PROTEIN FUNCTION PREDICTION USING FAMILY-SPECIFIC STRUCTURAL MOTIFS

Tanarat Kietsakorn

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 School of Pharmacy (Division of Chemical Biology and Medicinal Chemistry).

Chapel Hill
2011

Approved by:

Alexander Tropsha, Ph.D.

Michael Jarstfer, Ph.D.

Scott Singleton, Ph.D.

Denis Fourches, Ph.D.

©2011
Tanarat Kietsakorn
ALL RIGHTS RESERVED

ABSTRACT

TANARAT KIETSAKORN: Protein Function Prediction using
Family-specific Structural Motifs
(Under the direction of Alexander Tropsha, Ph.D.)

Protein function prediction using structural motifs is expected to be more reliable and informative than using global sequences/structures or sequence motifs.

In the first part of this thesis, we report a novel application of two structural motif-based methods, FFSM and CASIM, for predicting family-specific structural motifs and conserved key residues in Metallo-dependent phosphatase (Metallophos) structures. We also introduced the novel function prediction approach based on 3D-1D Cumulative Support Profiles, which represents degree of conservation of amino acid residues specific to Metallophos family.

In the second part of this thesis, we present novel structural motif-based approaches for function annotation of protein tyrosine kinase (PTK) sequences. This is the first report of non-traditional function inference, from structure to sequence to function.

Compared to other state-of-the art methods, our approaches were able to reveal more comprehensive information such as the 3D structure of the potential active site including key residues.

ACKNOWLEDGEMENTS

First of all I would like to express my appreciation to my thesis advisor, Dr. Alexander Tropsha. My graduate study could not be completed without him. He has introduced me to the field of bioinformatics, my area of interest, and provided me with guidance, encouragement and patience throughout my graduate study.

Special thanks go to Dr. Denis Fourches. I am grateful for his time and effort in assisting this research project. His valuable help involved in nearly all aspects of my work; from the early state of developing my proposal until the final state of proofreading the thesis. In addition, I would like to thank him for allowing me to use his computational tool called 'CASIM', which effectively solved many interesting and challenging bioinformatics problems that I encountered during my thesis.

As well, I would like to thank our inter-departmental collaborations; Dr.Wei Wang, Dr. Jun Huan, Dr. Deepak Bandyopadhyay and all members in motifSpace project in UNC Computer Science Department. Their hard work and focus in developing the 'FFSM' method, another main bioinformatics tool used in my study, deserved to be noted.

I would like to thank the other two of my committee members; Dr. Michael Jarstfer and Dr. Scott Singleton for their support and valuable scientific advice.

I thank Dr. Ashutosh Tripathi and Stephen Bush for helping me proofread my thesis.

I also thank all my lab mates, students and staffs in the Division of Chemical Biology and Medicinal Chemistry for their friendship and support.

I would like to express my gratitude to Thai government for their financial support.

Finally, my gratitude was dedicated to my beloved parents, husband, daughter, sisters and the rest of my family for their unconditional love.

TABLE OF CONTENTS

LIST OF TABLES

Table

LIST OF FIGURES

Figures

LIST OF ABBREVIATIONS

1D	one-dimensional
3D	three-dimensional
AAT	Aspartate aminotransferase
AD	Almost Delaunay
ADH	Alcohol dehydrogenase
AFP	Automated Function Prediction
b	Maximum background occurrence
BSGC	the Berkeley Structural Genomics Center
Cα	alpha carbon
CASIM	Conserved Adjacent Simplex Miner
cPLA2	cytosolic phospholipase A_2
CSA	Catalytic Site Atlas
CRP	Catalytic Residue Prediction
CSP	Cumulative Support
DMSO	Dimethyl sulfoxide
DT	Delaunay tessellation
EC	Enzyme Classification scheme
ESA	Exposed surface areas
ETA	Evolutionary Trace Annotation
f	Minimum support
FFSM	Fast Frequent Subgraph Mining
GASPS	Genetic Algorithm Search for Patterns in Structures

GO	Gene Ontology
HMM	hidden Markov models
Metallophos	Metallo-dependent phosphatases
NAD^+	Nicotinamide adenine dinucleotide
PDB	Protein databank
PLA_2	Phospholipase A_2
PROTMAN	PROTein MANager
PTK	Protein tyrosine kinases
RMSD	Root-mean-square deviation
SCOP	Structural Classification of Proteins
SNAPP	Simplicial neighborhood analysis of protein packing
$sPLA_2$	secreted Phospholipase A_2

CHAPTER I

INTRODUCTION

1.1 Introduction to protein function prediction

The knowledge of protein function is necessary to understand the machinery of life and translate this knowledge into drug discovery. There has been an exponential increase in the number of available protein sequences and structures resulting from genome sequencing and structural genomics projects, respectively; however, the function of many proteins still remain unknown. Consequently, there is a growing challenge of developing computational tools to predict functions of these proteins of unknown functions and focus the costly and time-consuming experimental work towards hypothesis validation rather than random (or serendipitous) exploration.

1.1.1 Automated Function Prediction requires the 'gold standard' of functional label.

The first step for any Automated Function Prediction (AFP) tools is to clarify the definition of protein function. Proteins are essential biological macromolecules that perform their functions in every process within cells ranging from sub-cellular to the whole-organism level. Thus, the definition of protein function is not very well-defined and may be explained in various aspects. For example, function of protein kinases can be described by many cellular functions in which they are involved or by a smaller scope of molecular function as transferases[1]. To allow protein functions to be understood and predicted *in silico*, it is

important to provide the machine with a standardized functional term.

Databases of protein function classification have been created using specific terms for different aspects of protein function. The two most widely-used schemes for protein sequences are Gene Ontology (GO)[2] and Enzyme Classification scheme (EC)[3]. GO categorizes protein function by controlled protein annotation vocabularies in terms of molecular function, biological process and cellular component. Molecular function is referred to the task performed by an individual protein whereas biological process composes of a variety of molecular functions, and cellular component indirectly addresses protein function in the context of sub-cellular structures, location and macromolecular complexes. EC scheme is a 4-level hierarchical functional classification for enzymes, based on the type of chemical reactions they catalyze. Each protein is associated with an EC number, which consists of 4 digits, 1 for each level. The first digit represents 6 main chemical reactions that the enzymes catalyze (oxidoreductases, transferases, hydrolases, lyases, isomerases or ligases). The second and third numbers describe the subclass and sub-subclass of the overall reaction, whereas the last number usually reflects the substrate specificity of the reaction. While GO is applicable for a variety of proteins, EC is limited to enzymes only. However, most approaches for enzyme function prediction rely on EC annotation[4-8]. That is because EC annotation provides higher enzyme-annotation coverage, and has been used as a gold standard in most enzyme databases. Moreover, EC annotation can be related to GO annotation using the web service referred as "ec2go" provided by GO website[2].

Function classification of protein structures is less studied compared to those of protein sequences. For instance, there is an approach, PDBsum[9], which annotates functions of protein structures (PDB chains) according to GO term and EC number of their

corresponding UniProt sequences. In another effort, Bandyopadhyay[10] reported the application of Fast Frequent Subgraph Mining (FFSM) for function inference of protein structures using family and superfamily definitions of Structural Classification of Proteins (SCOP) database[11] to define protein function. SCOP classifies proteins based on their three-dimensional structural similarity through the levels of class, fold, superfamily, family, domain and species. Actually, the relationship between SCOP classification and protein function is not obvious. SCOP classification is based on global structural (fold) similarity of a single domain, not functional similarity. A single SCOP family may be related to more than one function. For instance, the SCOP family of AAT-like (AAT: Aspartate aminotransferase) corresponds to two remote functions; lygase and transferase. In some other cases, a given function occurs in different SCOP folds (Beta-lactamase). I believe that proteins can perform multiple tasks and some of them are performed by their substructure (e.g. motifs), which explains why one function can be detected in proteins with different SCOP folds. However, SCOP does not allow this interconnection. Since SCOP is a hierarchical classification, proteins with different fold types will always be classified into different superfamily and family. Therefore, SCOP fold, superfamily and family are not ideal levels for investigating the relationships between structures and functions. Although domains are the basic unit of protein structure, function and evolution, using SCOP-domain level for function annotation is appropriate for single domain proteins only. *In this thesis, I report novel approaches for predicting function of two protein groups; Metallo-dependent phosphatases (Metallophos) in Chapter 2 and protein tyrosine kinases (PTKs) in Chapter 3. EC annotation has been used in a case of PTKs. However, SCOP annotation has been applied to describe function of Metallophos members due to the following reasons: (1) their EC annotations are not well*

studied, (2) Methallophos family contains only one-domain proteins, and (3) their SCOP definition is well adopted by most bioinformatic studies.

1.1.2 Insufficiency of global sequence or structural similarities for protein function inference

Most of AFP tools assume that proteins with similar sequences or structures usually share common function. Consequently, the function of a protein of unknown function is typically inferred from its homologous proteins of known function. This classical approach for inferring protein function typically relies on sequence similarity analysis, also known as homology-based annotation transfer[12]. The most popular methods in this category are sequence similarity search tools such as BLAST[13], or profile-based similarity search tools based on profile hidden Markov models (profile HMMs)[14]. Thornton suggested that function inference by sequence similarity is most reliable when the pair-wise sequence identity is above 40%[15-17]. Skolnick reported a threshold of 40% and 60% sequence identity as cutoffs for accurate function transfer between proteins that respectively share first three digits and all four digits in EC classification scheme[4]. Therefore, the major limitation of homology-based annotation transfer appears when the sequence similarity falls below a certain similarity cutoff. However, there are known exceptions to those recommended global similarity rules. For instance, melamine deaminase and atrazine chlorohydrolase share 98% sequence identity, but catalyze different reactions[18]. The authors suggested that the nine amino acids that differ between those two proteins are indeed responsible for their functional difference.

It is well known that a three-dimensional (3D) protein structure is well conserved compared to its sequence[19]. Consequently, structure conservation, if detected, may sometimes provide critical clues for function inference even when sequence-based

approaches fail or become unreliable. For instance, MJ0882, a hypothetical protein from *Mehanococcus jannaschii* has no detectable sequence similarity to any protein sequence in the Protein Databank (PDB). However, global structure comparison based on fold similarity by DALI[20] suggested that the protein was probably a methyltransferase because its crystal structure had a similar fold to many methyltransferases in the PDB, and this activity was subsequently confirmed by biochemical experiments[21]. However, it should be pointed out that proteins with similar folds may also have different functions[15]. For instance, proteins with the TIM barrel fold may carry out more than 60 different enzymatic functions. On the other hand, fold similarity does not always imply similar function; for instance, different o-glycosyl glucosidases belong to seven fold types[19]. Obviously, neither sequence nor global structure similarity is globally applicable for reliable function inference. A probable cause for those exceptions are as the following: proteins with highly similar sequences or structures may not share the same function because of divergent evolution where residues responsible for function have changed while most of their sequences or structures remained unchanged. In contrast, two proteins with low overall sequence or structural similarity may have the same function because their active sites could have remained conserved throughout the evolution unlike their remaining regions. *This assumption leads to the emerging concept of function inference through motifs (local similarity), the main focus of this thesis.*

1.1.3 The importance of motifs (local similarity) for function inference

Although global sequence or structure comparison approaches continue to be popular for function inference, some experimental evidence suggests that protein function can be correlated with the presence of local patterns of amino acid residues, or *motifs* shared by

proteins with similar function either at the sequence or structure levels. Motifs could be defined as highly conserved sets of residues that form similar patterns and often represent functionally important regions such as active or binding sites, or regions defining the overall protein fold. Therefore, local similarity analysis to identify either sequence or structural motifs could be useful for predicting protein function and/or identifying functionally significant sites.

Typically, sequence motifs are derived from multiple sequence alignments of proteins with similar function. Of the approaches implementing these motifs, PROSITE patterns[22] is the most widely used for inferring function. Other methods such as PRINTS[23] and Scan2S[24] were aimed to improve the predictive performances of PROSITE patterns. PRINTS uses the occurrence of multiple motifs (forming *fingerprints*) to reach better sensitivity whereas Scan2S includes secondary structure constrains to achieve better precision. However, all of these methods are capable of detecting only sequence-ordered motifs.

The review of AFP approaches by Chen[25] suggested that sequence-based approaches were able to provide high confidence only when pair-wise sequence identity between two proteins was in the safe zone (higher than 40%). However, when pair-wise sequence similarity fell into the twilight (20-30%) and midnight zones (below 20%), the AFP methods based on global structure similarity and local structure similarity were more applicable, respectively. Function inference by *local structural motifs* is likely to be more reliable than using global structural similarity because 3D arrangements of functionally important residues (e.g., in the active sites) are significantly more conserved than the entire fold[26]. Structural motifs are thus represented as local 3D templates containing conserved amino acid residues. The identification of motifs that are conserved among a given family of proteins requires the

systematic analysis of their 3D structures. For example, the Genetic Algorithm Search for Patterns in Structures (GASPS)[27] deduces motifs from multiple sequence alignments of homologous proteins. These motifs are then converted into 3D patterns by SPASM[28], which represents each residue in the motif by two points: the Cα carbon atom and the side-chain geometrical centroid. MSDmotif[29] uses enriched motifs integrating 3D structurally conserved patterns and super-secondary structural and sequence motifs; these motifs are classified into 13 types, based on specific patterns of hydrogen bonding, φ/ψ and χ angles. Evolutionary Trace Annotation (ETA)[30] identifies evolutionary important residues from phylogenetic trees of homologous protein sequences, and then maps those residues onto the structure to generate 3D templates. However, most of the structural motif based approaches described above rely on multiple sequence alignments. Thus, these methods inherit the limitations of sequence-motif based approaches. The sequence-independent AFP methods have been aimed to obtain information missing at the sequence level. Only few methods are in this category: (1) the 3D template searches[31, 32] (enzyme active site template, ligand binding site template, DNA-binding site template and reverse template searches), and (2) FFSM [33, 34].

1.2 Overview of Chapter 2

Structural motifs are considered much more conserved and informative than their corresponding sequence motifs. However, only few structural motif-based approaches have been addressed the problem of Automated Function Prediction (AFP) using the information of 3D protein structures alone. The limitation is due to the difficulty of local similarity comparison. *In this chapter, we report an application of two sequence-independent methods, FFSM and a novel CASIM (Conserved Adjacent Simplex Miner)*[35]*for predicting family-*

specific-structural motifs and conserved key residues. These two methods were implemented based on computational geometry technique known as Delaunay Tessellation (DT)[36, 37]. FFSM was developed earlier in collaboration with colleagues in the UNC Computer Science Department. Currently, the method was applied for predicting protein family-specific structural motifs only[10, 33, 34, 38, 39]. CASIM has been developed and implemented in the PROTMAN (PROTein MANager) program in our research group, and its application is first reported herein. We present a successful case study of Metallophos family. We are able to identify the Metallophos family specific residue packing patterns (Metallophos-specific motifs) using FFSM and CASIM. The identified Metallophos-specific motifs were found at the metal-binding active sites in the training-set members and the test proteins of known functions. We discuss the complementarities between the two approaches for the identification of family specific packing motifs and their use for the automated predicting function and conserved key residues (likely functionally important residues) for proteins of unconfirmed functions.

1.3 Overview of Chapter 3

The number of protein sequences that have no function annotation are greatly exceeds the number of their structures. Thus, function prediction of protein sequences is critical. Currently, only sequence-based approaches have been used for function prediction of proteins at sequence levels while publicly available structural motif-based methods including FFSM are applicable for protein function prediction at structure levels only due to the difficulty of extracting meaningful information from protein structures. In this chapter, we develop approaches being able to apply family-specific structural motifs originally extracted

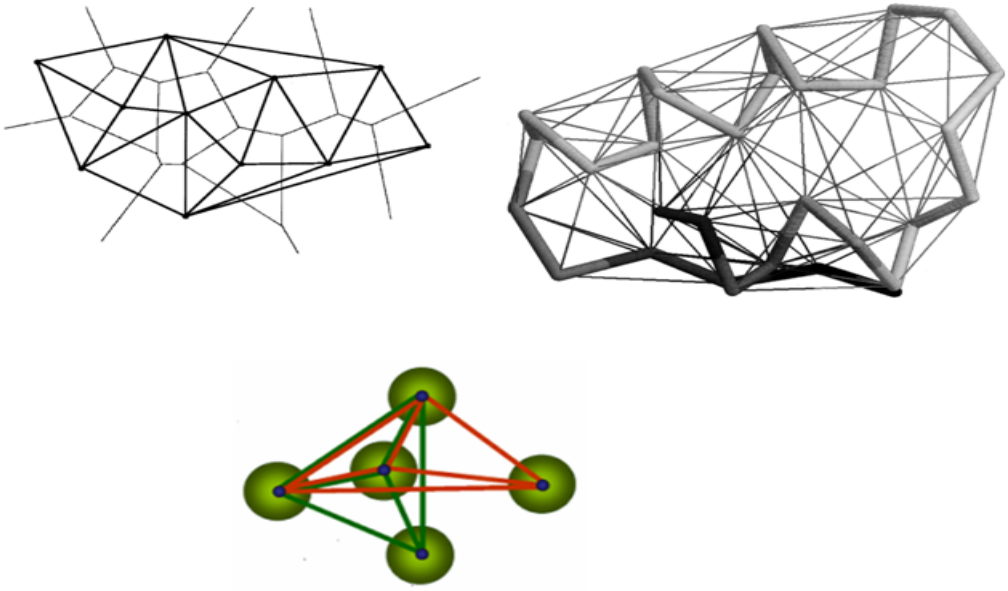
from protein structures to predict function and functionally important residues of protein sequences. We applied FFSM to identify structural motifs (frequent subgraphs) conserved in a given protein family. However, structural motifs represent three-dimensional structures; thus they cannot be directly mapped onto the linear string of protein sequences. We converted those identified structural motifs into sequence patterns, which can be easily matched on protein sequences by uncomplicated text mining algorithm. Our approaches were successfully applied for function inference of PTK family.

1.4 Introduction to Delaunay and Almost-Delaunay tessellations

Delaunay tessellation (DT) is a fundamental computational geometry structure related to the Voronoi tessellation. In Voronoi diagram[40], the space is partitioned into cells, each of which consists of one node and the points that are nearest to that node than to any other nodes. DT connects nodes in Voronoi diagram. DT and Voronoi diagram in two dimensions are illustrated in **Figure 1.1A**. In three-dimensional space, DT generates an aggregate of space-filling, non-overlapping irregular tetrahedra or *simplices*, preserving an empty sphere property. Each Delaunay simplex defines objectively and uniquely four nearest neighbors as vertices of a tetrahedron. Logically, the entire Delaunay structure could be described as a network of contacts between nodes thus forming a connected graph.

Our research group has pioneered the use of DT in protein structure analysis[36, 37]. The aggregate of Delaunay simplices representing a protein could be also regarded as a network of contacts between residues that can be described by a connected graph where residue-vertices can be labeled by their conventional names and the edges can be labeled by the physical distance between points representing residues (see **Figure 1.1B**). A protein

structural family can then be described by a family of labeled graphs where each graph represents a protein member of the family. However, protein structure coordinates are imprecise. The errors may occur due to measurement imprecision or atomic motions. Since DT represents a node as a certain point, it is not robust to perturbation. Small change in point coordinates may change the set of nearest neighbors. To improve DT algorithm for protein structure analysis, Bandyopadhyay and Snoeyink introduced a DT-based approach called Almost Delaunay (AD)[41]. Instead of presenting each amino acid as a precise point, AD allows the movement of a point with parameter ε while still preserving the empty sphere property. The protein graphs constructed by AD are termed AD edge graphs (see **Figure 1.1C**), which contains both DT edges and the new AD edges. It is reported that the AD approach helps recover greater number of more specific motifs that DT with a relatively minor loss in computationally efficiency.



(C)

Figure 1.1: (A): Voronoi tessellation and Delaunay tessellation (DT) in 2D space[42]. The Voronoi polihedra is shown by thin lines and the corresponding DT is shown by thick lines. **(B):** DT of crambin (PDB ID: 1crn) in 3D space[42]. The backbone of the protein is shown by thick lines whereas

the DT is shown by thin lines. (**C):** Illustration of AD edges; vertex can move within bounding sphere with radius ε[41].

1.5 Introduction to FFSM

Based on the assumption that amino acid residues responsible for protein function are encoded in family structural motifs, Huan *et al* at UNC developed the FFSM method focusing on finding structural motifs in protein families[33, 34]. FFSM identifies recurrent *frequent* subgraphs from family members modeled as AD graphs. Accordingly, those family-specific subgraphs or *fingerprints* correspond to structural motifs in protein structures. It is shown that this method was capable of capturing local packing motifs characteristic of protein structural and functional families[10, 33, 34, 38, 39]. The concept of FFSM can be briefly described as follow. FFSM represents each protein structure in the family of interest as an AD graph consisting of nodes and edges. Every node in the graph characterizes distinct amino acid residues in that protein and has the residue type as its label. Edges are distinguished and labeled according to AD algorithm and their lengths.

FFSM restricts the subgraph (the sub-structural pattern of a protein graph; see **Figure 1.2**) to a fully rigid interconnected subgraph referred as a *clique*. A clique is a graph where each node has degree n-1 where n is the number of nodes and degree is the number of edges incident with it. According to FFSM implementation, the sub-structural patterns identified by FFSM are not limited to only quadruplets. To eliminate the redundant subgraphs, FFSM selects only the maximal frequent subgraph (a graph that is not part of any larger frequent subgraph).

A subgraph of the entire AD graph is considered frequent if its '*minimum support*' value (i.e., a fraction of family members that contain this subgraph) is higher than a user-defined threshold (e.g., 90%). However, those frequent subgraphs become family-specific

subgraphs or *fingerprints* if and only if they are rarely found in the other proteins of a diverse '*background*' database (other proteins outside a target family).

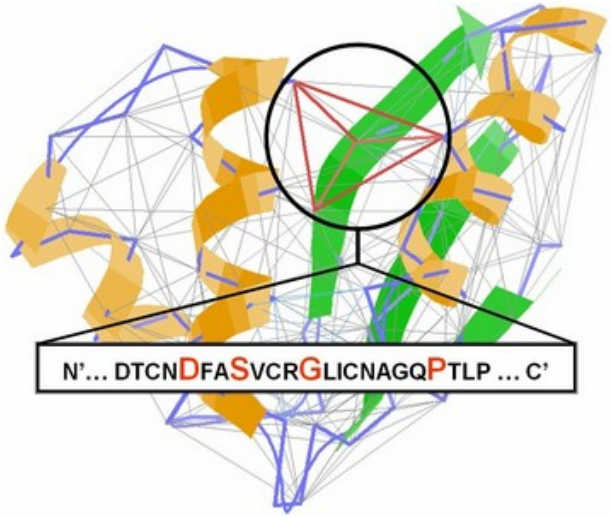


Figure 1.2: An Almost Delaunay (AD) graph of a protein structure; a subgraph DSGP (showed in red) is a sub-structural pattern of that protein graph[43].

1.6 Introduction to CASIM

The novel CASIM approach, implemented in the PROTMAN (PROTein MANager) program package by Fourches[35], has been developed to improve the performance of DT/AD based approaches for effective identification of family structural motifs.

In this CASIM method, each protein structure in the family of interest is modeled as a DT or AD graph (vertices are Cα atoms or side chain centroid of amino acid residues). Unlike FFSM that defines a motif as a fully interconnected subgraph, CASIM describes motifs as ensembles of neighboring Delaunay tetrahedral (see **Figure 1.3**). Thus, we expected to recover motifs missed by FFSM or vice versa. In CASIM, a motif can involve one or several neighboring tetrahedra sharing a common face, a common edge, a common vertex or having a spatial proximity according to a user-defined geometrical distance cutoff between the centroid of Delaunay simplices (e.g., 10Å) (see **Figure 1.4**). The neighborhood

of all these Delaunay quadruplets is determined using a second Delaunay tessellation of the tetrahedron centroid. Each tetrahedron has a unique nomenclature based on the alphabetical order of its residue-vertices. Similarly, motifs involving several tetrahedra possess a unique and single nomenclature based on their composition and the alphabetical order as well. For instance, a motif shown in **Figure 1.4** involves four neighboring Delaunay tetrahedra encompassing eight residues: DGGL, GGLL, GHIL and CHIL; thus, its unique name is CHIL-DGGL-GGLL-GHIL. Moreover, the motifs retrieved by CASIM provide additional information. Each CASIM motif is characterized by a series of constitutive and geometrical descriptors to enhance its specificity: the motif's exposed surface areas (ESA); its overall volume; number of involved residues; contact types between residues (peptide bond or geometrical proximity edge); the chirality of its constitutive tetrahedra; the overall SNAPP score[44]**;** chain characteristics (single chain or interfacial motif); presence/absence of organic ligands inside or in the proximity of the motif. In addition, all combinations of sub-motifs [CHIL-DGGL, DGGL-GGLL, CHIL-DGGL-GGLL, etc. for the example shown in **Figure 1.3**] involving one, two or three tetrahedra are also investigated to define families of motifs.

