University of Alberta

In-silico Characterization and Prediction of Protein-Small Ligand Interactions

by

Ke Chen

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Software Engineering and Intelligent Systems

Department of Electrical and Computer Engineering

Fall 2011
Edmonton, Alberta

ABSTRACT

Proteins, which participate in virtually every process within cells, implement many of their functions through interactions with various ligands. Although a substantial effort in characterization and prediction of protein-ligand interactions was observed in the past two decades, these subjects remain far from completion. This dissertation focuses on computational (in-silico) analysis and prediction of the protein-small ligand interactions, with particular emphasis on the protein-nucleotide interactions. We start by analyzing regularities, referred to as the interaction patterns, in the atomic-level protein-small ligand interactions which lead to the discovery of ten interaction patterns that cover majority of the known interactions. The discovery of these interaction patterns demonstrates that protein-ligand interactions can be predicted for a given protein. Next, we performed an extensive comparative analysis of the predictive performance of ten representative methods that predict binding residues and binding sites for small organic ligands. Our results reveal that although the predictive quality of these methods was significantly improved during the past decade, there is still a large room for further improvements, particularly when predicting for certain types of the organic compounds. We also found a few limitations of the existing methods which motivate the development of new predictors of the protein-small organic ligand interactions. Consequently, we proposed two methods that address prediction of the protein-nucleotide interactions. We selected nucleotides from among the organic compounds because they are highly abundant and ubiquitous (they are involved in a wide range of biological processes), and thus they

constitute an important and challenging problem. The first method predicts nucleotide binding residues from protein sequences, and the second method identifies the binding sites from protein structures. We empirically demonstrate that both, the sequence-based and the structure-based, methods significantly improve predictions over the existing state-of-the-art solutions. Our study aims to help with the characterization and annotation of biological functions of proteins and elucidation of the molecular-level mechanisms of cellular activities, and it provides tools that can be used to implement improved molecular-docking based rational drug discovery protocols.

This thesis is dedicated to my parents and Jamboo.

For their endless love, support and encouragement

ACKNOWLEDGMENTS

First and foremost, I would like to express my gratitude to my supervisor Dr. Lukasz Kurgan for providing a good research environment, offering invaluable discussions, and for his guidance during my entire graduate program.

I would like to thank the members of my examination committee, Drs Kihara, Musilek, Reformat, and Tuszynski, for their involvement and insightful feedback that helped me to improve the quality of this dissertation.

I would like to express my appreciation to my lab mates for their valuable discussions, participation in many collaborative projects, and for being great everyday companions.

Last but not least, I would like to thank my family and friends for their continuing encouragement and understanding.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF PUBLICATIONS

The following list enumerates the author's publications pertinent to this dissertation.

Chen, K., and Kurgan, L. (2009) Investigation of atomic level patterns in protein-small ligand interactions. *PLoS ONE*, 4, e4473.

Chen, K., Mizianty, M., Gao, J., and Kurgan, L. (2011) A critical comparative assessment of predictions of protein binding sites for biologically relevant organic compounds. *Structure*, 19:613-621.

Chen, K., Mizianty, M., and Kurgan, L. Prediction and analysis of nucleotide binding residues using sequence and sequence-derived structural descriptors. Submitted (in revision)

Chen, K., and Kurgan, L. Prediction of nucleotide-binding sites and residues using local structure similarity. Submitted

