995.

48. D. Bandyopadhyay, J. Huan, J. Liu, J. Prins, J. Snoeyink, W. Wang, A. Tropsha. Structure-based function inference using protein family-specific fingerprints. Protein Sci. 15(6):1537-43, 2006.

49. Tropsha A, Singh RK, Vaisman II, Zheng W. Statistical geometry analysis of proteins: implications for inverted structure prediction. Pac Symp Biocomput. 614-23, 1996.

50. Turk, V., Turk, B., Turk, D. (2001) *EMBO J*, 20(17). 4629-4633.

51. Otto, H.H., Schirmeister, T. (1997) *Chem. Rev.* 97, 133-171.