In order to define family-specific motifs, CASIM adopts the concept of FFSM approach described under Section 1.5. The motifs are specific to the family if they are found in significant numbers of protein members of the family and are rarely found in other proteins.

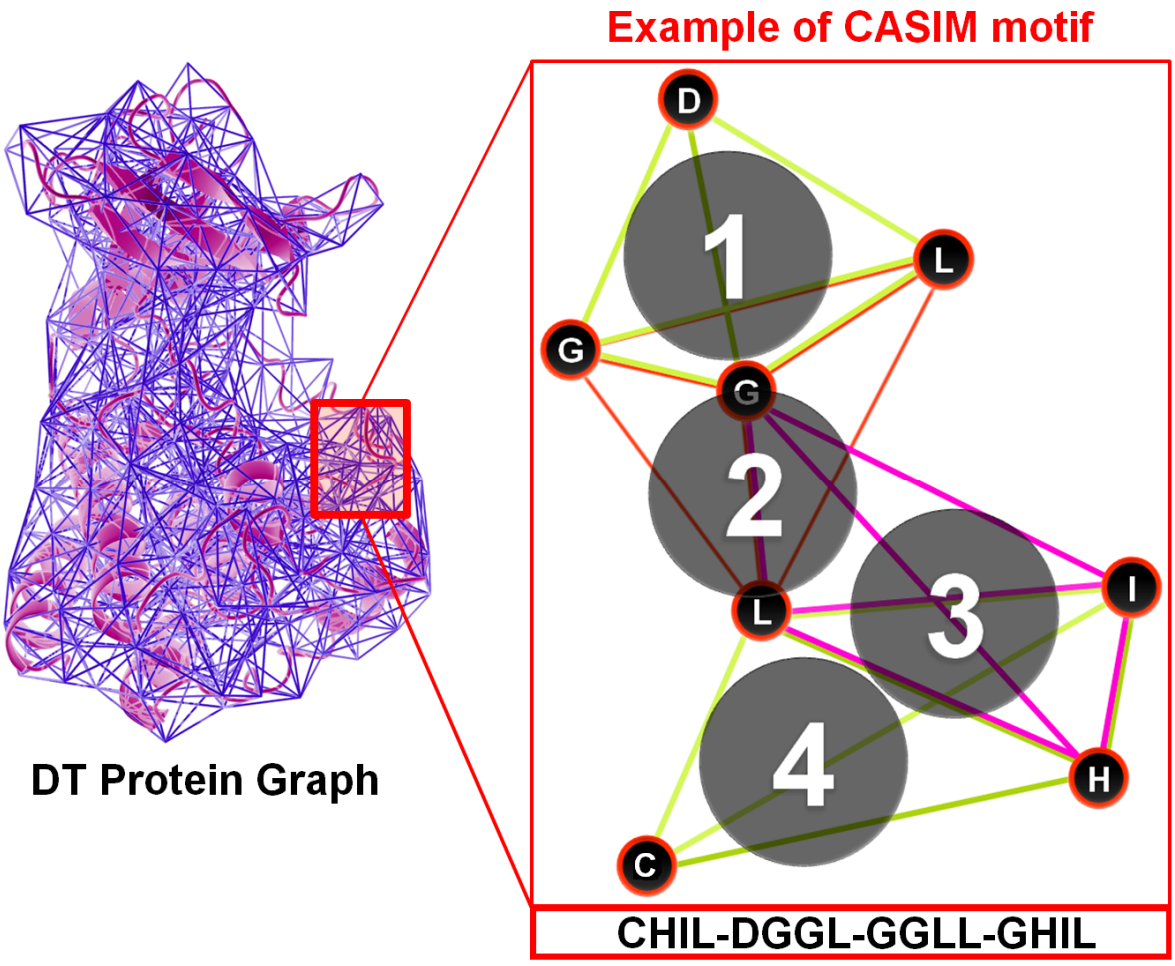


Figure 1.3: CASIM structural packing motifs retrieved from the DT of a protein: example of the motif CHIL-DGGL-GGLL-GHIL involving four neighboring simplicial tetrahedral.

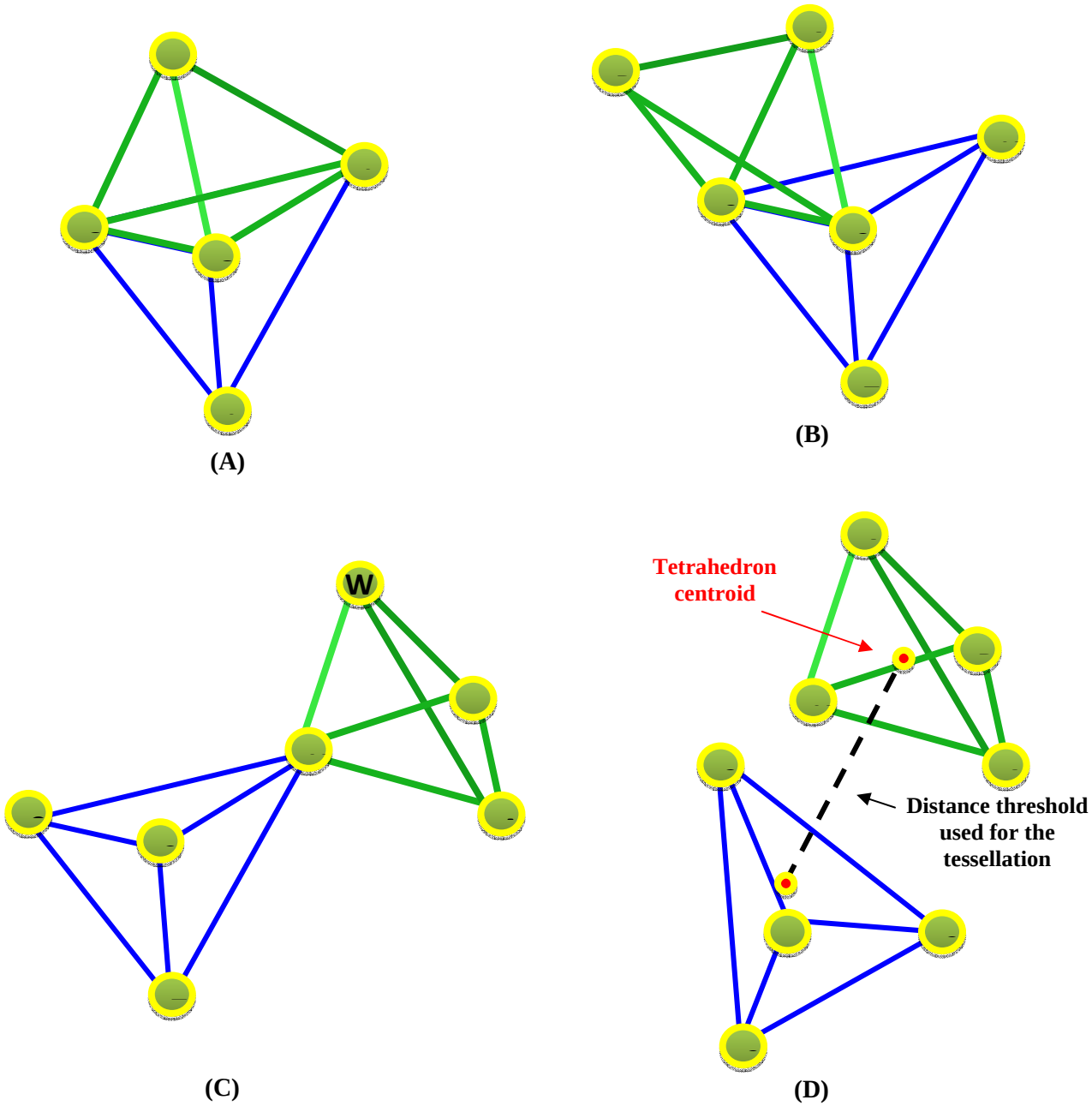


Figure 1.4: The 2nd order tessellation allows CASIM to retrieve complex neighborhood relationships for the 1st order Delaunay tetrahedra. Different types of tetrahedra are retrieved. They can share: (A) a common face; (B) a common edge; (C) a common node); (D) nothing.

CHAPTER 2

PROTEIN FUNCTION PREDICTION AT THE STRUCTURAL LEVEL BASED ON PROTEIN FAMILY-SPECIFIC STRUCTURAL MOTIFS AND CONSERVED KEY RESIDUES

2.1 Introduction

Although, several Automated Function Prediction (AFP) approaches have been reported during the recent decades, there are still several challenging problems remaining. **First,** it is accepted that a three-dimensional (3D) protein structure is better conserved than its sequence[19]. In addition, function inference by *structural motifs* is likely to be more reliable than using global structural similarity because 3D arrangements of functionally important residues (e.g., in the active sites) are significantly more conserved than the entire fold[26]. However, only few structural motif-based methods based on structural data alone have been developed[31-34]. The limitation comes from several problems such as the computational difficulty of local structural alignment and comparison or scanning on the large scale of protein structure database. **Second,** besides the assessment of the overall function of a given protein and/or its active site, predicting functionally important residues is also a critical part of AFP. The knowledge of these key residues can improve the understanding of protein function and thus to facilitate drug discovery. Several researchers have addressed this problem. For example, Youn et al.[45] and Cilia and Passerini[46] applied Support Vector Machine (SVM) to distinguish active site and non-active site residues labeled by several sequence- and structure-based features such as conservation profiles, physical and chemical

properties, amino acid composition and atomic density. From our standpoint, it is also challenging to address this problem based on structural data alone. **Third,** predicting function of proteins of unknown functions especially those with low sequence identity (less than 20%) to proteins of known functions are still the ultimate aim for all AFP researches.

The goal of this study is to investigate those three challenging problems. We focus on a structure-based function inference using both structural motifs and functionally important residues. We also applied this strategy for predicting function of proteins of unknown functions having low sequence identity (less than 20%) to proteins of known functions. We report an application of two sequence-independent structure-based methods, FFSM and a novel CASIM for predicting both family-specific structural motifs and conserved key residues. Currently, function inference by FFSM reported earlier was based on the occurrence of family-specific structural motifs only[10, 33, 34, 38, 39]. We extended the application of FFSM for predicting conserved key residues as well. In order to improve the efficiency and specificity of DT graph mining approach, we have incorporated FFSM with a novel CASIM approach (report herein for the first time). CASIM defines a novel type of structural packing motifs as an ensemble of neighboring Delaunay tetrahedra (where vertices are side chain centroids of amino acid residues). In addition, CASIM has been implemented to provide more comprehensive information for the identified family motifs.

We tested our approaches on the superfamily of Metallo-dependent phosphatases obtained from a manually curated database of the Structural Classification of Proteins (SCOP)[47]. This superfamily also known as metallo-phosphoesterase is related to the Pfam family of Metallophos (PF00149, description: Calcineurin-like phosphoesterase). They are a group of enzymes that catalyze the removal of a phosphate group from their substrates. The

Metallophos family members include both mono- and diphosphoesterases possessing two catalytically essential metal cations (e.g., magnesium, manganese, iron, zinc) in their active sites[48]. These enzymes play a critical role in a number of cellular processes[49-52] especially in the propagation of intracellular signals making them viable drug targets for such diseases as diabetes, cancer, cardiovascular disorders and others as discussed in a recent important review[53].

We found that almost 40% of protein structures in Metallophos superfamily (SCOP 1.7.1 release) were proteins with unconfirmed Metallophos function. In this study, we classified Metallophos structures into 2 categories; (1) a group of proteins having known Metallophos functions, and (2) a group of proteins of unconfirmed Metallophos functions. We have applied both FFSM and CASIM to the group of proteins having known Metallophos functions to identify Metallophos-specific structural motifs. We found that both methods were capable of identifying similar motifs but CASIM was more computationally efficient. We also showed that the predicting motifs were rarely found in proteins outside the family; this observation guaranteed the specificity of the identified Metallophos motifs. We combined the second group of proteins of unconfirmed Metallophos functions in SCOP 1.7.1 with new Metallophos members added in the newer version of SCOP 1.7.3 and 1.7.5. We selected only proteins having sequence identity less than 20%, compared to our training set, into the external set. By combining the data, the external set had both proteins of known Metallophos functions and unconfirmed Metallophos functions. We then determined whether proteins in the external set can be annotated as Metallophos proteins based on the occurrence of the identified Metallophos-specific motifs. We also predicted conserved key residues in those proteins. We validated our predicted results on a group of known Metallophos proteins

having support data from the primary literatures. We compared our predicting performance with several publicly available methods such as a sequence-based search (Pfam)[14], 3D template searches[31, 32, 54] (i.e. enzyme active site template and reverse template searches) and the Catalytic Residue Prediction (CRP)[45]. Furthermore, we predicted function and conserved key residues of proteins of unconfirmed Metallophos function having sequence identities less than 20% (midnight zone) compared to the training set. The studies reported herein showed that our predicted results are in agreement with the published results and are comparable to those from the benchmark methods. This observation illustrates the power of our methodologies for addressing the challenging issues of predicting function and key residues of proteins of unconfirmed function based on structure information alone.

2.2 Methods

2.2.1 Training set of Metallo-dependent Phosphatases

We have compiled a dataset of 84 PDB chains from 9 different families in the Metallo-dependent Phosphatases (Metallophos) superfamily (SCOP ID 65300 from SCOP release 1.7.1). Only 4 families of known Metallophos functions (families of Purple acid phosphatase-like, 5'-nucleotidase (syn. UDP-sugar hydrolase) N-terminal domain, Protein serine/threonine phosphatase and DNA double-strand break repair nuclease) containing 53 entries were used to generate a training set. However, the identification of frequent subgraphs requires the deletion of nearly identical structures to avoid any statistical bias. Thus, the public server PISCES[55] was used for additional training set curation. PISCES provides user with an efficient service for culling sets of protein sequences using different thresholds such as the maximum pair-wise sequence identity (measured using the PSI-BLAST algorithm in

three iterations) or the crystal structure resolution. In this study, we used a 90% sequence identity cutoff, resolution less than 3 Å and R-value better than 0.3. After the curation, the training set consisted of ten PDB chains **(**see **Table 2.1)**: 1s95A (PDB code: 1s95; chain A), 1g5bA, 1s70A, 1kbpA, 1ii7A, 1auiA, 1xzwA, 1uteA and 1qhwA sharing no more than 85% pair-wise sequence identities (see **Figure 2.1**).

Table 2.1: Metallophos training-set containing 10 protein chains

PDB ID	Chain	Protein name
1s95	A	Serine/threonine-protein phosphatase 5
1g5b	A	Serine/threonine-protein phosphatase
1s70	A	Serine/threonine-protein phosphatase PP1-beta catalytic subunit
1kbp	A	Iron(III)-zinc(II) purple acid phosphatase
1ii7	A	DNA double-strand break repair protein mre11
1aui	A	Serine/threonine-protein phosphatase 2B
1hp1	A	5'-nucleotidase
1xzw	A	Sweet potato purple acid phosphatase
1ute	A	Pig purple acid phosphatase
1qhw	A	Purple acid phosphatase from rat bone

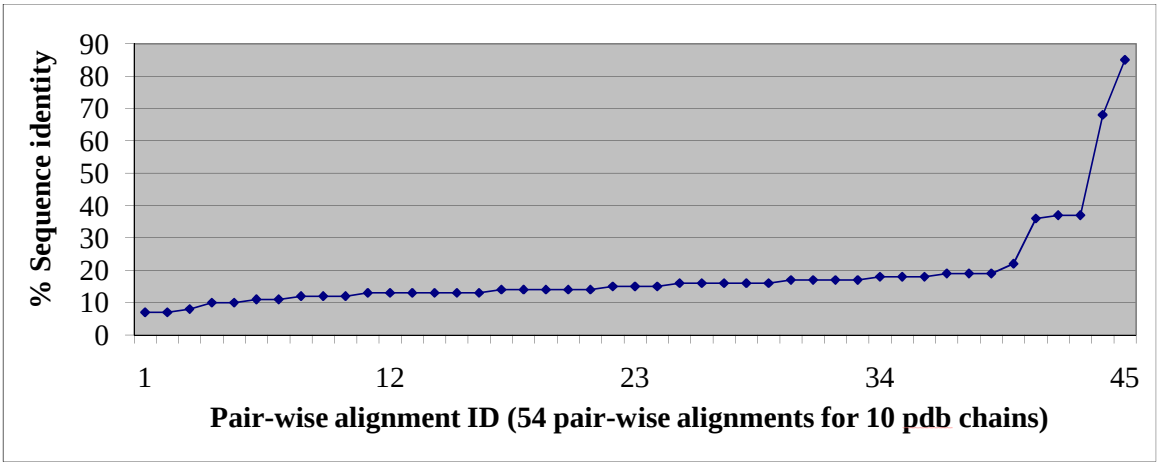


Figure 2.1: Distribution of pair-wise sequence identities in the Metallophos training-set: **(1) sequence length:** average = b= 311 amino acid residues, minimum = 219 amino acid residues, maximum = 378 amino acid residues; **(2) sequence identity:** average = 18.6%, minimum = 7%; maximum = 85%, only 5 pair-wise alignments have sequence identities > 30%.

2.2.2 Selection of the test set from the external dataset containing Metallophos members

We found that the majority of proteins in each of the other five Metallophos families in SCOP 1.7.1 (families of YfcE-like, TT1561-like, Hypothetical protein aq_1666, DR1281-like and Phosphoesterase-related) were proteins of unconfirmed Metallphos function. We included all members in those five families into the external dataset. We combined this dataset with new 74 Metallophos entries added in the new release of SCOP 1.7.3 and SCOP1.7.5. Then, the representative proteins were retrieved using the same PISCES criteria applied to generate the training set (see Section 2.2.1). In order to illustrate the performance of our approach on remote homology detection, we selected only proteins having sequence identity less than 20% when compared to the training-set members. We retrieved 12 proteins into the test set. Three of them have known Metallophos functions according to the literature information (PDB chains: 3d03A (PDB code: 3d03; chain A)[56], 1s3lA[57] and 1t70A[58]). One had a function suggested from structures (2nxfA)[59] whereas the rest were proteins of

unconfirmed Metallophos function (3ck2A, 1su1A[60], 1xm7A, 1t71A, 2cv9A, 2yvtA, 1nnwA and 1uf3A).

2.2.3 Background dataset

In our subgraph mining-based approaches, frequent subgraphs retrieved from the training set of proteins of interest become common subgraphs if and only if they are rarely found in the proteins of a "*background*" database. In this study, the same PISCES criteria used to curate the training set (see Section 2.2.1) were applied to the PDB (May 2007 release) to build a background dataset. This dataset included 6,605 non-redundant protein chains excluding the 84 Metallophos proteins in SCOP 1.7.1 (see Section 2.2.1).

2.2.4 Identification of Metallophos-specific structural motifs using FFSM

The FFSM approach[33, 34] (see Section 1.3-1.5) was applied to mine Metallophos-specific structural motifs (non-redundant frequent common subgraphs) from a training set of 10 Metallophos proteins. Each protein structure in the training set was modeled as AD (Almost Delaunay) graph consisting of nodes and edges. In this study, motifs were restricted to fully interconnected subgraphs in which all nodes connect to each other. Other parameters were set for mining motifs from the graph representations of protein structures in the training set as follows:

- **Nodes** represent alpha carbons (Cα) of amino-acid residues. There are 20 possible types of nodes based on the 20 natural types of amino acid residues.
- **Edges** which connects two adjacent nodes was determined according to the AD technique with ε=0.1. The edges were classified into 10 types; 5 types of AD

edges (edge length for type1 to type5 are 0-4, 4-6, 6-8.5, 8.5-10 and 10.5-12.5 Å, respectively) and 5 types of distance constraints between non-contacting residues (edge length for type6 to type10 are 0-4, 4-6, 6-8.5, 8.5-10 and 10.5-12.5 Å, respectively).

- **Minimum size of the motif** was set to 4 amino acid residues
- **Minimum support (f)** of that subgraph is the minimum fraction of family members in the training set that must contain that subgraph
- **Maximum background occurrence (b)** is the maximum fraction of proteins in the background dataset that contain a subgraph of interest. The value of b was set to 0.1% by default.

A subgraph is considered frequent if its 'minimum support' (f) value is higher than a user-defined threshold (e.g., f=0.9; the motif presents in at least 90% of the family members). However, those frequent subgraphs become frequent common subgraphs (motifs) if and only if they are rarely (below certain frequency threshold) found in proteins of a 'background' dataset (b= 0.1%: found in no more than seven proteins out of 6,605 non-redundant protein chains in the background dataset).

However, the main concerns that need to be underlined are as follows: (1) FFSM used in this study recognized only a structural packing motif of fully interconnected subgraph, and (2) the method reported only maximal subgraphs (graphs that are not part of any larger frequent subgraphs). Although, this motif definition facilitates the computational task and assures the motif specificity, it increases the possibility of missing motifs that are not fully interconnected subgraphs or/and are only substructures of a large maximal subgraph.

2.2.5 Identification of Metallophos-specific structural motifs using CASIM

CASIM (see paragraph 1.6) was applied to mine Metallophos-specific structural motifs (frequent common Delaunay tetrahedral) from a training set of 10 Metallophos proteins. Each protein structure was modeled as DT (Delaunay tessallation) graph consisting of nodes and edges. Nodes represent side chain centroids of amino acid residues. There are 20 possible types of nodes based on the 20 natural types of amino acid residues. Unlike FFSM that defines a motif as a fully interconnected subgraph, CASIM describes motifs as ensembles of neighboring Delaunay tetrahedral. Thus, we expected to recover motifs missed by FFSM or vice versa.

To reach the goal of efficient and fast function annotation, CASIM was applied to identify Metallophos-specific motifs as follows (see **Figure 2.2**): 1**)** each family member in the training set is tessellated to obtain a list of its constitutive CASIM structural motifs, 2) The lists of motifs for all family members are processed to build pattern matrices where each row corresponds to a protein, and each column corresponds to the type of the motif. These matrices contain the occurrences of each motif's type in every protein of the training set, using a specific sparse matrix implementation (only non-zero values are stored for efficiency). All motifs included in a given pattern matrix involve the same number of constituent Delaunay tetrahedra.

The identification of motifs is fast and optimized. The parameters of minimum support (f) and maximum background occurrence (b) described under Section 2.4 were also adopted to retrieve motifs. Only motifs occurring with at least a given user-defined support (f) value (e.g., f=0.9; the motif presents in at least 90% of the family members) are retrieved. As discussed above, although certain motifs may be well conserved in a family, it does not

imply that they are specific to this family. Each family-specific motif is required to occur with high support (i.e., in significant number of protein members of a family) but low background (i.e., in a very small number of all other proteins). Therefore, the algorithm applies a 'background' frequency filter to obtain the list of conserved motifs. Frequent motifs are retained if and only if they are rarely found in the 'background' dataset. The maximum background occurrence (b) was set to 0.1% by default (less than 7 proteins with our current database). The background checker implemented in the CASIM software (running on a standard Dual-Core PC) requires less than a second to retrieve all necessary information concerning the motifs in the background set; this high computational efficiency is achieved because all possible motifs present in the background proteins have been pre-calculated and stored in a database.