LIST OF ABBREVIATIONS

AA: Amino Acid

ACC: Accuracy

ADP: Adenosine diphosphate

AMP: Adenosine monophosphate

ATP: Adenosine triphosphate

AUC: Area under the ROC Curve

FN: False Negatives

FP: False Positives

GDP: Guanosine diphosphate

GTP: Guanosine triphosphate

MCC: Matthews Correlation Coefficient

PDB: Protein Data Bank

PREC: Precision

PSSM: Position Specific Score Matrix

RBF: Radial Basis Function

REC: Recall

ROC: Receiver Operating Characteristic

SPEC: Specificity

SVM: Support Vector Machine

TN: True Negatives

TP: True Positives

CHAPTER 1 Introduction

Proteins are biochemical compounds which are essential parts of every organism and which participate in virtually every process within cells. They perform a wide variety of biological functions, maintaining the cell shape, catalyzing biochemical reactions, neutralizing antigens during the immune response, and serving as signal receptors during cellular signal transduction, to name just a few (Howard and Hyman, 2007; Gutteridge and Thornton, 2005; Gilman, 1987). The knowledge of structures and functions of proteins is crucial for elucidation of the mechanism of cellular activities and for the development of molecular-docking based rational drug discovery protocols (Meng et al., 2011; Brooijmans and Kuntz, 2003). With the development of high-throughput sequencing techniques, a tremendous quantity of protein sequences was determined in the past three decades. For instance, as of April 5th**,** 2011, the UniProt Archive (UniProt consortium, 2010), a comprehensive publicly accessible protein sequence database, contains 26,004,569 non-redundant protein sequences. In contrast, mostly due to the difficulties in protein expression, purification, and crystallization, our knowledge of structures and functions of proteins is limited, i.e., as of April 2011, the Protein Data Bank (PDB) (Berman et al., 2000), the most comprehensive database of tertiary structures of macromolecules, includes only 67,001 protein structures. Moreover, the annotation of biological functions of proteins in PDB is incomplete. The wide and growing gap between the number of known protein sequences and the number of known protein structures with the annotated biological functions motivates the development of computational tools for protein sequence analysis, protein tertiary structure prediction, and protein function annotation; the latter is the focus of this thesis.

Proteins perform their biological functions through interactions with various molecules, including nucleic acids (DNA and RNA), metals, carbohydrates, other proteins, and small organic compounds. The interactions with the large ligands, such as protein-DNA, protein-RNA, and protein-protein interactions have been

systematically investigated (Ellis et al., 2007; Luscombe et al., 2001; Jones and Thornton, 1996) and dozens of computational methods have been developed for the prediction of DNA and RNA binding sites (Murakami et al., 2010; Gao and Skolnick, 2008) and identification of protein-protein interaction interfaces (Fiorucci and Zacharias, 2010). In contrast, the protein-small ligand interactions were not yet systematically studied. To this end, this dissertation is focused on computational analysis and prediction of protein-small ligand interactions, with special emphasis on protein-small organic compound interactions. Interactions between proteins and small organic compounds are of particular interest because they find applications in elucidation of mechanisms of numerous cellular activities, such as cellular signaling, regulation of cell cycles, and growth of neurons (Mukherjee et al., 2010; Popova et al., 2010; Whittard et al., 2006), to name just a few. The small organic compounds also constitute more than 80% of the drugs approved by the U.S. Food and Drug Administration (Wishart et al., 2008). Consequently, the knowledge of their interactions with proteins, including their binding sites, plays a crucial role in the molecular docking-based rational drug discovery (Meng et al., 2011; Brooijmans and Kuntz, 2003).

1.1 Existing solutions

The studies concerning prediction of the protein-small ligand interactions could be categorized into to two classes based on their inputs. The first class of methods takes the protein sequences as the input and predicts the binding residues, i.e., amino acids in the protein sequence that interact with a given ligand. The second class of predictors takes the protein tertiary (three-dimensional) structure as the input and generates the coordinates of the binding sites, i.e., location of the amino acids that interact with a given ligand. The first class includes approaches that aim at the prediction of metal-binding residues (Horst and Samudrala, 2010), ATP-binding residues (Chauhan et al., 2009), GTP-binding residues (Chauhan et al., 2010), and carbohydrate-binding residues (Chou et al., 2010). The second class of methods could be further subdivided into three groups, approaches based on