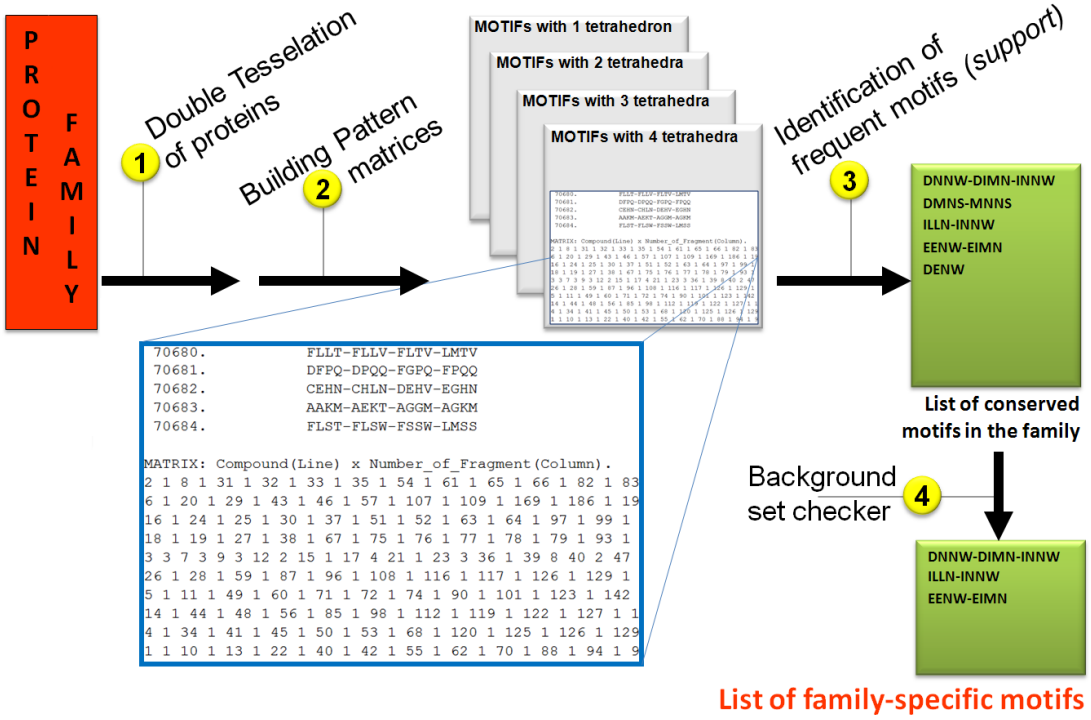


Figure 2.2: The overall workflow to identify family-specific structural motifs using the CASIM approach.

2.2.6 Cumulative Support Profiles for protein function inference

Since each amino acid residue in a protein is surrounded by other residues, there is an interesting question of characterizing the environment of each residue and investigating the structural similarity between the neighborhoods of each residue (especially the functionally significant ones) for a given target protein vs. a set of proteins (such as a family of proteins with the same function).

One simple yet powerful approach to comparing residue environments between protein structures is the use of so called 3D-1D profiles. Originally proposed by Eisenberg[61], this approach translates various parameters of a residue's environment in 3D to a sequence-specific profile where each residue in the sequence is given some sort of score reflecting its 3D environment. 3D-1D profiles have been used in fold recognition[61] or protein model quality assessment[62]. In our previous studies, we employed similar concept to compare proteins using profiles based on four-body statistical potentials generated with the help of Delaunay tessellation[63].

Here, we suggest a novel approach, called **'Cumulative Support Profiles (CSP profile)'**. In order to generate the CSP profile for every Metallophos protein, all CASIM motifs involving from one to four neighboring Delaunay tetrahedra were calculated for the entire training set. Then, the support value (defined here as the number of family members possessing the given motif, i.e., ranging from 1 to 10 in this case) of each quadruplet occurring in a given protein was calculated. In addition, support values of all possible motifs involving two, three and four neighboring tetrahedra were calculated. The cumulative support of a given quadruplet is equal to the sum of the support values of all motifs involving this specific quadruplet (note that for CSP profile, support values are expressed not as

frequencies but as numbers of occurrences). For example, if the quadruplet DGGH has a support value equal to 6 (i.e., it occurs in 6 out of 10 proteins in the training set) and the motif DGGH-GGHN has a support of 2, the partial cumulative support of DGGH is 6+2 = 8. This procedure is repeated for all motifs involving the quadruplet DGGH to calculate a total value of the cumulative support for this quadruplet.

This score can be calculated for any Delaunay quadruplet in any protein of the training set. If a quadruplet occurs frequently in a family, and so are its neighbors, its cumulative support is expected to be high. Thus, the cumulative support provides a quantitative assessment of the conservation of each Delaunay quadruplet of residues within a protein family. Similar consideration could then be applied to each amino acid residue to calculate its individual cumulative support: the latter is equal to the sum of the cumulative supports of all quadruplets involving this particular residue. Finally, the cumulative support values for each residue can be plotted against the residue number in the sequence to obtain the protein cumulative support profile (see **Figure 2.7**) where the peaks correspond to residues with the highest conserved 3D environment in the protein family.

2.3 Results

2.3.1 Metallophos-specific structural motifs identified by FFSM and CASIM

Both FFSM and CASIM approaches were independently utilized for identifying structural motifs conserved in the training set of Metallophos members but found in no more than 0.1% (b=0.1) in the background dataset of 6605 protein chains.

Small sets of motifs have been identified by FFSM and CASIM using multiple minimum support (f) values (see **Table 2.2**): (1) FFSM retrieved 31 to 12 motifs when f

value was increased from 0.8 to 1.0, respectively; (2) CASIM retrieved 13 motifs at f = 1.0. Both FFSM and CASIM detected the same set of eight residues in the training-set members.

Table 2.2: Number of Metallophos-specific structural motifs retrieved from the family training set (column 2) at given support (f) values

Methods	Metallophos structural motifs
FFSM	
f=0.8	31 motifs (8 residues)
f=0.9	27 motifs (8 residues)
f=1.0	12 motifs (8 residues)
CASIM	
f=1.0	13 motifs (8 residues)

The occurrence distribution of those motifs in the training set members and the background database (see **Figure 2.3**) revealed their good specificity to discriminate the Metallophos family. For instance, by using FFSM (**Figure 2.3**, top) at f = 0.8, there is a maximum of two motifs (among the 31 selected ones) which are present in some proteins of the background database (49 proteins possessed one motif and seven proteins possessed two motifs at most) whereas Metallophos training set members possess at least 18 motifs out of 31 (six Metallophos proteins contained all 31 motifs). Thus, there is a significant difference between the minimum number (18) of structural motifs found in any training set protein and the maximum number (2) of motifs found in any protein of the background dataset. The results obtained from CASIM are in the same direction (**Figure 2.3**, bottom). This observation clearly indicates the great specificity of FFSM and CASIM motifs for the Metallophos family. It should also be mentioned that the execution times required by both programs were formally similar (less than five minutes). However, FFSM runs on a cluster

server via client scripts under Linux. On the contrary, CASIM is executed locally (all calculations presented in this study were performed on a Dual-Core PC).

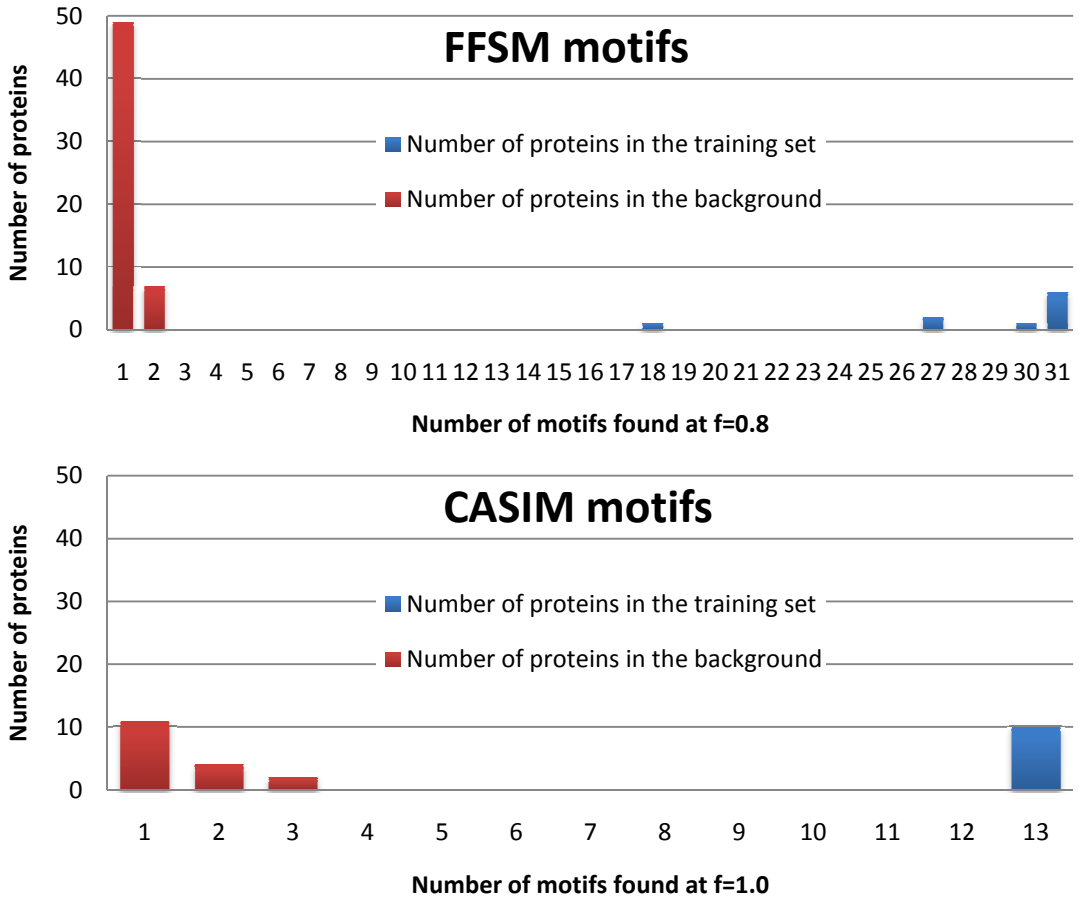


Figure 2.3: Distribution of Metallophos-specific structural motifs retrieved by FFSM (top) and CASIM (bottom) in the training set (blue) and the background dataset (red).

The examples of two Metallophos-specific motifs detected by CASIM are given in the **Table 2.3**, specifically their annotations: **DDHH-DGGH-DGHN-DHHN** of connectivity type *1122* (called motif **A**) and **DDGH-DDHH-DGGH-DGHN** connectivity type *1222* (called motif **B; cf. Figure 2.4**). Briefly, these two CASIM motifs A and B included seven residues involved in four neighboring Delaunay tetrahedra: for instance, motif A involves the tetrahedra DDHH, DGGH, DGHN and DHHN. Metallophos protein residues which are included in these two motifs are shown in **Table 2.3**. A rapid analysis suggests that motifs A

and B involve exactly the same residues but importantly, their types are different (1122 for motif A, 1222 for motif B): the tetrahedral connectivity between these residues is different because of the types of graph edges (i.e., an edge represents either a peptidic bond or geometrical proximity in 3D space) between the vertex-residues of the tetrahedra (see **Figure 2.4**). The nomenclature 1122 reflects that two Delaunay tetrahedra out of four included in motif A are of type 1 and the other two are of type 2. Here the tetrahedron DDHH type 1 (type 1 means that all four vertex-residues are not consecutive in the protein sequence) in the motif A not present in the motif B, whereas the tetrahedron DDHH type 2 (type 2 means that two out of four vertex-residues are consecutive in the protein sequence) is present in motif B.

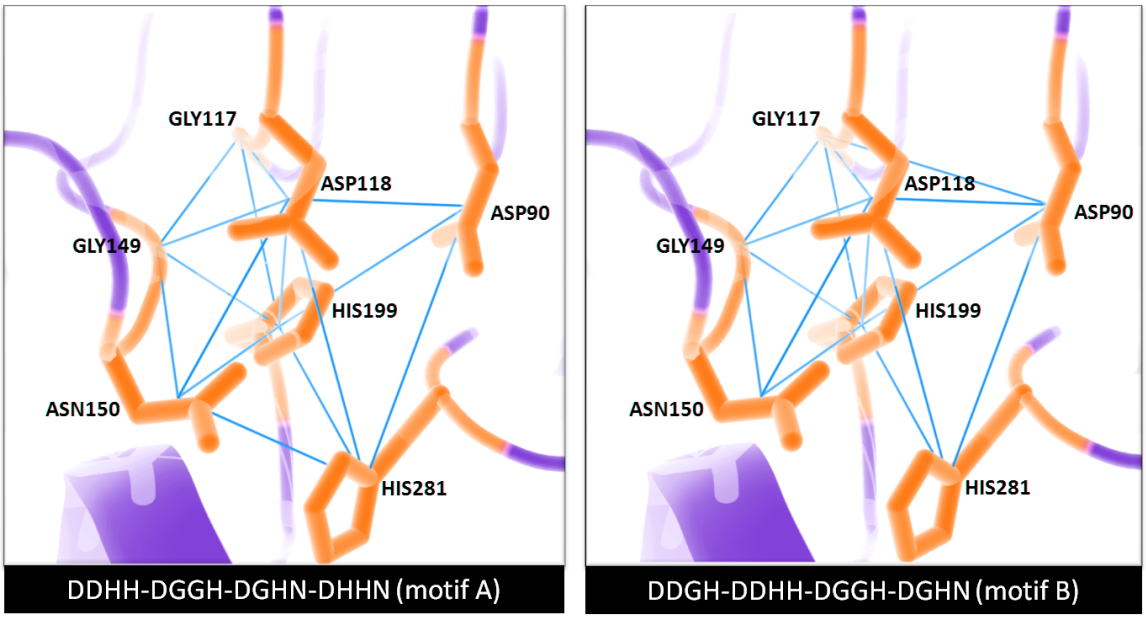


Figure 2.4: Metallophos-specific motifs (DDHH-DGGH-DGHN-DHHN and DDGH-DDHH-DGGH-DGHN) retrieved by CASIM plotted on serine/threonine-protein phosphatase 2B (PDB chain: 1au1A, a training-set member): both motifs involve the same residues but different constitutive Delaunay neighbor tetrahedra, improving their specificity to recognize Metallophos activity.

Table 2.3: Metallophos-specific structural motifs retrieved by CASIM in the ten training-set members and also in hypothetical protein YfcE (PDB chain: 1su1A). ESA = Exposed Surface Area ($Å^2$), ESA1; Volume in $Å^3$.

MOTIF DDHH-DGGH-DGHN-DHHN									
Protein	**Volume**	**ESA**	**TYPE 1122**						
1auiA	37.3	131.4	ASP77	GLY104	ASP105	GLY136	ASN137	HIS186	HIS268
1g5bA	41.4	139.4	ASP20	GLY48	ASP49	GLY74	ASN75	HIS139	HIS186
1hp1A	45.0	148.5	ASP16	GLY58	ASP59	GLY90	ASN91	HIS192	HIS227
1ii7A	45.1	149.8	ASP8	GLY48	ASP49	GLY83	ASN84	HIS173	HIS206
1kbpA	60.5	192.2	ASP15	GLY43	ASP44	GLY80	ASN81	HIS166	HIS203
1qhwA	38.0	131.0	ASP10	GLY47	ASP48	GLY86	ASN87	HIS182	HIS217
1s70A	54.3	177.3	ASP64	GLY91	ASP92	GLY123	ASN124	HIS173	HIS248
1s95A	37.4	129.8	ASP67	GLY95	ASP96	GLY127	ASN128	HIS177	HIS252
1uteA	65.7	201.5	ASP12	GLY49	ASP50	GLY88	ASN89	HIS184	HIS219
1xzwA	58.4	188.1	ASP16	GLY44	ASP45	GLY81	ASN82	HIS167	HIS204
1su1A	**33.8**	**122.0**	**ASP9**	**GLY36**	**ASP37**	**GLY72**	**ASN73**	**HIS105**	**HIS127**
MOTIF DDGH-DDHH-DGGH-DGHN									
Protein	**Volume**	**ESA**	**TYPE 1222**						
1auiA	37.3	131.4	ASP77	GLY104	ASP105	GLY136	ASN137	HIS186	HIS268
1g5bA	41.4	139.4	ASP20	GLY48	ASP49	GLY74	ASN75	HIS139	HIS186
1hp1A	60.9	181.0	ASP16	GLY58	ASP59	GLY90	ASN91	HIS192	HIS227
1ii7A	45.1	149.8	ASP8	GLY48	ASP49	GLY83	ASN84	HIS173	HIS206
1kbpA	38.2	131.4	ASP15	GLY43	ASP44	GLY80	ASN81	HIS166	HIS203
1qhwA	55.1	170.5	ASP10	GLY47	ASP48	GLY86	ASN87	HIS182	HIS217
1s70A	56.4	184.3	ASP64	GLY91	ASP92	GLY123	ASN124	HIS173	HIS248
1s95A	52.8	167.3	ASP67	GLY95	ASP96	GLY127	ASN128	HIS177	HIS252
1uteA	65.7	201.5	ASP12	GLY49	ASP50	GLY88	ASN89	HIS184	HIS219
1xzwA	50.9	161.9	ASP16	GLY44	ASP45	GLY81	ASN82	HIS167	HIS204
1su1A	**58.2**	**176.9**	**ASP9**	**GLY36**	**ASP37**	**GLY72**	**ASN73**	**HIS105**	**HIS127**

Furthermore, volumes as well as exposed surface areas (ESA) indicate a great homogeneity of the motifs found in the ten training-set members (see **Table 2.3**). For example, volumes of DDHH-DGGH-DGHN-DHHN varied from 37.3 to 65.7 $Å^3$ whereas their ESA values were ranging from 129.8 to 201.5 $Å^2$. A deeper analysis shows that these variations are due to several conformational shifts of residue side chains like HIS221 in the 1ute protein, HIS323 in 1kbp, etc. Meanwhile, Multiple Structural Alignments of CASIM

motifs (performed by the TMAlign program[64] executed via PROTMAN interface) for the training-set members revealed a very good local alignment of the seven residues involved in the motifs. In **Figure 2.5**, the motif DDHH-DGGH-DGHN-DHHN is visualized in the PyMol program under the control of PROTMAN via python scripts. For each residue within the motif, its representative vertices (corresponding to side chain centroïds) were fairly well superimposed. The RMSD values for each residue were in the range 0.41-0.73Å whereas the overall RMSD value was equal to 0.59Å for the whole DDHH-DGGH-DGHN-DHHN motif.

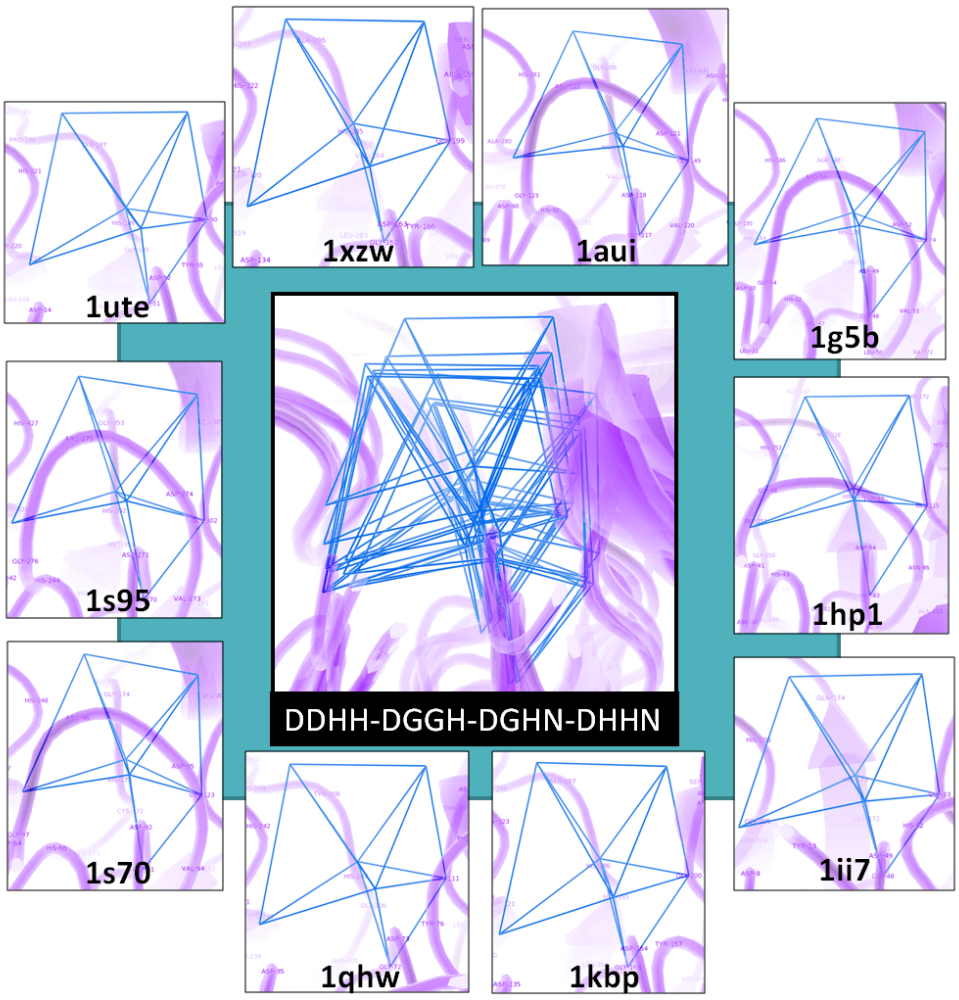


Figure 2.5: Visualization of the Metallophos-specific motif (DDHH-DGGH-DGHN-DHHN) retrieved by CASIM for all training set members.

2.3.2 Validation on test proteins of known function

The three test protein structures (PDB chains: 3d03A, 1z2wA and 1t70A) had known Metallophos functions according to the published results from experimental analyses (see **Supplementary data**). In addition, the catalytic site residues were also suggested for the test proteins by the authors, mostly from structure analysis and few from mutation analysis. These test proteins have only 9-19% sequence identity to the training set.

In this study, function inferences by CASIM and FFSM were relied on the appearance of Metallophos-specific motifs and the conserved key residues involved with those motifs (see **Table 2.4** and **Table 2.5**). As reported in Section 3.1, the family motifs and conserved key residues detected by CASIM (f=1.0) and FFSM (f=0.8) were 13 motifs with 8 residues and 31 motifs with 8 residues, respectively. CASIM was capable of detecting all family motifs in 3d03A whereas FFSM detected the majority of the motifs (25 from 31 motifs and 7 from 8 residues) in this protein. Compared to FFSM, CASIM captured the larger fraction of family motifs in 1z2wA. However, CASIM could not identify any family motifs in 1t70A, which was in turn identified by FFSM. Thus, by combining CASIM and FFSM (see **Table 2.4**, column 6, CASIM-FFSM), we could retrieve the family motifs in all test proteins. This observation suggested the benefit of combining two methods to recover more motifs from graph space.

CASIM-FFSM was able to detect the family motifs of 8 key residues corresponding to those found in the training set (see paragraph 3.1) in 3d03A and 1z2wA. The method only retrieved 1 motif in 1t70A. However, this single motif was highly specific to the family since it was not present in any proteins in the background dataset.

We evaluated the prediction performance of CASIM-FFSM with the published results and those from publicly available methods (see **Table 2.4** and **Table 2.5**); Pfam, reverse template and enzyme active site template searches, and Catalytic Residue Prediction (CRP). The common highlights between reverse template search and CASIM-FFSM are that they aim to predict protein function based on structure data alone and do not require any prior knowledge of functionally important residues. The reverse template was generated by breaking the query protein itself into many three-residue templates of neighboring residues. Then, these small templates were scanned against a representative set of structures in PDB. The enzyme active site template search and CRP assigned active site residues based on the data obtained from the Catalytic Site Atlas (CSA). Enzyme active site templates were manually derived templates of three to six residues. Each template consisted of one, two or three residues known to be catalytic, and one or more additional conserved residues relative to the catalytic residues. CRP was mainly developed for predicting catalytic residues instead of predicting protein function. The method predicted key residues from sequence and structure feathers using SVM.

Table 2.4: Function prediction on test proteins of known function using AFP methods. *For reverse template and enzyme active site template searches, we reported only the first hit of known function having the highest scoring template and are not found in our external set; resi = number of amino acid residues.*

PDB chain	Protein name	% Seq iden to the training set	CASIM (f=1)	FFSM (f=0.8)	CASIM-FFSM	Pfam	Reverse template	Enzyme active site template
3d03A	Glycerophosphodiesterase	10-16	13 motifs (8 resi)	25 motifs (7 resi)	38 motifs (8 resi)	Metallophos (2.3e-13)	1qhw (Metallophos)	4kbp (Metallophos)
1z2wA	Vsp29	11-19	5 motifs (8 resi)	1 motif (4 resi)	6 motifs (8 resi)	Metallophos (0.00047)	3dsd (Metallophos)	4kbp (Metallophos)
1t70A	DR1281	9-17	0	1 motif (4 resi)	1 motif (4 resi)	Metallophos (4.4e-6)	3jyf (Metallophos)	2dnj (Metallophos)

We found that CASIM-FFSM was able to detect the majority of published key residues in 3d03A and 1z2wA, and two published key residues in 1t70A. It is important to underline that other residues detected by us that were not found in the literatures were all neighbors of the published key residues, and were found at the metal-binding sites.

Pfam can also infer Metallophos function to all test proteins with high confidence (E-value less than 0.001). This function inference by Pfam was based on the presence of the HMM profile of Metallophos family (PF00149) obtained from the publicly available Pfam database.

The reverse template and enzyme active site template searches were applied for both function inference and catalytic residue prediction. In our study, we reported only a hit (matched structure) having the highest score template and were not members of our external dataset. We found that reverse template search and enzyme active site template search were able to provide hits for every test proteins. All hits given by both methods were Metallophos proteins. In addition, most predicted catalytic residues retrieved by both methods were similar to those reported in the literatures and those identified by FFSM-CASIM.

CRP was used for catalytic residue prediction with SVM-score threshold 2.5, a value reported by the authors to achieve almost 50% precision[45]. We found that the majority of the predicted residues were similar to those in the published and CASIM-FFSM results.

Table 2.5: Conserved key residues in test proteins of known PTK function: comparison of the key residues reported in the primary literatures and those from automated prediction methods. The predicted residues matching to the published residues are labeled in red.

PDB ID	Published results	Predicted residues			
		CASIM-FFSM	Reverse template search	Enzyme active site template search	CRP
3d03A	Asp8, His10, Asp50, Asn80, His156, His 195, His197	Asp8, Gly49, Asp50, Gly79, Asn80, His81, His156, His 195	Asp8, His156, Cys193.	Asp8, Asp50, Asn80, His81, His156, His195	Asp8, His10, His50, Asn80, His81, His156, His195, His197
1z2wA	Asp8, Asn39, Asp62, His86, His117	Asp8, Gly38, Asn39, Gly61, Asp62, His86, His115, His117	Asp8, His86, Gly114	Asp8, His10, Asn39, Asp62, His86, His115	Asp8, His10, His86, His115, His117
1t70A	Asp8, Glu37, Asn38, Asn65, His148, His173, His175	Gly64, Asn65, His66, His173	Asp8, Asn35, His148	Glu37, His148, Asp193, His175	Asp8, Glu37, Asn65, His66, His173

2.3.3 Predicting Metallophos function and conserved key residues in proteins of unconfirmed function

We predicted the Metallophos function and conserved key residues in 9 test protein structures; one putative (2nxfA), one uncharacterized (3ck2A) and seven hypothetical proteins (1su1A, 1xm7A, 1t71A, 2cv9A, 2yvtA, 1nnwA and 1uf3A). All of them fell into the midnight zone (less than 20% sequence identity) when compared to the training set. CASIM-FFSM was able to detect Metallophos motifs in the five following proteins (see **Table 2.6**).

Putative dimetal phosphatase LOC393393 (PDB code 2nxf) from *Danio rerio*

Protein LOC393393 shares 12-17% sequence identity to our training set. We inferred that this protein has a Metallophos function because it contains Metallophos motifs of 8 key residues. Our inference was corroborated by the following: (1) the authors of the structure suggested Metallophos function for this protein based on its similar topology to other Metallophos proteins[59], (2) six key residues detected by us were similar to those suggested by the authors (Asp13, Asp60, Asn96, His97, His228 and His265), (3) the two residues (Gly59

and Gly95) identified by us that were not mentioned in the literature were also found at the active site, and (4) Pfam found the Metallophos profile in this protein with high confidence (E-value 6.4e-06).

Conserved uncharacterized protein (predicted phosphoesterase COG0622) from *Streptococcus pneumoniae* TIGR4 (PDB code 3ck2)

The crystal structure of this conserved uncharacterized protein was released to the PDB by the Midwest Center for Structural Genomics (MCSG) in 2008. This protein shares 11-17% sequence identity to our training set. Pfam detected the Metallophos profile in this protein with low confidence (E-value 0.21). However, we were convinced that the protein has Metallophos function from the CASIM-FFSM results. We identified the majority of the family motifs containing 7 conserved key residues in this protein. In addition, the key residues (Asp11, Gly37, Asp38, Gly56, Asn57, His81 and His110) detected by us were present at the Mn^{2+}-binding sites. Three of them (Asp11, Asp38 and His110) were similar to those predicted by CRP, which identified 5 residues (Asp11 (SVM score of 3.72), His13 (3.41), Asp38 (3.59), His110 (3.82), His112 (2.90)).

Hypothetical protein aq_1665 (PDB code 1xm7) from *Aquifex aeolicus*

The crystal structure of hypothetical protein aq_1665 was deposited to the PDB by the MCSG in 2004 as a structural genomic target of unknown function. This protein shares low sequence identity (9-17%) to the training set. We inferred that this protein has Metallophos function based on the occurrence of 17 Metallophos motifs. Our function prediction was in agreement with that from Pfam, which found Metallophos profile in this protein with high confidence (E-value 6.2e-13). We also identified 8 conserved key residues

(Asp7, Gly49, Asp50, Gly77, Asn78, His79, His111 and His145) in this protein. Some of them were reported by other methods; (1) CRP identified 4 residues (His145, Asp50, His111 and Asp7)[45], and (2) the method predicting transition metal-biding in apo proteins by Babor's group identified 4 residues (Asp7, His9, Asp50 and His111)[65].

Hypothetical protein MPN349 (PDB code 1t71) from *Mycoplasma pneumoniae*

The crystal structure of hypothetical protein MPN349 was released by the Berkeley Structural Genomics Center (BSGC) to the PDB in 2004. This protein shares low sequence identity (7-18%) to the training set. Pfam was unable to provide any hit for this protein. However, we inferred Metallophos function to this protein due to the presence of 2 specific Metallophos motifs, which were not found in the background dataset. We also found that this protein share high similarity (38% sequence identity and DALI z-score 37.1) to protein DR1281 (PDB code 1t70), one of known Metallophos proteins in our test set. Our method predicted 5 conserved key residues (Gly70, Asn71, His72, His158 and His183). Three of them (Asn71, His72 and His183) were overlapped with those predicted by CRP, which identified 4 residues (Asp12 (SVM score of 3.16), Asn71 (2.68), His72 (2.96) and His 183 (2.55)).

Hypothetical protein YfcE (PDB code 1su1) from *E. coli*

The crystal structure of hypothetical protein YfcE was deposited to the Protein Data Bank (PDB) by the Midwest Center for Structural Genomics (MCSG) as a structural genomic target of unknown function in 2004. The protein shares 13-19% sequence identity to the training set. The presences of 18 Metallophos motifs involved with 7 conserved residues

highly suggested that the protein has Metallophos function. An example of 'DDHH-DGGH-DGHN-DHHN' motif detected by CASIM is visualized in **Figure 2.6**. One can see the presence of metal counter-ions and a phosphate group inside the motifs or in their close proximity. The prediction was supported by the following data: (1) structural and biochemical analysis by the authors revealed that the protein had the Mn^{2+} dependent phosphatase activity[60], (2) the authors suggested two metal binding sites: one included ASP9, HIS11, HIS129 and ASP37 whereas the other one consisted of ASP37, ASP73, HIS105 and HIS127. We identified 7 conserved key residues; five (Asp9, Asp37, Asn73, His105 and His127) were similar to those suggested by the authors whereas two were neighbors (Gly36, Gly72) to the reported residues. In this case, we did not employ Pfam results as a support data because Pfam incorporated protein YfcE into a seed used to generate the Metallophos profile.

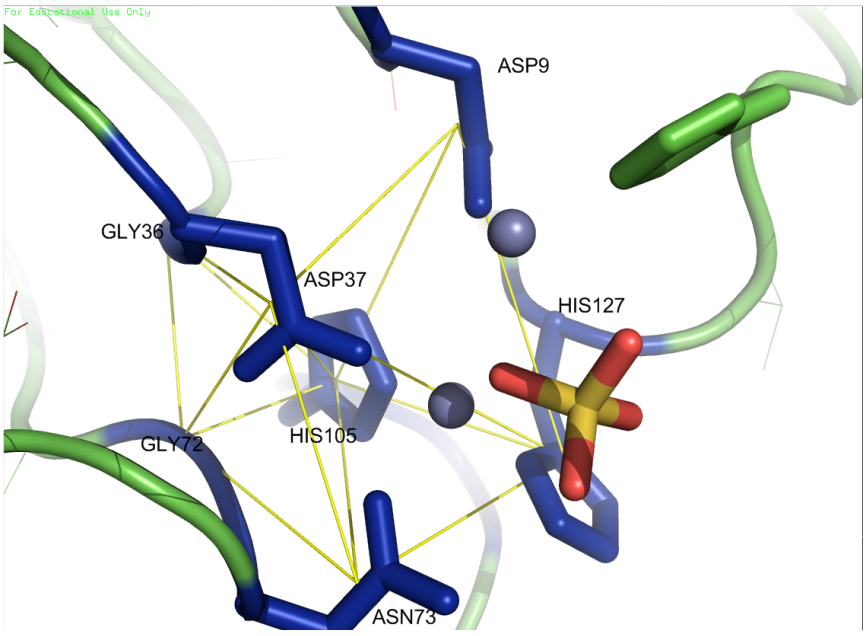


Figure 2.6: Visualization of the Metallophos-specific motif (DDHH-DGGH-DGHN-DHHN) retrieved by CASIM for protein YfcE (1su1A (1su1 chain A)).

Table 2.6: Function prediction on test proteins of unconfirmed Metallophos functions using CASIM-FFSM. *Metallophos-specific structural motifs detected in the training set by CASIM-FFSM (FFSM at f=0.8 and CASIM at f=1.0) consisted of 44 motifs involved with 8 conserved key residues (see **Table 2.2**).*

PDB chain	Protein name	% Seq iden to the training set	#Metallophos motifs detected by CASIM-FFSM	# Key residues detected by CASIM-FFSM
2nxfA	Putative dimetal phosphatase LOC393393	12-17	34 motifs	8 residues
3ck2A	Conserved uncharacterized protein (predicted phosphoesterase COG0622)	11-17	17 motifs	7 residues
1xm7A	Hypothetical protein aq_1665	9-17	17 motifs	8 residues
1t71A	Hypothetical protein MPN349	7-18	2 motifs	5 residues
1su1A	Hypothetical protein YfcE	13-19	18 motifs	7 residues

2.3.4 Cumulative Support Profiles: test case to YfcE

Cumulative Support Profiles (CSP profiles) were obtained for each Metallophos protein of the training set. For illustration, the profile of 1auiA (Serine/threonine-protein phosphatase 2B) is shown in **Figure 2.6**. One can see significant peaks for 7 to 10 residues implying that both these residues and their environments are highly conserved within the family. The vast majority of these residues are situated in the metal-binding site area.

We have further investigated the CSP profile of the YfcE protein as a case study (see **Figure 2.7**). Eleven significant peaks could be identified in this profile: among them, seven corresponded to the seven residues (ASP9, GLY36, ASP37, GLY72, ASN73, HIS105 and HIS127) that have been retrieved by CASIM in the two family-specific motifs DDHH-DGGH-DGHN-DHHN and DDGH-DDHH-DGGH-DGHN and also correspond to the metal-binding site of YfcE. These residues detected by the CSP profile are in perfect agreement with those identified in both FFSM and CASIM motifs. The remaining peaks may imply

residues that are critical for maintaining the overall 3D structure of the protein and thus relatively well conserved within the family.

To validate the method using the CSP profile, ten proteins were chosen randomly from the background set. Then, the CSP profile was generated for 1su1A (hypothetical protein YfcE) based on the Metallphos motifs detected in those 10 proteins. The results showed that no peaks were retrieved, as expected (see **Figure 2.8**).

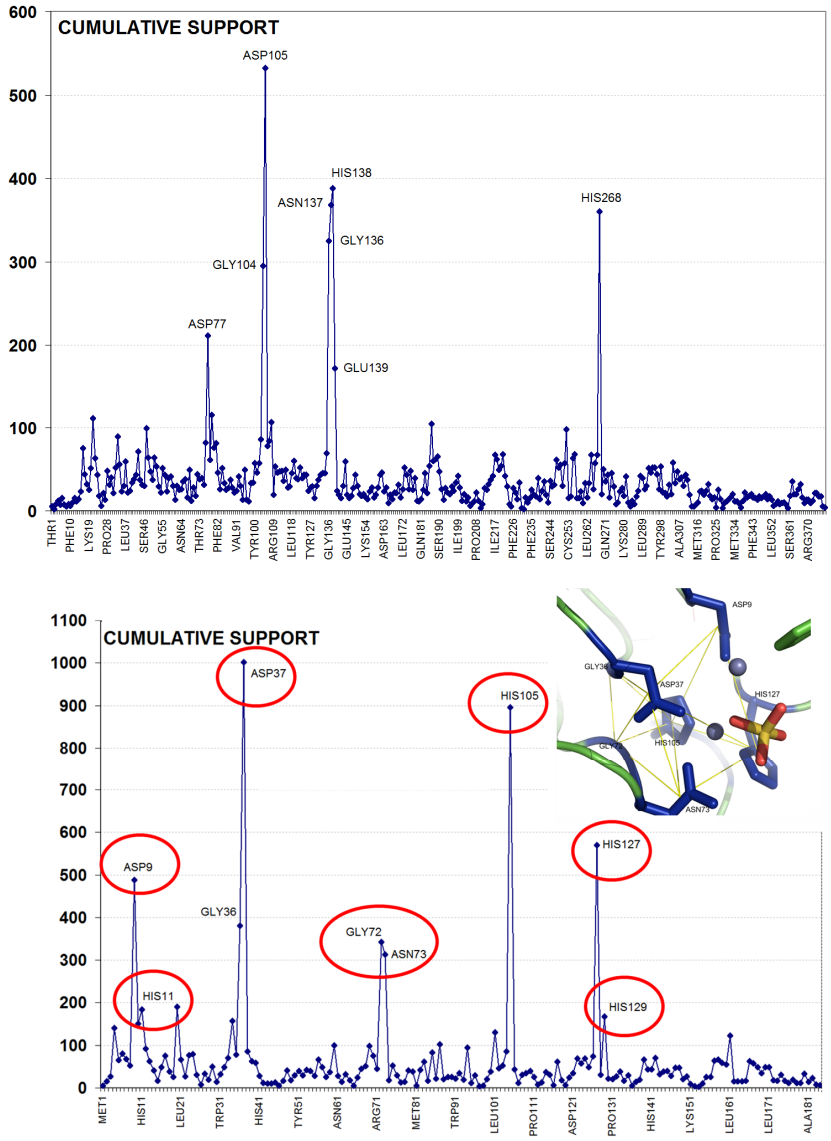


Figure 2.7: The CSP profiles of the serine/threonine-protein phosphatase 2B (a training-set member 1auiA, top) and hypothetical protein YfcE (PDB chain: 1su1A, bottom). X-axis = protein sequence, Y-axis = cumulative support scores.

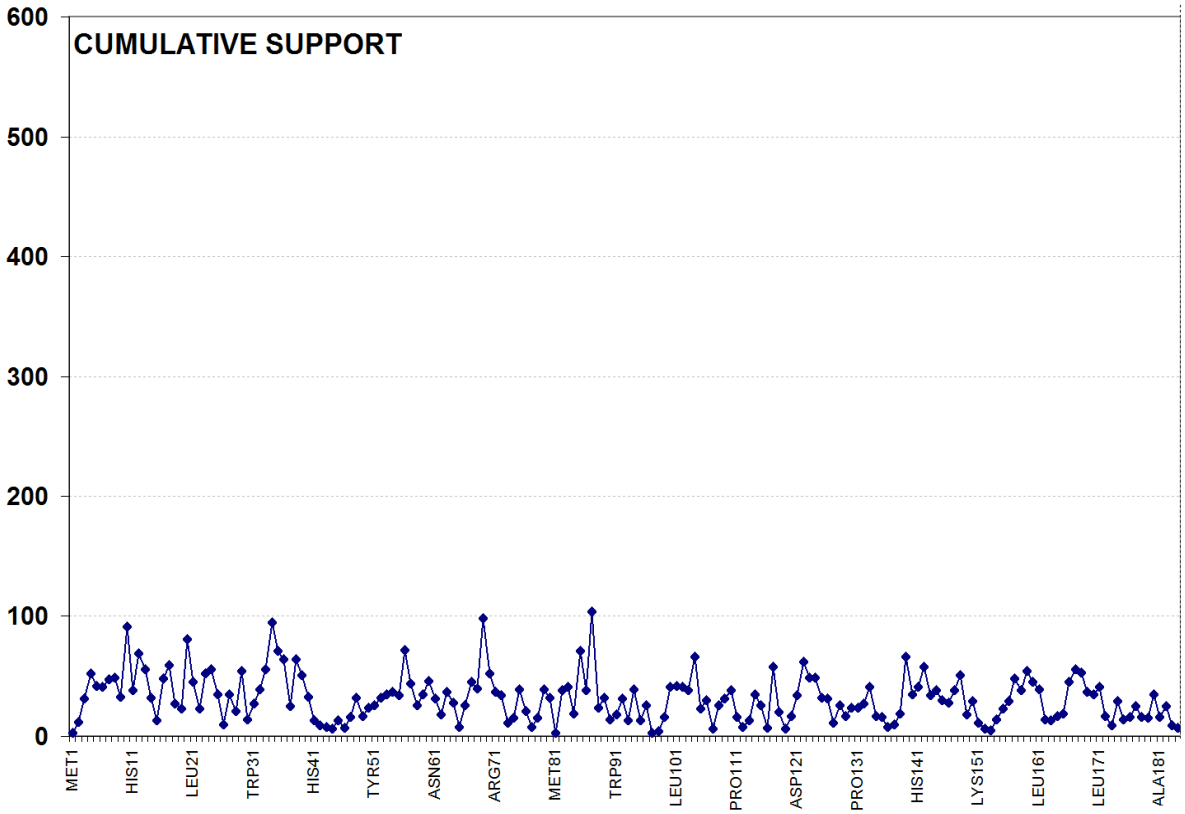


Figure 2.8: The CSP profile generated from background dataset is not present in the test protein YfcE (PDB chain: 1su1A). Ten proteins have been randomly choosen from the background set. The cumulative support profile has been generated for 1su1 using this dataset. As shown above, no remarkable peaks are retrieved.

2.4 Discussion

It is well accepted that proteins accomplish their functions using only a relatively small part of their structures that are highly conserved compared to any other regions. Motifs could be defined as highly conserved amino acid residues forming similar patterns that often represent functionally important regions. Many studies indicated that structural motifs are applicable for protein function annotation especially for detecting of remote homologues. However, few methods can predict structural motifs or conserved key residues based on structural properties alone. The limitations come from the difficulty of local structural alignments and comparison, and data mining on a large scale of protein structure database. In turn, many structural-motif based approaches rely on both sequence- and structure-based features. Theoretically, sequence information is more informative because the number of available protein sequences greatly exceeds the number of available structures. However, the development of the sequence-independent methods needs to be investigated in order to effectively exploit the 3D structure data missing at the sequence level. For example, the sequence-dependent methods, which depend on multiple sequence alignments of the family members, are capable of detecting a sequence motif only if (1) the motif can be aligned, and (2) if amino acid residues in that motif are conserved in terms of following the same order in the primary sequence. Actually, a protein is not a linear string containing one letter amino acid abbreviations as represented by multiple sequence alignments but it is a linear chain of amino acids folding into a unique three-dimensional structure. This implies that, first, family motifs should preserve both amino acid compositions and 3D packing patterns. Second, the amino acid compositions in the family motifs neither are necessary to be aligned in multiple sequence alignments nor follow sequence order. In this point of view, identification of

structural motifs from structure information alone is more challenging. That is because the identified structural motifs can represent both conserved residue compositions and their packing patterns but are not restricted to have similar sequence conservation.

Bandyopadhyay *et al* previously reported the application of FFSM, the sequence-independent AFP method, for function inference of proteins at structural levels. The family function was inferred to the test protein if significant numbers of family-specific motifs present in that protein[10]. In this study we extend the application of FFSM for predicting both structural motifs and conserved key residues in Metallophos proteins. However, the main concerns of FFSM are that the method defines a motif as a maximal subgraph (a graph that is not part of any larger frequent subgraph) in which every node connects to each other. This motif definition increases the possibility of missing motifs that are not fully interconnected subgraphs or/and are only substructures of a highly rigid motif (e.g., a large maximal subgraph). The novel CASIM has been developed to provide additional information retrieved from FFSM, and is expected to detect motifs missed by FFSM or vice versa. The application of CASIM for predicting protein function and conserved key residues is being reported for the first time herein. The idea behind the novel CASIM method is that; (1) the method defined the motif as an ensemble of neighboring Delaunay tetrahedral. This motif definition is different from the rigid structure of fully interconnected subgraph adopted by FFSM, (2) the substructure of a larger motif is also taken into account if that substructure is specific to the family, and (3) the method guarantees the motif conservation by providing a series of constitutive and geometrical descriptors such as amino acid composition, volume and ESA; in addition, the final set of family motifs can then be analyzed, visualized instantaneously in

the PROTMAN software (see **Figures 2.5** and **2.6**) via the PyMol[66] program so that the motif matching and location can be revealed on protein structures.

Both FFSM and the novel CASIM were able to capture structural motifs in the Metallophos proteins members by means of graph mining. Therefore, the methods can bypass the process of multiple structure alignments or aligning of local structures. We have showed that Metallophos motifs retrieved by both FFSM and CASIM are very specific to this family represented by a training set of ten protein members. We were also able to check the specificity of the identified motifs by scanning the motifs on the large set of 6,605 non-redundant protein structures outside the Metallophos family. These motifs, discovered in complimentary fashion by both approaches, included a set of eight conserved key residues. It is important to underline that Metallophos-specific structural motifs could not be simply detected or visualized on the family structures by multiple structure alignments of the training set. However, based on the known eight conserved key residues detected by us, we could easily reveal the family motifs on the training-set structures. Interestingly, these conserved key residues occurring in all training-set members can be reasonably well superimposed, and are located around the metal-binding sites (see **Figure 2.9**). The fact that these eight residues represent only about 2-3% of the entire amino acid residues in each training set member implies the efficiency of FFSM and CASIM that were able to detect low similarity in the training set.

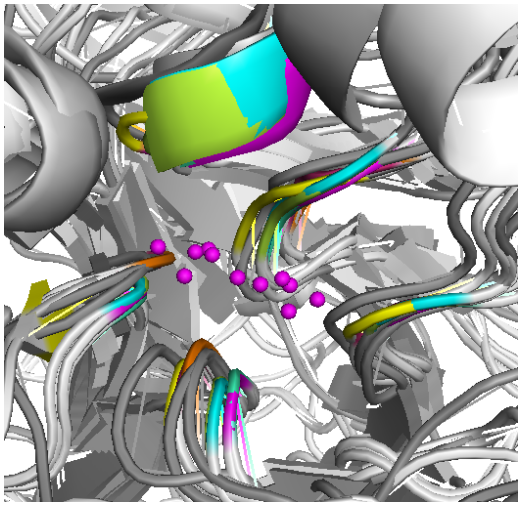


Figure 2.9: Metallophos-specific motifs retrieved by CASIM-FFSM correspond to structurally conserved protein regions at the metal binding sites: Multiple Structural Alignments of Metallophos training set (grey ribbons); Alignments of 8 amino acid residues involved in the Metallophos motifs (colored ribbons); Metal ions (magenta spheres).

We validated the specificity of the identified Metallophos motifs on the three known Metallophos proteins having literature support. It is important to emphasize that these test proteins have low sequence identities (< 20% sequence identity) to our training set. As discussed in Chapter 1, function inference by pair-wise sequence comparisons are unreliable at this sequence identity threshold. We showed that using only number of significant motifs for function inference might not always be suitable. For example, based on number of significant motifs, Metallophos function can be inferred to the test protein if that protein has at least 18 FFSM-motifs or 13 CASIM-motifs (see **Figure 2.3**). By using this cut-off value, only the test protein Glycerophosphodiesterase (PDB code: 3d03) can be annotated as Mettallophos members (see **Table 2.4**). However, we also inferred Metallophos function to protein Vsp29 (PDB code 1z2w) and protein DR1281 (PDB code 1t70) although the numbers of motifs in these proteins fall below the cut-off value. In case of Vsp29, we detected 8 amino acid residues corresponding to the entire 8 conserved key residues found in the training set. In addition, the majority of key residues present in Vsp29 were matched to those reported in the primary literature[67, 68]. In case of 1t70A, CASIM-FFSM retrieved only 1

motif. However, this single motif was highly specific to the family because it was not present in any protein in the background dataset. The limitation of using only significant numbers of motifs for function inference might be related to some technical limitations. For example, as discussed in the previous paragraph, FFSM selected only maximal subgraphs, which sometimes are excessively specific (e.g., a large and rigid subgraph) because the training-set members have high structure similarity compared to the test proteins. The family member will be treated as a false negative even if it consists of a majority part of the highly specific maximal subgraph. To overcome this problem, CASIM considered the partial pattern of the maximal subgraph as the motif if that partial pattern is specific to the family at the given minimum support and maximum background occurrence values. However, the results detected by CASIM (see **Figure 2.4**) showed that the definition of graph edges can sometimes generate two different motifs involving with the same amino acid residues. The present of any of them in the test protein might be sufficient for function annotation. By combining the two methods (CASIM-FFSM), we were able to identify Metallophos motifs in all test proteins. Our prediction was in agreement with the published results.