geometrical analysis, calculation of energy of protein-ligand binding, and threading using structural templates. This class includes the geometry-based SURFNET (Laskowski, 1995), PocketFinder (Hendlich et al., 1997), PASS (Brady and Stouten, 2000), LIGSITE^{csc} (Huang and Schroeder, 2006), PocketPicker (Weisel et al., 2007), ConCavity (Capra et al., 2009), and Fpocket (Le et al., 2009), the energy-based Q-SiteFinder (Laurie and Jackson, 2005), and the threading-based Findsite (Skolnick and Brylinski, 2008). Additionally, MetaPocket (Huang, 2009) is based on a consensus of the geometry- and energy-based approaches. Although our recent survey shows that the quality of the predicted binding sites of small organic compounds was significantly improved during the last decade (Chen et al., 2011), the success rates for certain ligand types are relatively low and there is a need and plenty of room for further improvements.

1.2 Outline, thesis statements, and contributions

This dissertation is focused on computational analysis and prediction of protein-small ligand interactions. Since the small ligands that are known to interact with proteins, i.e., which are deposited in complexes with proteins in PDB, are diverse and thus are likely to interact with proteins in different ways, we first *investigate whether there are generic interaction patterns that occur across different protein-small ligand interactions*. Our study demonstrates that the interactions between proteins and different small ligand types are dissimilar and no generic pattern is observed across all of these interactions. However, we note that some interaction patterns occur frequently for certain ligand types. This result implies that although it would be challenging (or virtually impossible) to build a well-performing generic model that predicts all types of protein-small ligand interactions, it should be feasible to develop a solution that accurately predicts a certain type of protein-small ligand interactions. Before we embark on the development of new predictive methodologies, we *analyze the state of the art of existing computational methods for the prediction of the protein-small ligand interactions*

to investigate whether there is a need for development of novel methods. To this end, we perform an extensive comparative survey of existing methods that predict interactions with small organic compounds. The protein-small organic compound interactions attract more attention when compared with the interactions with other small ligands, like protein-metal ion, protein-inorganic anion, and protein-inorganic cluster interactions. Our survey indicates that new and improved methods are needed and shows that the predictive performance of the top performing methods varies substantially between different types of the organic compounds. Our survey also reveals that consensus predictions lead to improvements and that certain approaches generate predictions with favorable quality. Motivated by the conclusions of the survey and our earlier result, which demonstrates that predictions should be performed for a specific type of small ligands, we develop new solutions to tackle prediction of protein-nucleotide[1] interactions. We focus on the nucleotides since they are highly abundant and ubiquitous and since the knowledge of the protein-nucleotide interactions is crucial for the protein function annotation, elucidation of the mechanism of cellular activities, and molecular docking-based rational drug discovery (Barrell et al., 2009; Goto et al., 2002). Consequently, our third hypothesis is *whether a novel strategy for the improved prediction of protein-nucleotide interactions, when compared with the existing solutions, can be built.*

The major contributions of this work include:

1. We show that the interactions between proteins and different small ligand groups are governed by many different patterns. Therefore, there is no common interaction pattern that occurs in all protein-small ligand complexes. We systematically categorize the protein-small ligand interactions and we discovered ten patterns that cover significant majority of the protein-small ligand complexes in PDB.

[1] nucleotides are one of the types of the small organic ligands

2. We perform an extensive comparative analysis of the predictive performance of ten representative predictors of the protein-small ligand interactions. We found that the recent binding site predictors significantly improve predictions over the older solutions. However, the predictive quality of the existing methods is not satisfactory for certain small ligand types. We also derive a couple of observations that are useful for the development of new predictors, which we utilize to develop our solution.

3. We develop five accurate sequence-based predictors that identify binding residues for the five most abundant nucleotides in the PDB, including ATP, ADP, AMP, GTP, and GDP. Our predictors are shown to significantly improve over the existing sequence-based methods.

4. We develop a novel predictor that identifies nucleotide-binding sites from protein structures. Our method is empirically shown to significantly improve predictions over the existing structure-based predictors, including the geometry-, energy-, and threading-based approaches.