We compared our results with those from the sequence-based (Pfam), dual sequence and structure-based (enzyme active site template and CRP) and sequence-independent structure-based (reverse template) techniques. The prediction performance of CASIM-FFSM was comparable to those of well-known automated prediction methods. However, it is important to underline the following comparisons. First, our training set consisted of only 10 representative proteins whereas the Pfam hmm profile of Metallophos family was generated from the seed of 330 protein sequences. We found that the hmm profile built from our training set of 10 proteins was not present in any test protein. In addition, Pfam is well

known for predicting protein function whereas our method is suitable for predicting both protein function and conserved key residues. On the other hand, CRP affords good performance on predicting catalytic site residue. However, the method does not offer an option of function inference. Second, our prediction did not require any prior knowledge of functionally important residues. In contrast, the enzyme active site template search and CRP generated the models based on the knowledge of catalytic site residues in the Catalytic Site Atlas (CSA). Third, although reverse template search is more sensitive than CASIM-FFSM, the method has some limitations. The concept of reverse template is to break the query protein structure into a set of 3-residue templates. Then, each template is scanned against the representative set of protein structures. Therefore, High sensitivity of reverse template search probably comes from the small size, flexibility and diversity of reverse templates. However, according to their small size, a 3-residue template might not be desirable for characterizing motifs or functionally important resides. Moreover the method can possibly select meaningless residues to build a template. An example was found in case of protein Rv0805. The Metallophos function of this protein was confirmed by structural and biochemical analysis[69]. The authors of the Rv0805 structure (PDB chain: 2hy1A) reported that the protein was a dimeric, $Fe3^{+}$ -$Mn2^{+}$ binuclear phosphodiesterase based on structural and biochemical analysis. Mutational analysis revealed the active site metals co-ordinateded by conserved aspartate, histidine and asparagine residues. They proposed the structure of the catalytic core in which Asp21, His23, Asp63 and His209 co-ordinate Fe^{3+} whereas Asn97, His169, Asp63 and His207 co-ordinate Mn^{2+}. The structure of Rv0805 (PDB code: 2hy1) was deposited in SCOP 1.7.5. This protein has 13-20% sequence identities to our training set. Therefore, 2hy1was excluded from our test set of known Metallophos function since we focused only on

Metallophos proteins having less than 20% sequence identities to the training set. As reported in Section 3.1, the family motifs and conserved key residues detected by CASIM and FFSM were 13 motifs with 8 residues and 31 motifs with 8 residues, respectively. CASIM and FFSM were able to detect the entire family motifs of 8 key residues in 2hy1A. Among those 8 conserved key residues identified by CASIM-FFSM, five of them (Asp21, Asp63, Asn97, His169, His207) were similar to those suggested by the authors of the structure whereas the other three (Gly62, Gly96, His98) were neighbors to those five amino acid residues. However, the active site residues (Thr138, His186, and Leu201) predicted by reverse template search for 2hy1 did not match to any residues in the published results.

We then predicted Metallophos function and its conserved key residues in nine proteins of unconfirmed functions manually curated into the Metallophos superfamily in SCOP database. These test proteins have remote homology to our training set (sequence identities less than 20%). Thus, their function cannot be simply inferred to the training set using sequence comparison. CASIM-FFSM was capable of detecting the Metallophos motifs and key residues in the structures of 2nxfA, 3ck2A, 1su1A, 1xm7A and 1t71A, but did not detect any motifs in the structures of 2cv9A, 2yvtA, 1nnwA and 1uf3A. Pfam detected Metallophos profile (E-value 0.0015) in 2yvt. However, Pfam identified the hmm profiles of other families in 2cv9 (PGA_cap family, E-value 0.081) and 1nnw (Libosomal_L36e family, E-value 0.25), and did not detect any Pfam profile in 1uf3. Enzyme active site template search failed to afford any hits to the four test proteins missed by us. On the other hand, the first hits of known function for the three test proteins, 2cv9, 2yvt and 1nnw, provided by reverse template search are Metallophos proteins. High sensitivity of reverse template search probably comes from the small size, flexibility and diversity of reverse templates. In contrast,

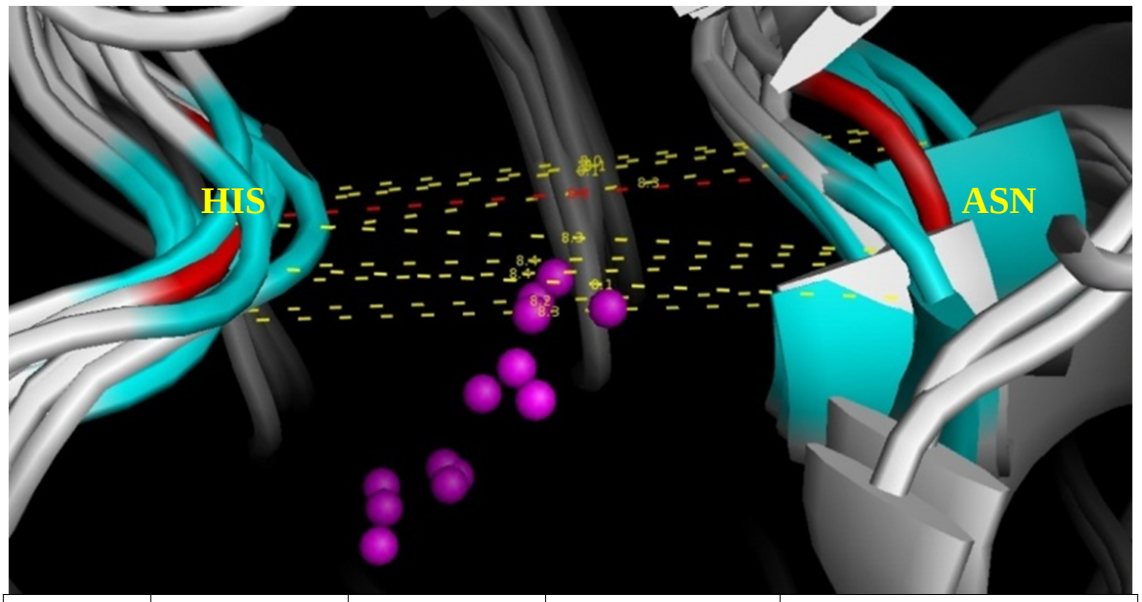
it is important to underline that high sensitivity of the method can sometimes provide false positive results as discussed in the previous paragraph.

Herein, we have presented another novel approach using CSP profiles for predicting the likely conserved key residues. In addition, we found that the Metallophos proteins share specific CSP profiles that are not present in other proteins. Thus, we propose that CSP profile could help detecting and visualizing most conserved residues in protein families and serve as a fast and efficient additional tool for function annotation.

Investigation of structural motif conservation at a sequence level has disclosed some interesting observations. Metallophos-specific structural motifs detected by CASIM-FFSM (FFSM at f=0.8 and CASIM at f=1.0) consisted of 44 motifs involved with 8 conserved key residues (see **Table 2.2**). We have transformed each structural motif into PROSITE-like signatures. Following this strategy we retrieved 44 signatures (31 and 13 signatures related to FFSM and CASIM motifs, respectively). In order to generate the signatures, we mapped residues involved in a given structural motif onto the primary sequences of the training set members and then calculated the distance between those two adjacent residues in each training set member. In a sequence signature, amino acid residues encoded in the structural motif are represented by the standard one-letter codes. Numbers inside figure brackets represent the first and last sequence position of a range of amino acids separating two sequential residues within the motif. For example, the CASIM structural motif DDHH-DGGH-DGHN-DHHN was transformed into the following sequence signature: 'D.{27,42}GD.{25,38}GN.{49,101}H.{33,82}H'. This signature consists of D, any 27 to 42 amino acids, G followed by D, then any 25 to 38 arbitrary amino acids, G, N, any 49 to 101 amino acids, H, any 33 to 82 amino acids, and then H.

In case of protein YfcE (PDB code 1su1), we expected that we would detect many sequence signatures (see Section 2.3.3) in 1su1A sequence as what we found in 1su1A structure (18 motifs containing 7 conserved key residues). Surprisingly, only the signature 'D.{27,42}D.{31,38}N.{48,135}H', transformed from a motif DHND, was present in 1su1A sequence. We then mapped residues involved with the motif DHND onto 1su1A sequence. We found that the signature 'D.{27}D.{35}N.{53}H' in 1su1A sequence is a subset of a corresponding signature 'D.{27,42}D.{31,38}N.{48,135}H' in the training set. We further investigation on the set of 17 motifs found in 1su1A structure but their corresponding signatures were not present in 1su1A sequence. A deeper analysis of these results showed that we could not detect some of those corresponding signatures in 1su1A sequence because of only one non-equivalent residue range. For example, the signature D.{34}GN.{53}H in 1su1A sequence is not matched with the corresponding signature 'D.{24,37}GN.{110,135}H' in a training set because the length between asparagine and histidine in 1su1A sequence is much shorter. It follows that there is low similarity at the sequence level for highly similar structural motifs. To get deeper insights we used multiple structure alignments to map the 7 selected residues corresponding to Metallophos-specific structural motifs onto both the training-set members and the 1su1A structure. Then, we compared the physical distances between pairs of residues with their distances at the sequence level. Examples of distances between certain asparagines and their adjacent histidines are given in **Figure 2.10**. We found that the distances between these two residues in the structures of the training set members and in the 1su1A structure are almost constant: 8.0 - 8.4 Å. However, the distances at the sequence level represented by ranges of residues varied from 110 to 135 residues for the training set members whereas only 53 amino acids

were present between these two residues in the 1su1A sequence. The case studie of YfcE demonstrate how function annotation cannot be obviously achieved at the sequence level only and that local family-specific structural motifs are better suited for efficient function prediction.



PDB ID	ASN	HIS	Distance (Å)	#AA residues between ASN and HIS
1xzw	200	322	8.2	121
1aui	150	281	8.1	130
1g5b	75	186	8.3	110
1hp1	116	252	8.1	135
1ii7	84	206	8.3	121
1kbp	201	323	8.3	121
1qhw	112	242	8.4	129
1s70	124	248	8.0	123
1s95	303	427	8.1	123
1ute	91	221	8.4	129
Min			**8.0**	**110**
Max			**8.4**	**135**
Signature: N.(110,135)H				
1su1	**73**	**127**	**8.3**	**53**
Signature: N.(53)H				

Figure 2.10: Measurement of the distances between two residues (histidine (HIS) and the adjacent asparagine (ASN); **blue:** HIS and ASN in the training set members, **red:** HIS and ASN in 1su1 structure) involved in Metallophos-specific structural motifs retrieved by CASIM-FFSM.

2.5 Conclusions

In this study, we have addressed several challenging problems in the area of automated function prediction as follows: (1) implementation of CASIM-FFSM and CSP-profile search, DT- based methods for predicting protein function based on structure information alone, (2) identification of local similarities (motifs) without aligning of local structures, and (3) prediction of function and the likely functionally important residues in proteins of unconfirmed function having remote homology to the training set. Using the Metallophos family as a test case, we have demonstrated that CASIM-FFSM is capable of detecting Metallophos-specific motifs in a small set of ten Metallophos structures. These family motifs are packed with inclusive information such as the geometry of the motifs, and amino acid types as well as the connection of those amino acid residues. We have established that the Metallophos family specific motifs include residues forming the metal-binding active sites in the training-set members. These family motifs are found in all five test proteins having known Metallophos function according to the literature information. In most cases, the authors of the structures also hypothesized about functionally important residues based on manually structure analysis. The identified key residues detected by CASIM-FFSM that are similar to those from the published results support the experimental hypothesis. The predicting performance of CASIM-FFSM is comparable to the current state-of-the art methods; Pfam, active site template and reverse template searches and CRP. However, the following aspects need to be taken into accounts: (1) our method provides some information that are not captured by Pfam such as 3D structural motifs and their conserved key residues, (2) unlike active site template search and CRP, CASIM-FFSM does not require any prior knowledge of active site residues, and (3) CASIM-FFSM gives more information about the

structural motifs compared to those provided by the reverse template search. Furthermore, we infer Metallophos function and predict conserved key residues that are hypothesized as likely functionally important residues in five proteins of unconfirmed functions having sequence identity less than 20% to the training-set members. We also have verified that structural motifs are better suited for automatic function annotation compared to the corresponding sequence patterns derived from structural motifs. Finally, we have developed a novel approach for generating 3D-1D Cumulative Support Profiles that afford fast and automated identification and visualization of amino acid residues that are conserved within protein families.

2.6 Supplementary data

The published data for test proteins of known Metallophos functions

Glycerophosphodiesterase structure (GpdQ, PDB code 3d03)[56] from *Enterobacter aerogene*

The authors of the structure applied structural, spectroscopic and kinetic techniques to disclose the plausible catalytic mechanism of the protein. They suggested that the amino acid residues involved in the catalytic site were Asp8, His10, Asp50, Asn80, His156, His 195 and His197. In addition, mutation study at Asn80 showed the contribution of this amino acid residue to reactivity.

Recombinant mouse mVps29 (PDB code 1z2w)[67, 68]

Elucidation of the crystal structure of recombinant mouse mVps29 revealed that the protein had similar fold to Metallophos proteins. In addition, mutational analysis of human

Vsp29[67] showed that the enzymatic activity was reduced by alanine substitutions at Asp8, Asn39, Asp62, His86 and His117.

DR1281 structure (PDB code 1t70)[58] from *Deinococcus radiodurans*

The Metallophos function of the DR1281 structure was confirmed by the structural and enzymatic studies. The authors also proposed conserved residues involved in metal binding and catalytic activity based on structure analysis. Those residues were Asp8, Glu37, Asn38, Asn65, His148, His173 and His175.

CHAPTER 3

A NOVEL APPROACH FOR PROTEIN FUNCTION PREDICTION AT THE SEQUENCE LEVEL BASED ON FAMILY-SPECIFIC STRUCTURAL MOTIFS

3.1 Introduction

At present, most Automated Function Prediction (AFP) approaches have been developed for assessing protein function at the sequence level because of the following major reasons. **First,** the number of protein sequences without known function greatly exceeds the number of their structures[70], and thus is critical. **Second,** it is well known that protein structures are more conserved and informative than their corresponding sequences[19]. However, it is not completely clear whether using structural information alone is better than relying on sequence information for inferring protein function reliably. The main concern is due to the limitation of available structural data. **Third,** as we discussed in Chapter 1, extracting meaningful information from protein structures are limited by some computational technical difficulties (e.g., multiple structure alignments, aligning of local structures, and scanning of motifs on the large scale of protein structure database). Therefore, studying the relationships between protein sequences, structures and function is traditionally based on the *sequence-to-structure-to-function* paradigm.

In Chapter 1, we reported that FFSM and the novel CASIM, two structure-based AFP approaches used in our study, were able to detect Metallophos-specific structural motifs and the key residues being responsible for biological function of the family, which were

successfully applied for function prediction of Metallophos. CASIM-FFSM overcomes major computational problems (see Chapter 1) by means of graph mining. However, currently, FFSM, CASIM and other publicly available structural motif-based methods are applicable for protein function prediction at structural level only.

In this chapter, we present a novel concept of nontraditional protein function prediction, from structure to sequence to function. We tested our approach on the family of protein-tyrosine kinases (PTKs), a well-studied and well-defined group of proteins. PTKs are enzymes that catalyze phosphorylation reactions by removing the γ-phosphate group from ATP and covalently attaching it to a hydroxyl group of tyrosine site in the substrate. They are key enzymes in many signal transduction pathway[71]. *We formulated the new approach of protein function prediction at sequence level based on family-specific structural motifs.* We applied FFSM to identify structural motifs (frequent subgraphs) conserved in PTKs (CASIM was not used in this study; the process of transforming structural motifs into sequence signatures is still under development.). As structural motifs representing three-dimensional objects could not be directly mapped onto the linear string of protein sequences, we converted those identified structural motifs into PROSITE-like signatures. We then determined the predicting power of the method by scanning those sequence signatures on the large scale of protein sequences. We benchmarked our method with several well-known, sequence-based function prediction methods (PROSITE, PRINTs and profile HMMs searches).

3.2 Methods

3.2.1 Training set of PTK structures

In the Structural Classification of Proteins (SCOP) database, PTK structures are classified as members of Protein kinases, catalytic subunit family (SCOP ID 88854), which includes both structures of PTKs and serine/threonine kinases.

In the area of bioinformatics, EC annotation is widely used to describe function of PTKs (EC 2.7.10: protein-tyrosine kinases). Therefore, we applied EC annotation as functional label to relate PTK structure, sequence and function. Functional label of PTK members were retrieved from the database of 'PDBsum'[9], which annotates functions of protein structures (PDB chains) according to GO term and EC number of their corresponding UniProt sequences. The main concern of PDBsum that needs to be taken in to account is that function annotation is assigned to a whole sequence rather than the structure. Therefore, PDB structure elucidated from a larger multi-domain protein sequence will often have an EC annotation although the catalytic domain has been deleted from the structure. For example, enzyme megakaryocyte-associated tyrosine-protein kinase belongs to EC 2.7.10.2, a family of non-specific protein-tyrosine kinase. Its sequence (Swiss-Prot ID: P42679) in one chain consists of 3 domains; Src homology 3 domain (SH3), Src homology 2 domain (SH2) and tyrosine kinase catalytic domain. Only tyrosine kinase catalytic domain is responsible for tyrosine kinase function. However, its structure (PDB ID: 1jwo chain A), which lacks of tyrosine kinase catalytic subunit is still assigned as a member of EC 2.7.10.2 (see **Figure 3.1**).

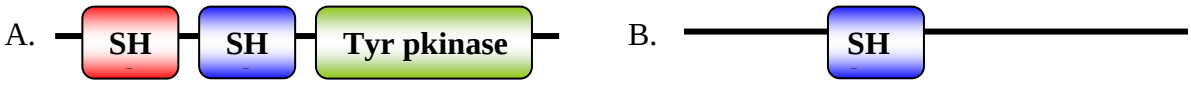


Figure 3.1: A: A protein sequence P42679 consists of 3 domains (SH3, SH2 and tyrosine kinase (Tyr pkinase) domains). **B:** A protein structure 1jwo chain A, a related structure of P42679, consists of only SH2 domain. Only Tyr pkinase domain is responsible for tyrosine kinase activity. However, 1jwo has incorrect annotation as a member of tyrosine kinase family (EC 2.7.10.2) by PDBsum.

In order to establish the dataset of homologous PTK structures sharing similar function annotated by EC number (2.7.10), we defined the function of PTK structures based on EC annotation in PDBsum. Then, those PDB chains were filtered against SCOP. Only PDB chains present in the family of Protein kinases, catalytic subunit family (SCOP ID 88854) were incorporated in the '*PTK-structural dataset'*. This process was aimed to avoid wrong annotation by PDBsum in the case of truncated structures. The PTK-structural dataset consisted of 61 unique protein chains with EC 2.7.10 (Protein-tyrosine kinases or Tyrosine kinases). The PISCES criteria used to generate Metallophos-training set in Chapter 2 (see section 2.2.2) was applied to select '*PTK-training set'*. This representative set of non-redundant entries (PTK-training set) contains 24 protein chains; 1agwA (PDB ID: 1agw, chain A), 1bygA, 1fpuA, 1rjbA, 1fvrA, 1i44A, 1jpaA, 1k2pA, 1k3aA, 1lufA, 1m17A, 1mp8A, 1mqbA, 1oecA, 1pkgA, 1qpeA, 1r0pA, 1sm2A, 1u4dA, 1u59A, 1vr2A, 1xbbA, 2hckA and 2src. Six mutant PDBs in this PTK-training set were curated by replacing the names of modified residues or mutated residues with their native-residue names (see **Table 3.1**). We assumed that the replacement does not change the geometry patterns of the structures. The pair-wise sequence identities between all members in the training set were 18-85.8% (see **Figure 3.2**).

Table 3.1: PTK training-set containing 24 proteins

PDB ID & Chain	Protein name	Mutation	PDB ID & Chain	Protein name	Mutation
2src_	Src		1agwA	FGFR-1	L457V, C488A, C584S
2hckA	HCK		1oecA	FGFR-2	
1qpeA	LCK		1vr2A	VEGFR-2	
1k2pA	BTK		1pkgA	Kit	Y568PTR, Y570PTR
1bygA	CSK		1rjbA	FLT3	
1fpuA	ABL1		1i44A	Insurin receptor	C981S, Y984F, D1161A
1lufA	MuSK		1k3aA	IGF-I receptor	Y1131PTR, Y1135PTR, Y1136PTR
1mp8A	FADK 1		1r0pA	HGF receptor	Y1194F, Y1234F, Y1235D, V1272L
1m17A	ErbB-1		1sm2A	ITK/TSK	
1jpaA	EPH-3	Y604F, Y610F	1u59A	ZAP-70	
1mqbA	ECK		1u4dA	ACK-1	
1fvrA	TIE-2		1xbbA	SYN	

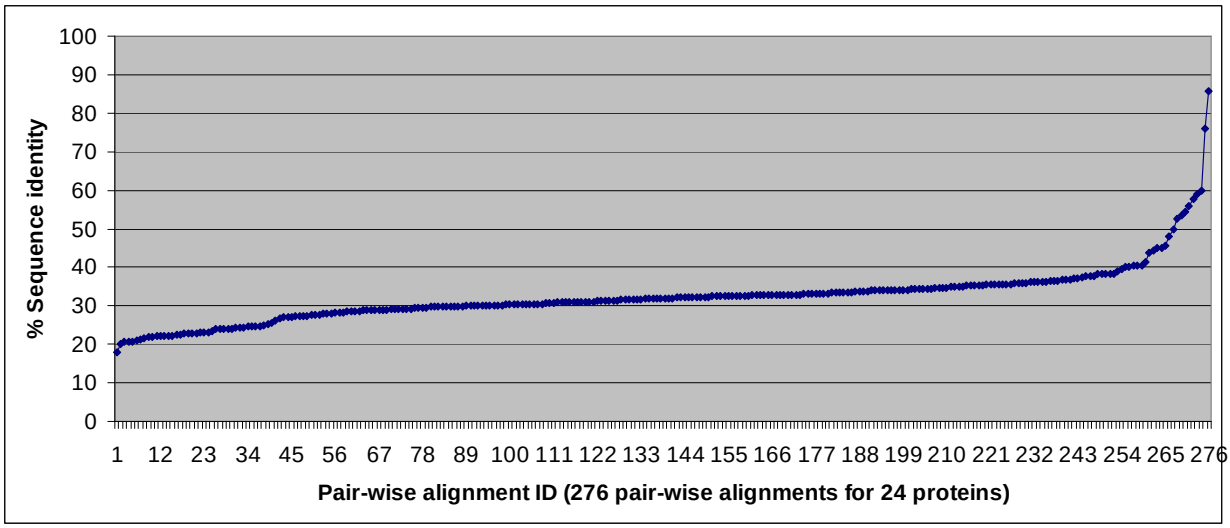


Figure 3.2: Distribution of pair-wise sequence identities in the PTK training-set: *(1) sequence length:* average = 283 amino acid residues, minimum = 245 amino acid residues, maximum = 449 amino acid residues; ***(2) sequence identity:*** average = 32.4%, minimum = 18%; maximum = 85.8%.

3.2.2 Background dataset

The same PISCES criteria used to curate the training set (see section 3.2.1) were applied to protein structures in the PDB (May 2007 release) to build a background dataset. This dataset included 6,605 non-redundant protein chains excluding the 61 PTK structures in PTK-structural dataset.

3.2.3 Identification of structural motifs from PTK-training set using FFSM

The same criteria (see Section 2.2.4) used to identify Metallophos motifs by FFSM were applied to mine non-redundant structural motifs in PTK-training set with a given minimum support (f) values varied from 0.5 to 1.0 and maximum background occurrence (b) value 0.1%.

3.2.4 Transformation of structural motifs into sequence signatures

Each structural motif identified in the previous step was converted into a PROSITE-like sequence signature. For example (see **Figure 3.3**), consider a motif containing 5 residues; S, P, D, W and C, these residues were mapped onto the primary sequences of protein chains in the family-training set. Then, the distance between those two adjacent residues in the motifs were calculated. In a sequence signature, those five amino acid residues were represented by the standard one-letter code. The numbers inside curry brackets represented the number of arbitrary amino acids between two adjacent residues. If there were two numbers inside the curry brackets, the former represented the minimum number of residues whereas the latter represented the maximum number of residues. If a motif corresponded to more than one sequential order pattern, this motif would be converted into more than one sequence signature as well.

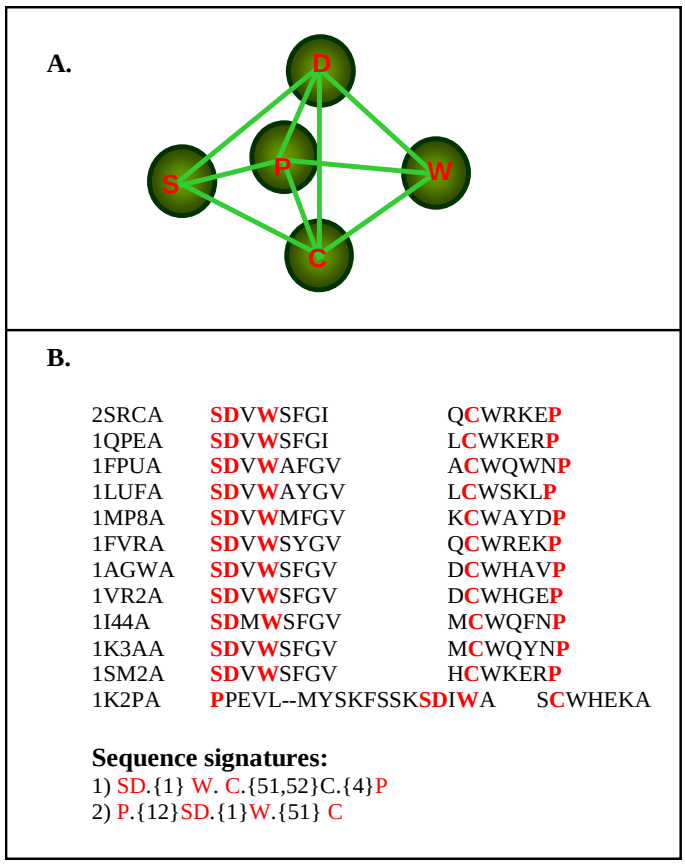


Figure 3.3: Transformation of a structural motif into corresponding sequence signatures.
A: A motif size 5 containing SER, ASP, TRP, CYS and PRO.
B: Deriving sequence signatures by mapping residues in the motif (red) onto primary sequences of the training set (showed examples of some members in the training set). The numbers inside the curry bracket represented the number of arbitrary amino acids between two adjacent residues in the motif.

3.2.5 Test set of protein sequences

A test set of protein sequences (TEST_SET_SEQ) was used to evaluate sequence signature conservation and prediction abilities of our tools on a family of PTKs. This set was obtained from the sequence database of SwissProt release 54.3 (October 2, 2007; 285335 unique protein sequences). After excluding protein sequences related to the PTK-training set, the TEST_SET_SEQ consisted of 285311 protein sequences that were classified into two groups:

(1) The group of '*true family members*' (448 PTK sequences)

(2) A group of proteins outside PTK family assigned as '*BACKGROUND_SEQUENCE_DATASET*' (284,863 protein sequences)

3.2.6 Determination of specific-pattern conservation using precision and recall

The conservation of each sequence signature in PTK sequences was measured on TEST_SET_SEQ dataset using precision and recall values calculated by the following formulas:

$$\%precision = TP/(TP+FP)*100$$

$$\%recall = TP/(TP+FN)*100$$

Here (see Section 3.2.5):

TP was the number of protein sequences in '*true family members*' correctly predicted as family members.

FP was the number of false positives or proteins sequences in '*BACKGROUND_ SEQUENCE_ DATASET*' incorrectly predicted as family members.

FN was the number of false negatives or protein sequences in '*true family members*' that were missed.

3.2.7 Using family-specific sequence fingerprints for function prediction of protein sequences

A '***fingerprint***' is referred as an ensemble of motifs related to conserved regions in a protein family. Each motif is unique; however some features of each motif can be overlapped with other motifs in the fingerprints. Function annotation using fingerprints is based on multiple-motif matching, which is more flexible and powerful than a single-motif approach[10, 72]. Only sequence signatures specific to PTK sequences (% precision at least 90% in the *TEST_SET_SEQ* and presented less than 0.03% in the *BACKGROUND_SEQUENCE_DATASET*) referred as ***'sequence motifs'*** were incorporated in the ***'PTK-specific sequence fingerprints'***. Precision-Recall (PR) curve was applied for selecting the minimum number of sequence motifs that the family members needed to have. Using PR curve afforded more accuracy than its related structure, a Receiver Operator Characteristic (ROC) curve, in an unbalanced dataset[73] especially when the number of negative samples extremely exceeded the number of positive samples such as our TEST_SET_SEQ.

3.2.8 Benchmark methods

Three benchmark methods were used in this study.

(1) A sequence motif search of a PROSITE signature

PROSITE signature or pattern is a unique sequence motif attempting to characterize a short and well-conserved region, such as catalytic site and binding region. **PROSITE pattern** is a single regular expression generated from multiple sequence alignments. Each position can allow one or more amino acids, which are presented by the standard one-letter abbreviations. The acceptable and unacceptable amino acids for a given position are listed

inside the square brackets and the curly brackets, respectively. The letter "x" represented any arbitrary amino acids.

Two PROSITE patterns were used in this study: **(1)** *'PROSITE_pattern1'* was a PROSITE pattern of tyrosine protein kinase specific active-site (PS00109) obtained from PROSITE database[74] (release 20.37 of 23-Sep-2008; and **(2)** *'PROSITE_pattern2'* was constructed from 24 PDB sequences of PTK-training set using PRATT 2.1 program[75, 76] The predicting power of the method was determined by motif searching on TEST_SET_SEQ using 'ps_scan.pl' program[77] with default settings. Protein sequences having such a PROSITE pattern were assigned as 'hits'.

***PROSITE_pattern1*:** [LIVMFYC] – {A} - [HY] - x - D - [LIVMFY] - [RSTAC] - {D} – {PF} - N - [LIVMFYC])

PROSITE_pattern2:S-D-x-W-x-[FY]-G-[IV]-x-[LMV]-x-E-x(4)-[AG]-x(2)-P-[FWY]

(2) A sequence motif search of PRINTS fingerprints

PRINTS is a public database of protein motif fingerprints. The fingerprints are defined as a set of sequence motifs derived from conserved regions in multiple sequence alignments. The PRINTS fingerprints of PTKs (the tyrosine kinase catalytic domain signature; PR00109) were retrieved from PRINTS database[23]. This signature consists of five non-overlapped sequence motifs with 14, 19, 11, 23 and 23 amino acid residues. We searched for the PRINTS fingerprints of PTKs using a searching tool "FingerPRINTSCan" obtained from the fingerPRINTScan package[72]. Hits (protein sequences in TEST_SET_SEQ

containing the PTK fingerprints) were determined according to E-value, which is the expected number of occurrences of sequences scoring greater than or equal to the query's score. The lower the E value is, the more significant the score. E-value calculation depends on the size of the database characterized by the number of amino acid residues in the database and the length of the fingerprint. The database used in this study was the TEST_SET_SEQ containing 285311 protein sequences (see Section 3.2.5) and 1.04751085 X 10^8 amino acid residues.

(3) A protein sequence profile search of profile HMMs

Hidden Markov Models (HMMs) derive a profile or position-specific scoring from the multiple sequence alignment of protein sequence using gap and insertion scores. The profile displays position-specific information about the degree of conservation at various positions in the multiple alignments. Three profile HMMs of PTKs were used in this study. Two of them were Pfam profiles of PTKs (symbol: Pkinase_Tyr; Pfam ID: PF07714) obtained from Pfam database: one was a global alignment model (Pkinase_Tyr_ls.hmm) and another was a local alignment model (Pkinase_Tyr_fs.hmm). These two models were generated from 152 known PTKs. The third profile HMMs of PTKs was generated from the multiple sequence alignment of 24 PTK sequences in our PTK-training set using HMMER program[78].

3.2.9 Benchmarking analysis

The prediction performance of our method was compared to those of benchmarking methods (see Section 3.2.8) on the **TEST_SET_SEQ** using %precision and %recall.

3.3 Results

3.3.1 PTK-specific structural motifs and their related sequence signatures

Small set of PTK-specific structural motifs have been identified by FFSM (see **Table 3.2** column 2) at a given minimum support (f) values varied from 0.5 to 1.0 (e.g., f=0.5; the pattern presents in at least 50% of the training-set members). These specific motifs were found in no more than 0.1% of 6,605 protein structures in the background dataset. The motif sizes were between 4 to 9 amino acid residues.

Each structural motif was then transformed into the sequence signature. At f value 0.5 to 0.9, there were more sequence signatures than their structural motifs (see **Table 3.2**). That was because some structural motifs corresponded to more than one sequential order pattern.

Table 3.2: Number of PTK-specific structural motifs retrieved from PTK structures in the training set (column 2) and number of their corresponding sequence signatures (column 3) at given minimum support (f) values.

Minimum support (f)	# Structural motifs	# Sequence signatures
f=0.5	2956	2996
f=0.6	1728	1750
f=0.7	800	812
f=0.8	391	397
f=0.9	61	62
f=1.0	17	17

3.3.2 Conservation of the sequence signatures in PTKs

Determining family conservation of sequence signatures on TEST_SET_SEQ showed that many sequence signatures, derived from structural motifs retrieved by FFSM, were specific to PTK sequences with high precision. However, some sequence signatures, even derived from structural motifs occurring in most of the members of PTK-training set,

provided low precision. For example, among 17 sequence signatures translated from 17 structural motifs at f = 1.0, four of them obtained %precision less than 50% (see Figure 3.4). On the other hand, two signatures with the highest prediction accuracy, which were HRD.{37,45}W.{14}SD (99.18% precision and 81.47 %recall) and HRD.{37,45}W.{14} SD.{1}W (99.18 %precision and 80.58 %recall) were derived from structural motifs with f values only 0.6 and 0.5, respectively.

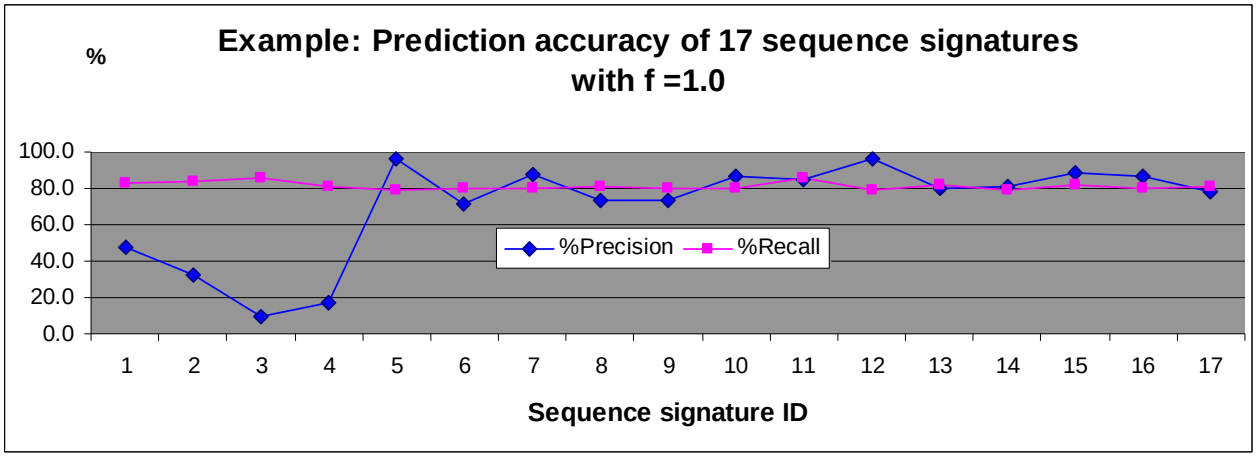


Figure 3.4: Prediction accuracy of sequence signatures derived from structural motifs at f=1.0. The prediction accuracy of each sequence signature was present by %precision (blue) and %recall (pink).

3.3.3 Prediction accuracy of FFSM-based models using PTK-specific sequence fingerprints for function inference of PTK sequences

The results from Section 3.3.2 implied that not all sequence signatures were suitable for inferring PTK function. Therefore, only the signatures specific to PTK sequences (% precision at least 90% in the TEST_SET_SEQ and present less than 0.03% in the BACKGROUND_SEQUENCE_DATASET) referred as ***'sequence motifs'*** were used in this study.

PTK-specific sequence fingerprint was defined as an ensemble of sequence motifs. We generated five sets of fingerprints; A, B, C, D, E and F according to the original set of

structural motifs with f= 0.5, 0.6, 0.7, 0.8, 0.9 and 1.0, respectively. For example (see **Table 3.2** and **Figure 3.5**), Fingerprint A consists of 1236 sequence motifs filtered from 2996 sequence signatures derived from 2956 structural motifs at f=0.5.

Minimum support (f)	**0.5 (12/24)**	**0.6 (14/24)**	**0.7 (17/24)**	**0.8 (19/24)**	**0.9 (22/24)**	**1.0 (24/24)**
# Structural motifs	2956	1728	800	391	61	17
# Sequence signatures	2996	1750	812	397	62	17
Select sequence signatures with precision ≥ 90%						
# Sequence motifs	1236	658	280	121	16	2
PTK-specific fingerprints	**A**	**B**	**C**	**D**	**E**	**F**
Models	A1	B1	C1	D1	E1	F1
	. .	. .	. .	. .	. .	F2
	A1235	B657	C279	D120	E15	
	A1236	B658	C280	D121	F16	

Figure 3.5: Design of PTK-specific fingerprints and FFSM-based models.

The *FFSM-based models* discriminated the family members from other proteins using at least a certain number of sequence motifs in the fingerprints. For example, model A112 required hits to have at least 112 unique sequence motifs in Fingerprint A, which contained 1236 unique sequence motifs. We accessed precision and recall values of each model on TEST_SET_SEQ, and applied Precision-Recall (PR) curve (see **Figure 3.6** and **3.7**) for model selection. Models providing high precision and recall values were present on the upper right-hand corner of the plots. FFSM-based approach successfully led to several models affording high accuracy with precision more than 90% and recall almost 90%.

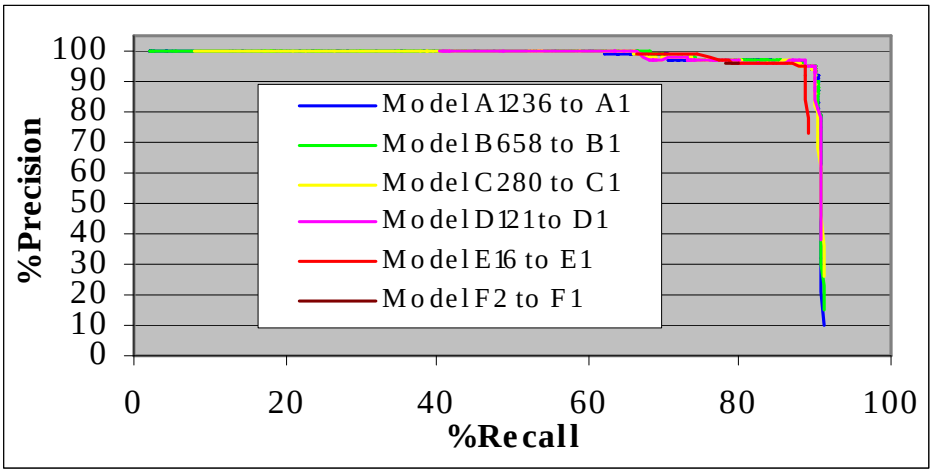


Figure 3.6: Model selection using PR curves. PR curves present %recall (x-axis) and %precision (y-axis) of all FFSM-based models for PTKs. The numbers of sequence motifs required in hits were reduced from left to right. For instance, the precision and recall values of model F2 was on the left hand of those of model F1.

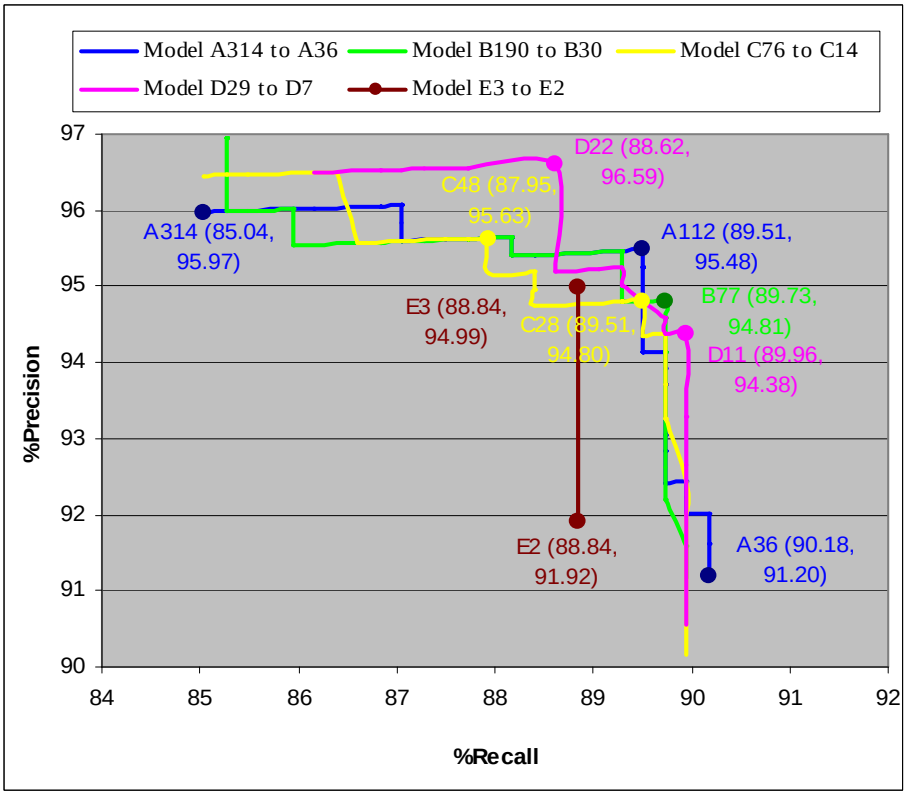


Figure 3.7: PR curves of models preserving recall almost 90% and precision more than 90%. Precision and recall values of some models having high precision and recall were given as examples.

3.3.4 Function inference of new PTK entries

The prediction abilities of FFSM-based models were tested by applying the model A112, one of the best models, to five new PTK sequences added to ExPASY-ENZYME database (enzyme.dat released on April 8, 2008). The model successfully retrieved all new PTK entries. The number of sequence motifs in the new PTK entries, PID O19064 (JAK2 from pig), PID Q5RB23 (JAK2 from Pongo), PID Q75R65 (JAK2 from chicken), PID Q17R13 (ACK-1 from bovine) and PID Q5U2X5 (ACK-1 from rat) were 500, 441, 500, 441 and 500, respectively.

3.3.5 Comparing prediction accuracy of FFSM-based and benchmark methods

Motif HRD.{37,45}W.{14}SD.{1}W, model A77 and A112 were used as the representatives of FFSM-based approach. In order to evaluate the prediction performance of FFSM-based models, we performed a benchmarking study with PROSITE, PRINTS and profile HMMs searches (see Section 3.2.8) because these approaches are very used worldwide and all based on a similar strategy, which are the use of retrieved family-specific patterns for function inference of external protein sequences. Prediction accuracies of the four methods were determined in terms of precision and recall on the same sequence database of TEST_SET_SEQ.

We compared the predicting power of FFSM-based methods with those of motif searches of two PROSITE signatures (see **Figure 3.8**); one obtained from PROSITE database and another was generated from 24 PTK sequences in our PTK-training set. We found that FFSM-based models and a method using PROSITE pattern1 provided highest coverage with almost 90% recall. However, using PROSITE pattern1 affords lowest precision compared to

other approaches and that was about 25-30% lower than FFSM-based approach. Using a more restrict form of PROSITE pattern2, which required higher number of residues in the motif, increased the precision but reduced the recall value dramatically. However, when compared the sequence motif search of PROSITE pattern2 with that of the single motif HRD.{37,45}W.{14}SD.{1}W, the latter achieved better performances in both recall and precision. These results were taking into account that the motif identified by us and PROSITE pattern2 were derived from the same training set. In addition, our FFSM motif contained only 7 residues, which was 3 times less than the number of residues in PROSITE pattern2 (21 residues).

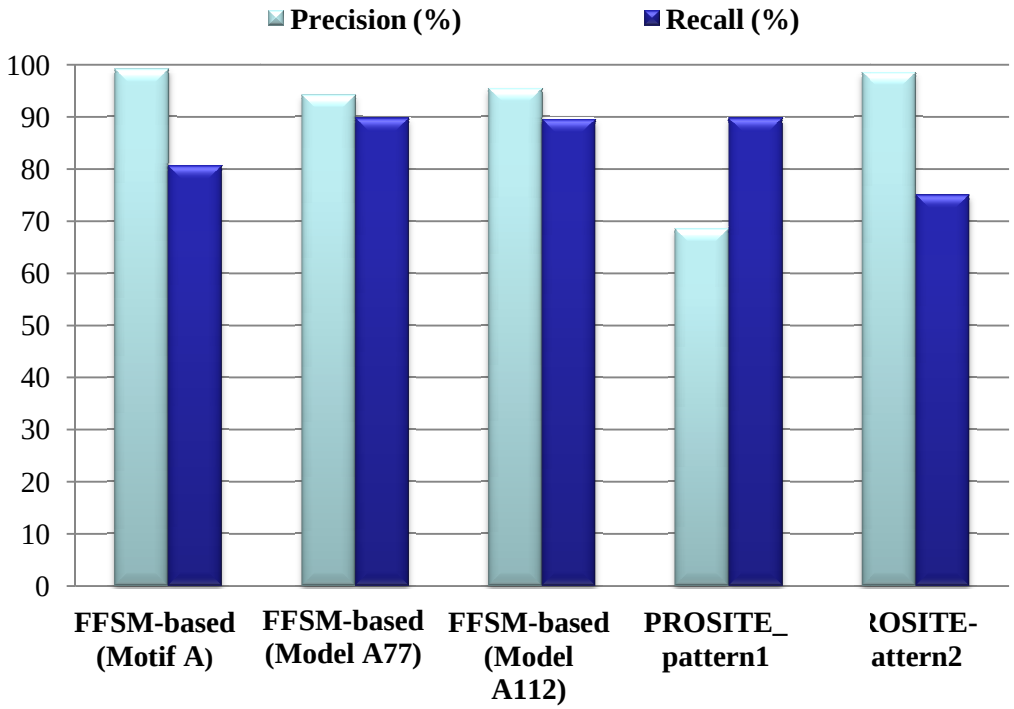


Figure 3.8: Precision and recall comparison of FFSM-based methods and motif searches of PROSITE and PRINTS. (1) FFSM-based methods; ***Motif A:*** HRD.{37,45}W.{14}SD.{1}W; ***Model A77*** (requires hits to have at least 77 sequence motifs from fingerprint A); ***Model A112***, and (2) motif searches of PROSITE signatures: ***PROSITE*_pattern1** (Tyrosine protein kinase specific active-site, PS00109, PTK signature obtained from PROSITE database); ***PROSITE_pattern2*** (PTK signature derived from PTK sequences in PTK-training set).

We used PR curves to compare the predicting power of FFSM-based models corresponding to Fingerprint A with those of FingerPRINTSCan of PTK fingerprints (PR00109, E-value cut off: $1X10^{-40}$ to 0.01) and profile HMMs searches of three profile HMMs of PTKs (E-value cut off: $1X10^{-100}$ to $1X10^{-20}$); one profile HMMs created from our PTK-training set, and two profile HMMs (a global and a local alignment models) obtained from Pfam database (see **Figure 3.9)**. Results showed that the precision and recall provided by FFSM-based models were higher than those of PRINTS methods and comparable to those of profile HMMs searches.