1.3 Organization of the thesis

The remainder of this dissertation is presented as follows. Chapter 2 includes background necessary to understand the issues discussed in the remaining chapters. Both the biological background concerning proteins and protein-ligand interactions and the background concerning computational methods that are used in this thesis are discussed. Chapter 3 describes goals and provides a detailed outline of the remainder of the thesis. Chapter 4 proposes several interaction patterns that are crucial for protein-small ligand interactions. Chapter 5 surveys the existing binding site predictors. Chapter 6 presents the sequence-based method that predicts protein-nucleotide interactions. Chapter 7 introduces the structure-based algorithm that identifies protein-nucleotide interactions. The dissertation is summarized in Chapter 8.

CHAPTER 2 Background

The following sections include background information necessary to understand the issues discussed in the subsequent chapters. Section 2.1 provides background concerning proteins, ligands, and their interactions. Section 2.2 describes the background concerning computational methods, including classification, clustering, and feature representation and selection. Section 2.3 introduces the preparation of datasets and evaluation protocols.

2.1 Background on proteins, ligands, and their interactions

2.1.1 Amino acids

Amino acids (AAs) are molecules that contain an amine group ($-NH_2$), a carboxylic acid group (-COOH) and a side-chain (R) that varies between different AAs. AAs are linked together by peptide bonds, which are formed between the amine group of one AA and the carboxylic acid group of the adjacent AA. Figure 2-1 shows the structure of AA and Figure 2-2 demonstrates the formation of a peptide bond between two consecutive AAs.

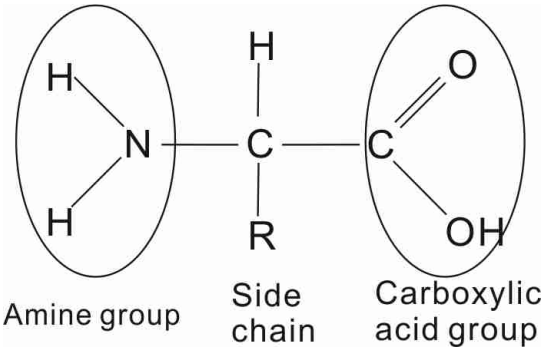


Figure 2-1: Atomic structure of amino acids.

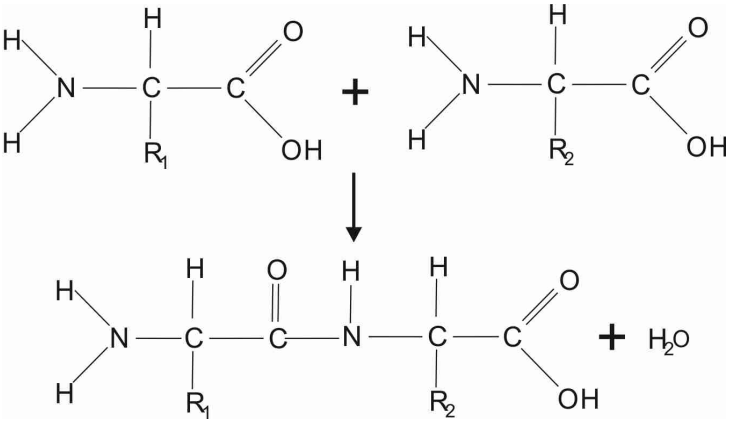


Figure 2-2: Formation of a peptide bond between two amino acids.

There are 20 different R groups and each corresponds to one AA type. The list of the 20 AA types is given in Table 2-1.

Table 2-1: The names of the 20 amino acid (AA) types, including their 3-letter and 1-letter encoding.