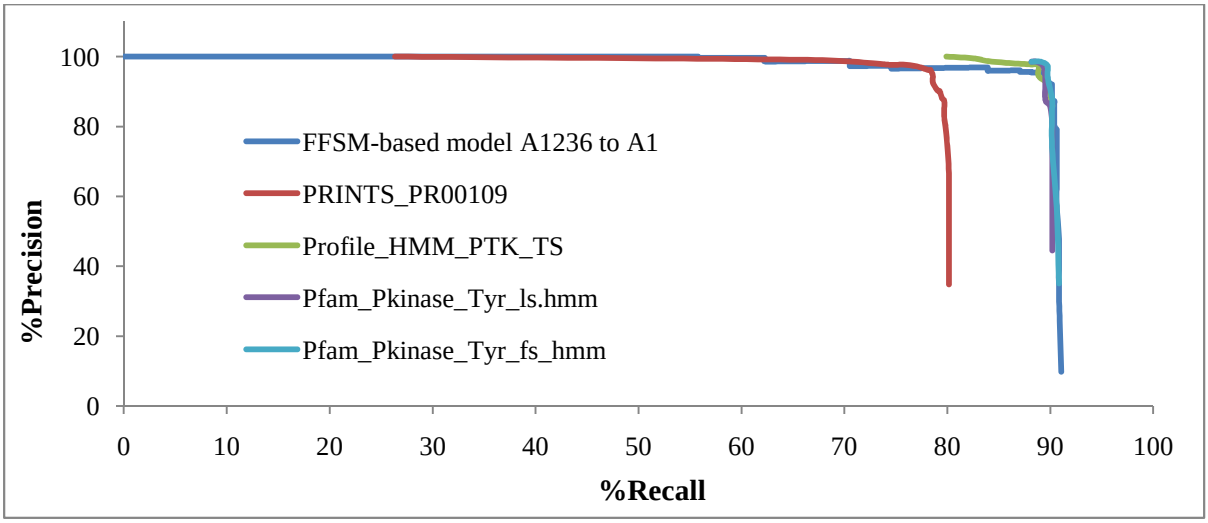


Figure 3.9: PR curves of FFSM-based models and benchmark methods. (1) FFSM-based model A1236 to A1, (2) PRINTS using tyrosine kinases catalytic domain signature (PR00109, E-value: E-value: $1X10^{-40}$ (left) to 0.01 (right)), (3) profile HMM of PTK (PR00109, E-value: $1X10^{-100}$ (left) to $1X10^{-20}$ (right)): (3A) generated from PTK-training set (Profile_HMM_PTK_TS), (3B) obtained from Pfam database; Pfam_pkinase_Tyr-ls.hmm (a global alignment model) and Pfam_pkinase_Tyr-ls.hmm (a local alignment model).

We further assessed the quality of precision measurement by analyzing the set of false positives. We used ‘EC 2.7.10’ (a group of PTKs) as a standard of ‘true’ annotation. We defined a set of ‘false positives with curated EC annotation’ as a set of false positives having known EC numbers (see **Table 3.3**). Proteins labeled with “uncharacterized”, “probable” or “putative” were excluded from this set. We then assigned penalty score for incorrect annotation at the first, second and third level of EC number with scale 3, 2 and 1, respectively, and consequently calculated the total penalty score for each approach. FFSM-based methods and profile HMMs searches provided lower total penalty score compared to PROSITE pattern1 and PRINTS searches. For the search of PROSITE pattern1, there were 28, 2 and 62 wrong annotations at the first, second and third level of EC number, respectively. FFSM-based methods, PRINTS and profile HMMs only gave wrong annotations at a third level of EC number, and all of their false positives fall into two groups of EC annotation; 2.7.11.1 (Non-specific serine/threonine protein kinase) and 2.7.11.25 (Mitogen-activated protein kinase kinase kinase). The results implied that PTKs may be related to these two groups of serine/threonine protein kinases than to any other proteins.

Table 3.3: Penalty score comparison of FFSM-based methods and benchmark methods. The representative models of PRINTS and profile HMMs searches were selected from their best models using PR curves.

Approaches	Methods	#FP	#FP with curated EC annotation	Total penalty score
FFSM-based	HRD.{37,45}W.14SD.{1}W	3	2	2
	A77	25	18	18
	A112	19	12	12
Sequence-motif based	PROSITE pattern1	185	92	150
	PROSITE pattern2	5	0	NA
	PRINTS (E-value: $1X10^{-12}$)	38	27	27
profile HMMs	Profile_HMM_PTK_TS (E-value: $1X10^{-30}$)	28	25	25
	Pfam_Pkinase_Tyr_ls.hmm (E-value: $1X10^{-70}$)	18	15	15
	Pfam_Pkinase_Tyr_fs_hmm (E-value: $1X10^{-80}$)	11	1	1

3.4 Discussion

A rise in the number of proteins having unknown functions have motivated the development of computational tools for predicting molecular function. Function annotation of protein sequences is more popular because (1) the number of protein sequences without function annotation is greatly exceeds the numbers of their structures, and (2) there is abundance of protein sequence data compared to a much smaller number of protein structures.

The conserved pattern of amino acid residues termed '*motif*' often represent functionally important regions that has been adopted for inferring protein functions by many studies[12]. In general, sequence motifs are derived from multiple sequence alignments of

proteins with similar functions. Structural motifs are more difficult to obtain directly from structural data because of some computational difficulties such as the problem of local structural alignments and comparison, and data mining on a large scale of protein structure database. Consequently, the major trend of protein function prediction at structural levels using structural motifs has been relied on sequence-to-structure-to-function pattern. Many studies reported that sequence conservation could be applied for assigning function to protein structures[79]. However, it needs to be underlined what has long been accepted that protein structure is more highly conserved than its sequence especially at the functionally important regions. In addition, our results in Chapter 2, which investigated function prediction of proteins at structural levels using structural motifs, indicated that function annotation cannot be completely achieved at that sequence level only (see Section 2.4).

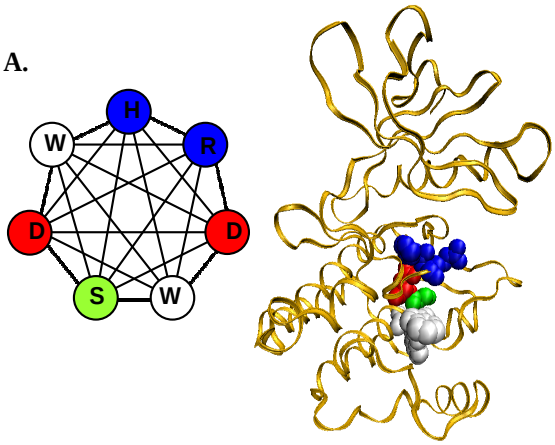
We have realized the important of function prediction of proteins at sequence level. The study reported in this chapter is the first attempt aimed to investigate the non-traditional concept of function annotation, from structure to sequence to function. We reported the novel approach for function prediction of proteins at sequence level using family-specific patterns derived from structural motifs originally extracted from protein structures. The idea behind this work is according to our trust that structural motifs are better candidates compared with sequence motifs for function inference because they represent both conserved residue compositions and their packing patterns but are not restricted to have similar sequence conservation (discussed under Section 2.4).

We applied FFSM on the small representative set of PTK structures to identify PTK-specific structural motifs, which were then transformed into the PROSITE-like signatures (sequence signatures). We have taken into account that the number of available protein

sequences greatly exceeds the number of available structures; sequence signatures derived from conserved patterns in three-dimensional protein structures may be either conserved in primary sequences or have no sequence conservation. We scanned those sequence signatures against the large scale of protein sequences (see Section 3.3.2). We found that sequence signatures derived from structural motifs highly conserved at structural level were not always conserved at sequence level. The results indicated the benefit of combining sequence and structural data in family-motif identification.

The two main problems of sequence motifs excised from multiple sequence alignments are the restriction of residue pattern in which the residue compositions need to be proximity in a primary sequence and have conserved sequential order. Our approach outperforms those limitations because the sequence signatures were derived directly from structural motifs in which the residue compositions only need to be contiguous in 3D space (regardless of residue proximity) and were sequence-order independent. Consequently, the structural motif could be related to more than one sequence signature. Therefore, as shown in **Table 3.2**, for f values ranging from 0.5 to 0.9, there were more sequence signatures than their structural motifs. In addition, the amino acid residues in the sequence signature were not required to be neighbors in a primary sequence. For instance, **Figure 3.10** illustrates a structural motifs and its corresponding sequence signature HRD.{37,45}W.{14}SD.{1}W, in which residue D and W were separated by 37 to 45 residues along a sequence. This signature was highly specific to PTK sequences with 99.18% precision and 80.58% recall.

A.



B. HRD.{37,45}W.{14}SD.{1}W

Figure 3.10: A structural motif and its sequence signature. A: A structural motif (left) corresponding to amino acid packing pattern visualized by Visual Molecular Dynamics (VMD) software on proto-oncogene tyrosine-protein kinase LCK: (PDB ID: 1qpe chain A) (right): The residue compositions are colored by their chemical types; white (TRP, TRP): non-polar; blue (HIS, ARG): basic; red (ASP, ASP): acidic; green (SER): polar. **B:** The related sequence signature.

We adopted the concept of using multiple sequence motifs (fingerprints) for protein function inference from PRINT approach[23], which aimed to improve the sensitivity of PROSITE[22] that infers protein function using only a single sequence motif. The family members do not need to comprise of all motifs in the fingerprints. However, the greater the number of sequence motifs in the fingerprints the hit has the more likely it is the family member. The results from PR curves (see **Figure 3.6**) showed that there was a risk of missing family members if the required number of motifs was high (low recall value), or on the contrary, there was a risk of retrieving too many false positives if the required number was too low (low precision value). When compared to the best sequence motif HRD.{37,45}W.{14}SD.{1}W, one of the best selected model A112 provided better recall but lower precision (see **Figure 3.8**). Using the fingerprint approach was likely to increase the coverage of known members but to reduce the precision of the method. It was desirable to guarantee high precision while allowing a limited loss in coverage. We suggested two approaches for function inference of PTK sequences; one to ensure precision using the single sequence motif and another to ensure coverage using the fingerprint approach.

We investigated the distribution of Fingerprint A, which was related to one of the best models A112, within the structures of PTK-training set using multiple structure alignments (see **Figure 3.11**). The results illustrated that the fingerprints were located at the same region of the training-set members and only present at the C-terminal lobe. The fingerprints also have conserved conformation with average RMSD 0.81 Å and standard deviation 0.22 Å. Visualization of Fingerprint A on the structure of LCK (1qpeA, see **Figure 311B**) showed that the fingerprint was present around known catalytic loop including an aspartic acid residue, which was believed to function as the catalytic base[80]. In addition, the fingerprints probably involved with structure stability of the catalytic loop through the effect of their non-polar residues adjacent to the catalytic loop.

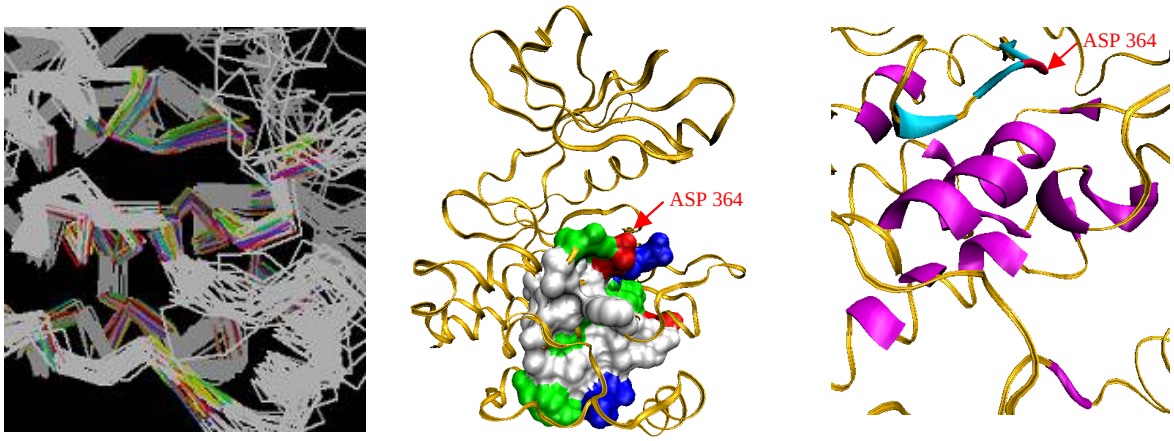


Figure 3.11: PTK-specific sequence fingerprints mapped on the structure of PTK-training set. **A:** Multiple structural alignments of protein chains in PTK-training set (white) performed by MultiProt software and visualized of Fingerprint A (color) by kinemages. **B:** The distribution of 31 amino acid residues of Fingerprint A on LCK (PDB ID: 1qpe chain A) viewed by VMD; **B1:** The residues in the fingerprint are colored by their chemical types; white: non-polar; blue: basic; red: acidic; green: polar. **B2:** The larger image of Fingerprint A colored by purple, blue and red where blue is a catalytic loop and red is a potential catalytic base, aspartic acid (ASP 364).

The results from both quantitative and qualitative assessments demonstrated that FFSM-based methods significantly outperformed the sequence motif searches of PROSITE patterns and PRINTS fingerprints and were comparable to the profile search of pfam HMMs for this PTK family. This result can probably be explained by the difference in their natures of implementation. *First*, the PROSITE pattern afforded high recall values because it allowed more than one amino acid residues at a given position. However, this also increased the risk of adding negative samples. Inferring function using a single pattern only may increase the risk of a wrong annotation if the protein function is related to conserved amino acid residues separated along a protein sequence. *Second*, the PRINTS fingerprints of PTK signature (PR00109) containing five sequence motifs were expected to provide better recall than PROSITE signature, which relied only on a binary decision of the presence or miss of one pattern. In contrast, these PRINTS signatures gave lower recall (PROSITE pattern1: %recall = 89.73%; PRINTS (E-value = $1X10^{-12}$): %recall = 79.02%) maybe due to their requirements concerning the conservation order among those five sequence motifs. Third, there was a limitation of deriving sequence motifs from sequence alignments, the strategy adopted by PROSITE and PRINTS. The residues in PROSITE pattern or each sequence motif in PRINTS signature needed to be neighbors in a primary sequence and required sequential order. Indeed, the family-spatial motif may consist of residues separated along the sequence but contiguous in 3D space and preserve more than one sequential order. Since our method derived family motifs from protein 3D structures, the method was independent from those limitations (see **Figure 3.12**). Profile HMMs represents position-specific scoring of amino acid residues in multiple sequence alignments instead of sequence motifs in order to reduce the restriction found in sequence motifs. Compared to profile HMMs searches, the results

from **Figure 3.9** and **Table 3.3** showed that our methods provided comparable but not exactly the same predicting power. Another example was found in case of protein EGFR (Swiss-Prot ID P55245). This PTK protein was detected by our model A112 but missed by all benchmark methods including profile HMMs searches. The results implied that our method may provide additional information missing from sequence alignment-based methods.

PTK-fingerprints detected by Model A112

PTK pattern detected by PROSITE (PS00109)

PTK fingerprint detected by PRINTS (PR00109: 5 sequence motifs)

Figure 3.12: PTK-specific patterns on protein LCK (PDB ID: 1qpe chain A): PTK-fingerprints detected by Model A112 (left), PTK pattern (PS00109) detected by PROSITE (middle), and PTK fingerprint (PR00109: 5 sequence motifs) detected by PRINTS.

3.5 Conclusions

In this study, we report a proof-of-concept study where FFSM, a structural-motif based approach, is applied for assessing the function of protein sequences. Tested on PTKs, the approach employed PTK-specific sequence fingerprints to infer function. The fingerprints were derived from structural motifs extracted from structural data using FFSM, and then refined by sequence data. We described the fingerprint as an ensemble of sequence motifs represented by regular expression structures. Therefore, we could easily utilize a simple text-

mining approach for the virtual screening of the fingerprints on a large scale of protein sequences. This technique fulfills the limitation of most structure-based approaches, which limit their application to only function inference of protein structures. In addition, the simplicity of our approach overcomes the difficulty of most structure-based approaches, which rely on complex and CPU-time consuming algorithm resulting in the drawback of runtime. In this study, our FFSM-based approach was able to detect PTK-specific fingerprints located around known active site including a potential active site residue, aspartic acid. The prediction was achieved without prior knowledge of functional site. We assessed the performance of FFSM-based approach and other sequence-motif based approaches (PROSITE pattern, PRINTS signature and profile HMMs) in function prediction of PTKs. The results indicated that our method obtains high prediction accuracy in both qualitative and quantitative assessment. Our approach was designed to provide additional information that may not be detected by simple sequence alignments. The study suggests the benefit of using our method in combination with other existing methods, which probably retrieves additional hits, increases the confidence of annotation or avoids incorrect annotation.

CHAPTER 4

SUMMARY AND FUTURE DIRECTIONS

4.1 SUMMARY

Proteins are integral to most biological processes and functions. Understanding of molecular details of protein function is fundamentally important for many research areas including drug discovery. With large amount of sequence data generated by genome sequencing projects, approximately less than one percent of it is experimentally verified for biochemical activities. In addition, 40% of the nearly 10,000 protein structures solved by Structural Genomics (SG) still have unknown function in the PDB[70]. Computational function prediction has become all the more critical in recent years by assisting and complementing wet-bench experiments in managing large scale genomic data and for providing further opportunity for the discovery of new protein as novel drug targets.

In general, computational approaches for protein function prediction infer protein function by finding proteins with global or local similarity at sequence or structural level. The conserved local patterns of amino acid residues referred as ***'motif'*** often represent functionally important regions. *Function inference using 'Structural motifs' are our special interest because of the following reasons.* **First,** 3D arrangements of functionally important residues are significantly more conserved than the entire sequence and structure[19, 26]. **Second,** structural motifs are capable of elucidating the molecular basis of function through a three-

dimensional (3D) structure containing only a few key amino acid residues; this information provides clues about functionally important regions and amino acid residues that could be valuable for the design of specific ligands and site-directed mutagenesis experiments. It is important to underscore that this information could not be retrieved directly from global similarity or sequence motifs searches. However in spite of great interest, identification of structural motifs directly from protein 3D structures alone is plagued by computational difficulties such as local structural alignments and comparison. As a result, structural motifs derived by most methods rely on sequence information. However, function inference by sequence-independent methods is still a major challenge in order to take full advantage of 3D structural data, which is missing at the sequence level. Only a couple of methods in this category including Fast Frequent Subgraph Mining (FFSM)) and reverse template search have been introduced during the recent decades for predicting structural motifs.

In chapter 2, we reported the new application of two sequence-independent methods, FFSM and the novel CASIM, for predicting not only family-specific structural motifs but also conserved key residues as well. We used these two conserved features for function inference of Metallophos structures. The goal was to improve the coverage and accuracy of function annotation compared with using either of those two features alone. Our approaches were able to capture structural motifs and key residues at the metal-binding active sites of Metallophos proteins in the training set and the test set. The identified motifs and residues were then utilized for function inference of proteins of unconfirmed Metallophos function having remote homology (less than 20% sequence identity) to the training set. In addition, we present a novel method for function inference using Metallophos-specific 3D-1D Cumulative Support Profiles (CSP).

In chapter 3, we reported the novel structural-motif based approaches for function prediction of PTKs at sequence level. This is the first such report of applying structure based methods for function annotation of protein sequences, which demonstrated the non-traditional concept of function inference, from structure to sequence to function. We identified PTK structural motifs from a small set of PTK structures. Each motif was then translated into a PROSITE-like sequence signature. We determined the predicting power of these signatures in the large scale of protein sequences. Signatures specific to PTK sequences were defined as 'sequence motifs'. We found that PTK-specific sequence motifs were located at the catalytic loop of PTKs and included an active site aspartic acid residue. We compared the predicting performance of two methods using our identified sequence motifs; a sequence motif search of single motif (PROSITE-like method) and a sequence motif search of multiple motifs (a fingerprints search, PRINT-like method). The first approach provided higher precision but lower recall. Both of our methods significantly outperformed PROSITE (a single motif search) and PRINTS (fingerprint search), the two sequence-motif based methods. We discussed the possible advantages of deriving motifs originally from 3D protein structures compared to originally from 1D protein sequences.

Compared to other benchmark methods excluding PROSITE and PRINTS, in general, our unique function prediction approaches in both part of the thesis provided comparable predicting power. However, with the advantage of using structural motifs for function inference, our approaches could divulge more comprehensive information potentially associated with protein function (e.g., the 3D structure of the small active site and how the identified key residues fit in that active site).

4.2 FUTURE DIRECTIONS

The results showed in Chapter 1 and Chapter 2 demonstrated successful case studies of novel structural motif-based function prediction of Metallophos proteins at structural level and PTKs at sequence level, respectively. Further investigation should be performed to demonstrate the predictive abilities of the suggested methods by applying these promising methods to different protein families. Selecting reliable data is one of the most important procedures for guiding method development. We recommend further study of enzymes because they are well-studied and well-defined group of proteins. In addition, many enzyme resources with well systematic collection containing carefully curated and continually updated data in various aspects are publicly available.

Another challenging aspect will be the identification of family-specific motifs for protein families, in which family members have different fold types and/or multiple chains and/or multiple domains. The idea behind this interest is that currently most publicly available AFP methods either sequence or structure-based are focused on one-domain or one-chain protein family. For example, (1) DALI, the most widely used structure-based AFP tool, relies on global structural (fold) similarities of single chain proteins, (2) Pfam, the most reliable sequence-based AFP tool, infers protein function based on sequence similarity at a domain level, and (3) our studies described under Chapter 1 and 2 involved only families of one-domain monomeric enzymes with similar fold type. *However, based on our interest in family-specific motifs (local similarity) for function inference, proteins should share similar function regardless of their fold similarity. In addition, there are some cases in which all members of a particular family always function as multi-domain or multi-chain proteins in vivo. This implies that their domain combination or chain combination are probably*

essential for the family function. This assumption is supported by the study of Bashton[81], which indicated that the concert of domains in the multi-domain proteins could either preserve the function of an individual domain or provide new function; therefore protein function of multi-domain proteins should be given to whole structures rather than just either domain. In order to test our assumption, it will be challenging to apply the methods used in the previous two chapters for investigating the functions of the diverse families, in which family members have different fold types and/or multiple chains and/or multiple domains. The goal is to predict the family-specific motifs and their potential functionally important residues in order to relate their structures and sequences with the family functions. Some interested protein families for a given EC classification are reviewed below. The training set of protein structures are retrieved from SCOP database, which classifies proteins based on their three-dimensional structural similarity through the levels of class, fold, superfamily, family, domain and species. We suggest the use of biologically active units of the proteins.

A. Families of one-domain monomeric enzymes: family members occur in more than one SCOP class therefore involving more than one fold type.

Two protein families, phospholipase A_2 and β-lactamase, will be selected to test if similar function occur in proteins in the same family but have different fold types is related to specific motifs.

A.1 Phospholipase A_2 (PLA_2, EC 3.1.1.4)

PLA_2 is an enzyme that catalyzes the hydrolysis of the middle ester bond of substrate phospholipids. The released product, arachidonic acid, is known as a precursor of eicosanoids, which are potent mediators of inflammation[82]. PLA_2 is involved with a broad

range of enzymes. The two major group of PLA_2 are $sPLA_2$ (secreted PLA_2) and $cPLA_2$ (cytosolic PLA_2). Typically, $sPLA_2$ have molecular weight between 13 and 15 kDa and consist of one domain formed by α-helices and containing a Ca^{2+} binding loop. The larger enzyme $cPLA_2$ (molecular weight around 85 kDa) consists of two domains: 1) the catalytic domain containing both α-helices and β-strands that are largely interspersed, 2) the Ca^{2+} lipid binding domain formed by β-sheets[83] (see **Figure 4.1**).

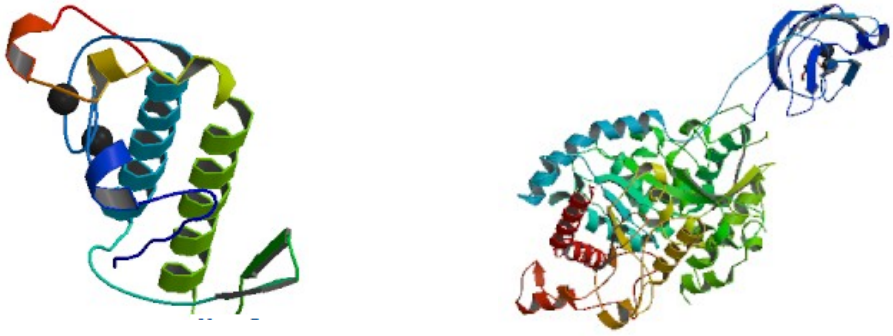


Figure 4.1: Structures of $sPLA_2$ (left; PDB ID: 1n28 chain A) and $cPLA_2$ (right; PDB ID: 1cjy chain A). Two calcium ions are represented with spheres.

The enzyme $sPLA_2$ and $cPLA_2$ have non-detectable sequence homology[84] and completely different three-dimensional structures: these two types of PLA_2 are classified into different fold types by SCOP database (see **Table 4.1**). Nevertheless, we are interested in the fact that they possess the same PLA_2 activity. *It will be interesting to test if there are common motifs specific to both $sPLA_2$ and $cPLA_2$.*

Table 4.1: SCOP classification of phospholipase A_2 ($sPLA_2$ and $aPLA_2$)

SCOP	$sPLA_2$	$cPLA_2$	
Class	All alpha proteins (ID: 46456)	Alpha and beta proteins (a/b) (ID: 51349)	All beta proteins (ID: 48724)
Fold	Phospholipase A2, PLA2 (ID: 48618)	FabD/lysophospholipase-like (ID: 52150)	C2 domain-like (ID: 49561)
Superfamily	Phospholipase A2, PLA2 (ID: 48619)	FabD/lysophospholipase-like (ID: 52151)	C2 domain (Calcium/ lipid-binding domain, CaLB) (ID: 49563)
Family	Vertebrate phospholipase A2 (ID: 48623)	Lysophospholipase (ID: 53645)	PLC-like (P variant) (ID: 49563)
Domain	Snake phospholipase A2 (ID: 48624) Phospholipase A2 (ID: 48637)	Cytosolic phospholipase A2 catalytic domain (ID: 53646)	Domain from cytosolic phospholipase A2 (ID: 49566)

A.2 β-lactamase (EC 3.5.2.6)

β-lactamases are enzymes that catalyze the hydrolysis of an amide bond in the characteristic β-lactam ring of β-lactam antibiotics such as penicillin and cephalosporin families. Based on their amino acid sequences, β-lactamases are grouped into four classes (A, B, C and D). Classes A, C and D act by a serine-based mechanism whereas class B requires zinc cations for their action[85].

In the SCOP database, proteins in class A, C and D have the same fold type, which is different from that of proteins in class B (metallo-β-lactamases) (see **Table 4.2**). However, all four classes have β-lactamase activities. *It will be interesting to test if there are family-specific motifs shared by class A, C and D, which are specific to Class B as well.*

Table 4.2: SCOP classification of β-lactamases (class A, B, C and D)

SCOP	β -lactamases (class A, C and D)	β-lactamases (class B)
Class	Multi-domain proteins (alpha and beta) (ID:: 56572)	Alpha and beta proteins (a+b) (ID: 53931)
Fold	beta-lactamase/transpeptidase-like (ID: 56600)	Metallo-hydrolase/oxidoreductase (ID: 56280)
Superfamily	beta-lactamase/transpeptidase-like (ID: 56601)	Metallo-hydrolase/oxidoreductase (ID: 56281)
Family	beta-lactamase/D-ala carboxypeptidase (ID: 56602)	Zn metallo-beta-lactamase (ID: 56282)
Domain	beta-Lactamase, class A (ID: 56606) AMPC beta-Lactamase, class C (ID: 56618) Class D beta-lactamase (ID: 56622)	Zn metallo-beta-lactamase (ID: 56283)

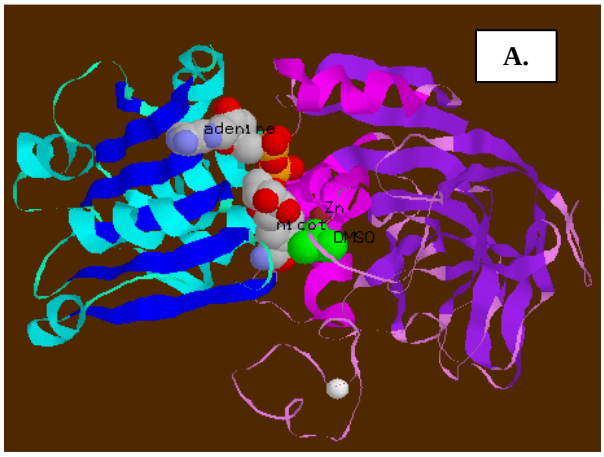
B. Families of dimeric enzymes in which each chain consists of two domains

B.1 Alcohol dehydrogenase (ADH) family (EC 1.1.1.1)

ADH family has been selected for the case study of proteins containing multiple domains and multiple chains. ADH[86] catalyzes the reversible oxidation of alcohols to their corresponding aldehyde or ketone with the concomitant reduction of NAD^+ to NADH. Here, we will focus on a group of zinc-containing ADHs, a homodimer that bind to two zinc cations per unit (chain), for the two following two reasons:

1) Each chain consists of two domains: the catalytic domain and the NAD^+-binding domain (see **Table 4.3**). The inter-domain interface forms a cleft containing the catalytic active site (see **Figure 4.2A**)[87] indicating the role of domain-combination for protein function.
2) The biological units of zinc-containing ADHs always exist as dimers; each dimer is formed by two NAD-binding domains packed together (see **Figure 4.2B**). Thus, it will be challenging to test if our structural motif based approaches can identify family-specific motifs at the domain-domain interface related to the

active site, and if there are family conserved motifs at the protein-protein interface.



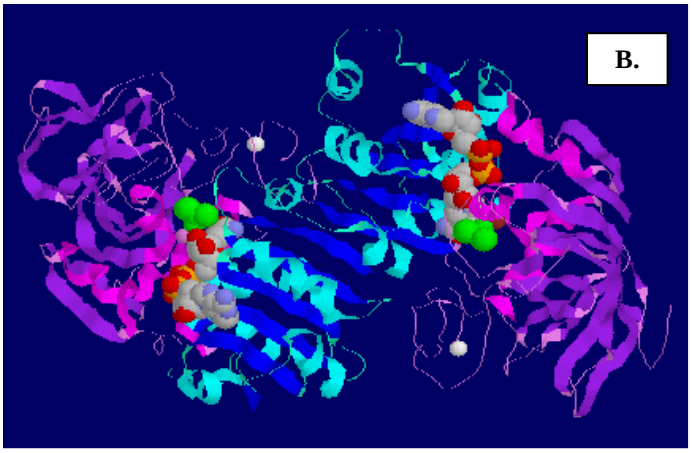


Figure 4.2: The structure of the ADH enzyme family (illustrated by horse liver alcohol dehydrogenase, PDB ID: 6ADH)[87]. A: The NAD^+-binding domain is shown with helix in cyan and sheet in blue. The catalytic domain has helix in magenta and sheet in purple. The substrate is dimethylsulphoside (DMSO) shown in green. The active Zn^{++} ions are in brown and white. NAD+ is colored based on CPK color scheme. **B**: Alcohol dehydrogenase dimer.

Table 4.3: SCOP classification of alcohol dehydrogenases (ADHs); ADH consists of two domains; catalytic domain and NAD^+-binding domain.

SCOP	Catalytic domain	NAD^+-binding domain
Class	All beta proteins (ID: 48724)	Alpha and beta proteins (a/b) (ID: 51349)
Fold	GroES-like (ID: 50128)	NAD(P)-binding Rossmann-fold domains (ID: 51734)
Superfamily	GroES-like (ID: 50129)	NAD(P)-binding Rossmann-fold domains (ID: 51735)
Family	Alcohol dehydrogenase-like, N-terminal domain (ID: 50136)	Alcohol dehydrogenase-like, C-terminal domain (ID: 51736)
Domain	Alcohol dehydrogenase (ID: 50137)	Alcohol dehydrogenase (ID: 51737)

In case of ADH proteins, which involve multiple chains, the training-set members need to have the same number of chains (all dimers for instance) and the latter will be accounted for structural motif identification. The idea is to study if the motifs are required to occur in all chains of the protein structure or only in one chain of the dimeric structure. Moreover, the study could not be complete without the study of interfacial motifs (see **Figure 4.3**).

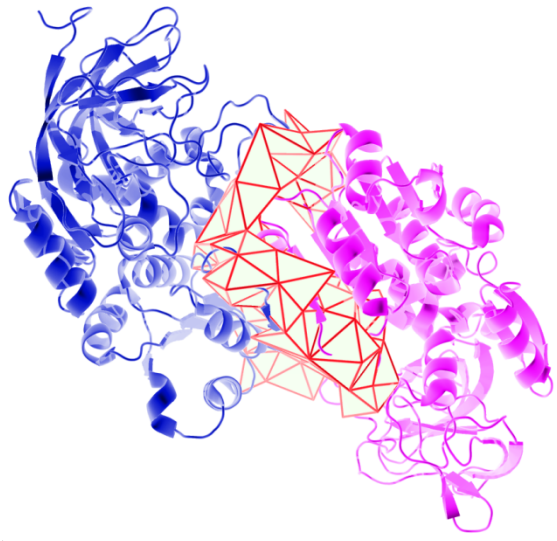


Figure 4.3: Alcohol dehydrogenase class IV sigma (PDB ID: 1D1T: chain A and B, with interfacial motifs (red).

The pioneering approaches reported in this thesis demonstrate the proof of concept of how to effectively exploit the 3D structural data for protein function prediction. In order to improve and build upon the approaches, we recommend the assessments of these approaches on a diversity of protein families. We also underline the limitation of current AFP methods that only aim to elucidate protein function at a domain or chain level. We further suggest our approaches for investigating family-specific motifs in protein families of scientific interest, in which the biological units of family members have different fold types and/or multiple chains and/or multiple domains. The completion of this study will be great interest for researchers in the field of protein function prediction in the years ahead.

REFFERENCES

1. Rost B, Liu J, Nair R, Wrzeszczynski KO, Ofran Y. Automatic prediction of protein function. Cellular and Molecular Life Sciences 2003;60:2637-2650.

2. Thomas PD, Mi H, Lewis S. Ontology annotation: mapping genomic regions to biological function. Curr Opin Chem Biol 2007;11:4-11.

3. Webb EC. Enzyme Nomenclature 1992. Enzyme Nomenclature 1992: Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology, Academic Press, San Diego, California 1992.

4. Tian W, Skolnick J. How well is enzyme function conserved as a function of pairwise sequence identity? J Mol Biol 2003;333:863-882.

5. Tian W, Arakaki AK, Skolnick J. EFICAz: a comprehensive approach for accurate genome-scale enzyme function inference. Nucleic Acids Res 2004;32:6226-6239.

6. Arakaki AK, Tian W, Skolnick J. High precision multi-genome scale reannotation of enzyme function by EFICAz. BMC Genomics 2006;7:315.

7. von Grotthuss M, Plewczynski D, Vriend G, Rychlewski L. 3D-Fun: predicting enzyme function from structure. Nucleic Acids Res 2008;36:W303-W307.

8. Syed U, Yona G. Enzyme function prediction with interpretable models. Methods Mol Biol 2009;541:373-420.

9. Laskowski RA. PDBsum new things. Nucleic Acids Res 2009;37:D355-D359.

10. Bandyopadhyay D, Huan J, Liu JZ, Prins J, Snoeyink J, Wang W, Tropsha A. Structure-based function inference using protein family-specific fingerprints. Protein Science 2006;15:1537-1543.

11. Andreeva A, Howorth D, Chandonia JM, Brenner SE, Hubbard TJ, Chothia C, Murzin AG. Data growth and its impact on the SCOP database: new developments. Nucleic Acids Res 2008;36:D419-D425.

12. Friedberg I. Automated protein function prediction--the genomic challenge. Brief Bioinform 2006;7:225-242.

13. Altschul SF, Gish W. Local alignment statistics. Methods Enzymol 1996;266:460-480.

14. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer ELL, Eddy SR,

Bateman A. The Pfam protein families database. Nucleic Acids Research 2010;38:D211-D222.

15. Lee D, Redfern O, Orengo C. Predicting protein function from sequence and structure. Nat Rev Mol Cell Biol 2007;8:995-1005.

16. Todd AE, Orengo CA, Thornton JM. Evolution of protein function, from a structural perspective. Current Opinion in Chemical Biology 1999;3:548-556.

17. Watson JD, Laskowski RA, Thornton JM. Predicting protein function from sequence and structural data. Curr Opin Struct Biol 2005;15:275-284.

18. Seffernick JL, de Souza ML, Sadowsky MJ, Wackett LP. Melamine deaminase and atrazine chlorohydrolase: 98 percent identical but functionally different. J Bacteriol 2001;183:2405-2410.

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

20. Holm L, Kaariainen S, Rosenstrom P, Schenkel A. Searching protein structure databases with DaliLite v.3. Bioinformatics 2008;24:2780-2781.

21. Kim SH, Shin DH, Choi IG, Schulze-Gahmen U, Chen S, Kim R. Structure-based functional inference in structural genomics. J Struct Funct Genomics 2003;4:129-135.

22. Sigrist CJ, Cerutti L, de CE, Langendijk-Genevaux PS, Bulliard V, Bairoch A, Hulo N. PROSITE, a protein domain database for functional characterization and annotation. Nucleic Acids Res 2010;38:D161-D166.

23. Attwood TK, Avison H, Beck ME, Bewley M, Bleasby AJ, Brewster F, Cooper P, Degtyarenko K, Geddes AJ, Flower DR, Kelly MP, Lott S, Measures KM, Parry-Smith DJ, Perkins DN, Scordis P, Scott D, Worledge C. The PRINTS database of protein fingerprints: a novel information resource for computational molecular biology. J Chem Inf Comput Sci 1997;37:417-424.

24. Skrabanek L, Niv MY. Scan2S: increasing the precision of PROSITE pattern motifs using secondary structure constraints. Proteins 2008;72:1138-1147.

25. Liu ZP, Wu LY, Wang Y, Zhang XS, Chen LN. Bridging protein local structures and protein functions. Amino Acids 2008;35:627-650.

26. Loewenstein Y, Raimondo D, Redfern OC, Watson J, Frishman D, Linial M, Orengo C, Thornton J, Tramontano A. Protein function annotation by homology-based inference. Genome Biol 2009;10:207.

27. Polacco BJ, Babbitt PC. Automated discovery of 3D motifs for protein function annotation. Bioinformatics 2006;22:723-730.

28. Kleywegt GJ. Recognition of spatial motifs in protein structures. J Mol Biol 1999;285:1887-1897.

29. Golovin A, Henrick K. MSDmotif: exploring protein sites and motifs. Bmc Bioinformatics 2008;9.

30. Kristensen DM, Ward RM, Lisewski AM, Erdin S, Chen BY, Fofanov VY, Kimmel M, Kavraki LE, Lichtarge O. Prediction of enzyme function based on 3D templates of evolutionarily important amino acids. Bmc Bioinformatics 2008;9.

31. Laskowski RA, Watson JD, Thornton JM. ProFunc: a server for predicting protein function from 3D structure. Nucleic Acids Res 2005;33:W89-W93.

32. Watson JD, Sanderson S, EzerskY A, Savchenk A, Edwards A, Orengo C, Joachimiak A, Laskowski RA, Thornton JM. Towards fully automated structure-based function prediction in structural genomics: A case study. Journal of Molecular Biology 2007;367:1511-1522.

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

34. Huan J, Bandyopadhyay D, Prins J, Snoeyink J, Tropsha A, Wang W. Distance-based identification of structure motifs in proteins using constrained frequent subgraph mining. Comput Syst Bioinformatics Conf 2006;5:227-238.

35. Fourches D, Tropsha A. PROTMAN (PROTein MANager). In: 2010.

36. Singh RK, Tropsha A, Vaisman II. Delaunay tessellation of proteins: Four body nearest-neighbor propensities of amino acid residues. Journal of Computational Biology 1996;3:213-221.

37. Tropsha A, Vaisman II, Cho SJ, Zheng W. A new approach to protein fold recognition based on delaunay tessellation of protein structure. Abstracts of Papers of the American Chemical Society 1996;212:71-COMP.

38. Bandyopadhyay D, Huan J, Prins J, Snoeyink J, Wang W, Tropsha A. Identification of family-specific residue packing motifs and their use for structure-based protein function prediction: II. Case studies and applications. J Comput Aided Mol Des 2009.

39. Bandyopadhyay D, Huan J, Prins J, Snoeyink J, Wang W, Tropsha A. Identification of family-specific residue packing motifs and their use for structure-based protein function prediction: I. Method development. J Comput Aided Mol Des 2009.

40. Okabe A, Boots B, Sugihara K. Spatial tessellations: concepts and applications of Voronoi diagrams. John Wiley, Chichester; 1992.

41. Bandyopadhyay D, Snoeyink J. Almost-Delaunay simplices: Nearest neighbor relations for imprecise points. In ACM-SIAM Symposium On Distributed Algorithms 2004:404-412.

42. Ilyin VA, Abyzov A, Leslin CM. Structural alignment of proteins by a novel TOPOFIT method, as a superimposition of common volumes at a topomax point. Protein Sci 2004;13:1865-1874.

43. Shah RR, Huan J, Bandyopadhyay D, Wang W, Tropsha A. Structure Based Identification of Protein Family Signatures for Function Annotation. In: 2004.

44. Tropsha A, Carter CW, Jr., Cammer S, Vaisman II. Simplicial neighborhood analysis of protein packing (SNAPP): a computational geometry approach to studying proteins. Methods Enzymol 2003;374:509-544.

45. Youn E, Peters B, Radivojac P, Mooney SD. Evaluation of features for catalytic residue prediction in novel folds. Protein Science 2007;16:216-226.

46. Cilia E, Passerini A. Automatic prediction of catalytic residues by modeling residue structural neighborhood. Bmc Bioinformatics 2010;11.

47. Lo CL, Ailey B, Hubbard TJ, Brenner SE, Murzin AG, Chothia C. SCOP: a structural classification of proteins database. Nucleic Acids Res 2000;28:257-259.

48. Mumby MC, Walter G. Protein serine/threonine phosphatases: structure, regulation, and functions in cell growth. Physiol Rev 1993;73:673-699.

49. Mansuy IM, Shenolikar S. Protein serine/threonine phosphatases in neuronal plasticity and disorders of learning and memory. Trends Neurosci 2006;29:679-686.

50. Stoker AW. Protein tyrosine phosphatases and signalling. J Endocrinol 2005;185:19-33.

51. Tonks NK. Protein tyrosine phosphatases: from genes, to function, to disease. Nat Rev Mol Cell Biol 2006;7:833-846.

52. Winder DG, Sweatt JD. Roles of serine/threonine phosphatases in hippocampal synaptic plasticity. Nat Rev Neurosci 2001;2:461-474.

53. McConnell JL, Wadzinski BE. Targeting protein serine/threonine phosphatases for drug development. Mol Pharmacol 2009;75:1249-1261.

54. Lee D, de Beer TA, Laskowski RA, Thornton JM, Orengo CA. 1,000 structures and more from the MCSG. BMC Struct Biol 2011;11:2.

55. Wang G, Dunbrack RL, Jr. PISCES: a protein sequence culling server. Bioinformatics 2003;19:1589-1591.

56. Hadler KS, Tanifum EA, Yip SHC, Mitic N, Guddat LW, Jackson CJ, Gahan LR, Nguyen K, Carr PD, Ollis DL, Hengge AC, Larrabee JA, Schenk G. Substrate-Promoted Formation of a Catalytically Competent Binuclear Center and Regulation of Reactivity in a Glycerophosphodiesterase from Enterobacter aerogenes. Journal of the American Chemical Society 2008;130:14129-14138.

57. Chen SF, Yakunin AF, Kuznetsova E, Busso D, Pufan R, Proudfoot M, Kim R, Kim SH. Structural and functional characterization of a novel phosphodiesterase from Methanococcus jannaschii. Journal of Biological Chemistry 2004;279:31854-31862.

58. Shin DH, Proudfoot M, Lim HJ, Choi IK, Yokota H, Yakunin AF, Kim R, Kim SH. Structural and enzymatic characterization of DR1281: A calcineurin-like phosphoesterase from Deinococcus radiodurans. Proteins-Structure Function and Bioinformatics 2008;70:1000-1009.

59. Bitto E, Wesenberg GN, Phillips GN, McCoy JG, Bingman CA. Protein Structure Data Summary. In: Center for Eukaryotic Structural Genomics (http://www.uwstructuralgenomics.org/gallery/2NXF.pdf); 2006.

60. Miller DJ, Shuvalova L, Evdokimova E, Savchenko A, Yakunin AF, Anderson WF. Structural and biochemical characterization of a novel Mn2+-dependent phosphodiesterase encoded by the yfcE gene. Protein Sci 2007;16:1338-1348.

61. Eisenberg D, Bowie JU, Luthy R, Choe S. Three-dimensional profiles for analysing protein sequence-structure relationships. Faraday Discuss 1992:25-34.

62. Eisenberg D, Luthy R, Bowie JU. VERIFY3D: assessment of protein models with three-dimensional profiles. Methods Enzymol 1997;277:396-404.

63. Tropsha A, Singh RK, Vaisman II, Zheng W. Statistical geometry analysis of proteins: implications for inverted structure prediction. Pac Symp Biocomput 1996:614-623.

64. Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on TM-score. Nucleic Acids Research 2005;33:2302-2309.

65. Babor M, Gerzon S, Raveh B, Sobolev V, Edelman M. Prediction of transition metal-binding sites from apo protein structures. Proteins-Structure Function and Bioinformatics 2008;70:208-217.

66. Delano W. PyMol v1.2. In: Schrodinger (http://www.pymol.org/); 2010.

67. Damen E, Krieger E, Nielsen JE, Eygensteyn J, van Leeuwen JEM. The human Vps29 retromer component is a metallo-phosphoesterase for a cation-independent mannose 6-phosphate receptor substrate peptide. Biochemical Journal 2006;398:399-409.

68. Collins BM, Skinner CF, Watson PJ, Seaman MNJ, Owen DJ. Vps29 has a phosphoesterase fold that acts as a protein interaction scaffold for retromer assembly. Nature Structural & Molecular Biology 2005;12:594-602.

69. Shenoy AR, Capuder M, Draskovic P, Lamba D, Visweswariah SS, Podobnik M. Structural and biochemical analysis of the Rv0805 cyclic nucleotide phosphodiesterase from Mycobacterium tuberculosis. Journal of Molecular Biology 2007;365:211-225.

70. Erdin S, Lisewski AM, Lichtarge O. Protein function prediction: towards integration of similarity metrics. Curr Opin Struct Biol 2011;21:180-188.

71. Hunter T, Cooper JA. Protein-tyrosine kinases. Annu Rev Biochem 1985;54:897-930.

72. Scordis P, Flower DR, Attwood TK. FingerPRINTScan: intelligent searching of the PRINTS motif database. Bioinformatics 1999;15:799-806.

73. Davis J, Goadrich M. The relationship between Precision-Recall and ROC Curves. In 23rd International Conference on Machine Learning (ICML), Pittsburgh, PA, USA 2006.

74. Hulo N, Bairoch A, Bulliard V, Cerutti L, Cuche BA, de Castro E, Lachaize C, Langendijk-Genevaux PS, Sigrist CJ. The 20 years of PROSITE. Nucleic Acids Res 2008;36:D245-D249.

75. Jonassen I, Collins JF, Higgins DG. Finding flexible patterns in unaligned protein sequences. Protein Sci 1995;4:1587-1595.

76. Jonassen I. Efficient discovery of conserved patterns using a pattern graph. Comput Appl Biosci 1997;13:509-522.

77. Gattiker A, Gasteiger E, Bairoch A. ScanProsite: a reference implementation of a PROSITE scanning tool. Appl Bioinformatics 2002;1:107-108.

78. Johnson LS, Eddy SR, Portugaly E. Hidden Markov model speed heuristic and iterative HMM search procedure. Bmc Bioinformatics 2010;11:431.

79. Sadowski MI, Jones DT. The sequence-structure relationship and protein function prediction. Curr Opin Struct Biol 2009;19:357-362.

80. Cole PA, Grace MR, Phillips RS, Burn P, Walsh CT. The role of the catalytic base in the protein tyrosine kinase Csk. J Biol Chem 1995;270:22105-22108.

81. Bashton M, Chothia C. The generation of new protein functions by the combination of domains. Structure 2007;15:85-99.

82. David AS, Edward AD. The expanding superfamily of phospholipase A2 enzymes: classification and characterization. Biochimica et Biophysica Acta 2000:1-19.

83. Marry FR. Phospholipases: Generation of lipid-derived second messengers. In: Sitaramayya A, editor. Introduction to cellular signal transduction; 1999. 146 p.

84. Leslie CC. Properties and regulation of cytosolic phospholipase A2. J Biol Chem 1997;272:16709-16712.

85. Bush K. Characterization of beta-lactamases. Antimicrob Agents Chemother 1989;33:259-263.

86. Hammes-Schiffer S, Benkovic SJ. Relating protein motion to catalysis. Annu Rev Biochem 2006;75:519-541.

87. Eklund H, Samma JP, Wallen L, Branden CI, Akeson A, Jones TA. Structure of a triclinic ternary complex of horse liver alcohol dehydrogenase at 2.9 A resolution. J Mol Biol 1981;146:561-587.