Name of AA	3-letter encoding	1-letter encoding	Name of AA	3-letter encoding	1-letter encoding
Alanine	ALA	A	Leucine	LEU	L
Arginine	ARG	R	Lysine	LYS	K
Asparagine	ASN	N	Methionine	MET	M
Aspartic acid	ASP	D	Phenylalanine	PHE	F
Cysteine	CYS	C	Proline	PRO	P
Glutamic acid	GLU	E	Serine	SER	S
Glutamine	GLN	Q	Threonine	THR	T
Glycine	GLY	G	Tryptophan	TRP	W
Histidine	HIS	H	Tyrosine	TYR	Y
Isoleucine	ILE	I	Valine	VAL	V

AAs are linked together by peptide bonds and form a polypeptide chain. When a protein is translated from messenger RNA, it is created from amine-terminus (N-terminus) to carboxyl-terminus (C-terminus). One or several polypeptide chains make up a protein. For instance, the ubiquitin, a regulator protein that is responsible for degrading and recycling of unwanted proteins, contains a single protein chain while hemoglobin, a protein that is responsible for the transportation

of oxygen, consists of four chains. An example polypeptide chain is given in Figure 2-3.

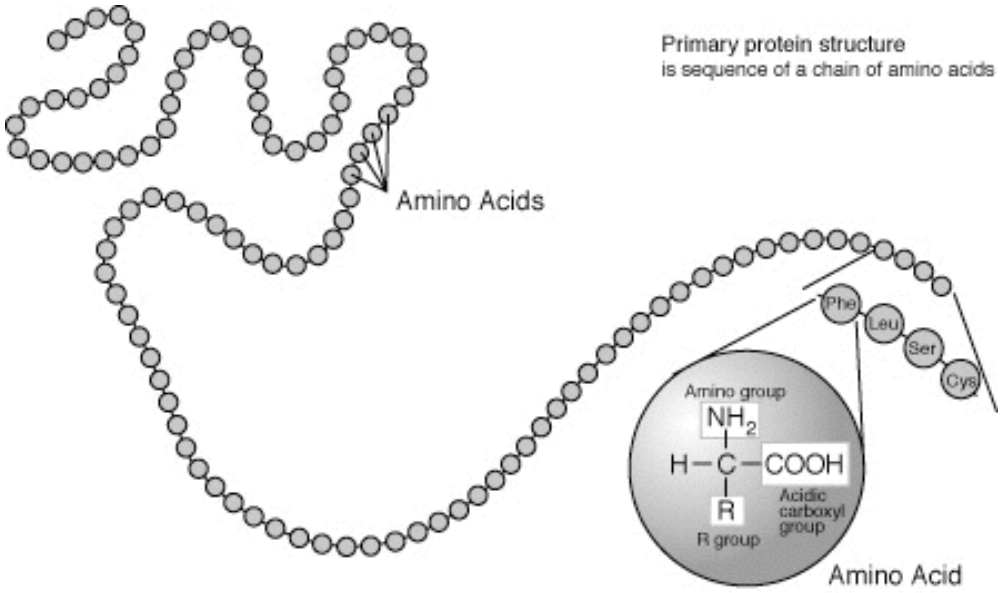


Figure 2-3: A polypeptide chain (adapted from Wikipedia public domain image resource) (Wikimedia Foundation, 2006).

2.1.2 Protein Structure Hierarchy

The protein structure is categorized into four levels: primary, secondary, tertiary, and quaternary (Linderstrøm-Lang, 1952).

- *Primary structure* is the basic level of the hierarchy and is defined as the linear sequence of amino acids that comprise one polypeptide chain.

- *Secondary structure* is the regular folding of spatially local regions within one polypeptide chain into particular, structural patterns. Alpha helix and beta sheet/strand are the two common types of the secondary structure elements. The remaining portions in the chain which are not classified to helix and strand are called coils. The secondary structures are usually maintained by hydrogen bonds between the carbonyl oxygen and the amide hydrogen of the peptide bond.

- *Tertiary structure* is a particular three-dimensional arrangement of all the amino acids in one polypeptide chain. This structure is usually the native and active conformation and is held together by multiple non-covalent interactions. The tertiary structure of a protein is represented by the coordinates of all atoms of the protein.

- *Quaternary structure* is the particular spatial arrangement, and interactions, between two or more polypeptide chains, which are called subunits. The quaternary structure complex could involve up to 60 subunits, e.g., in the viral capsids.

An example of the primary, secondary, tertiary and quaternary structures of tubulin protein is given in Figure 2-4.

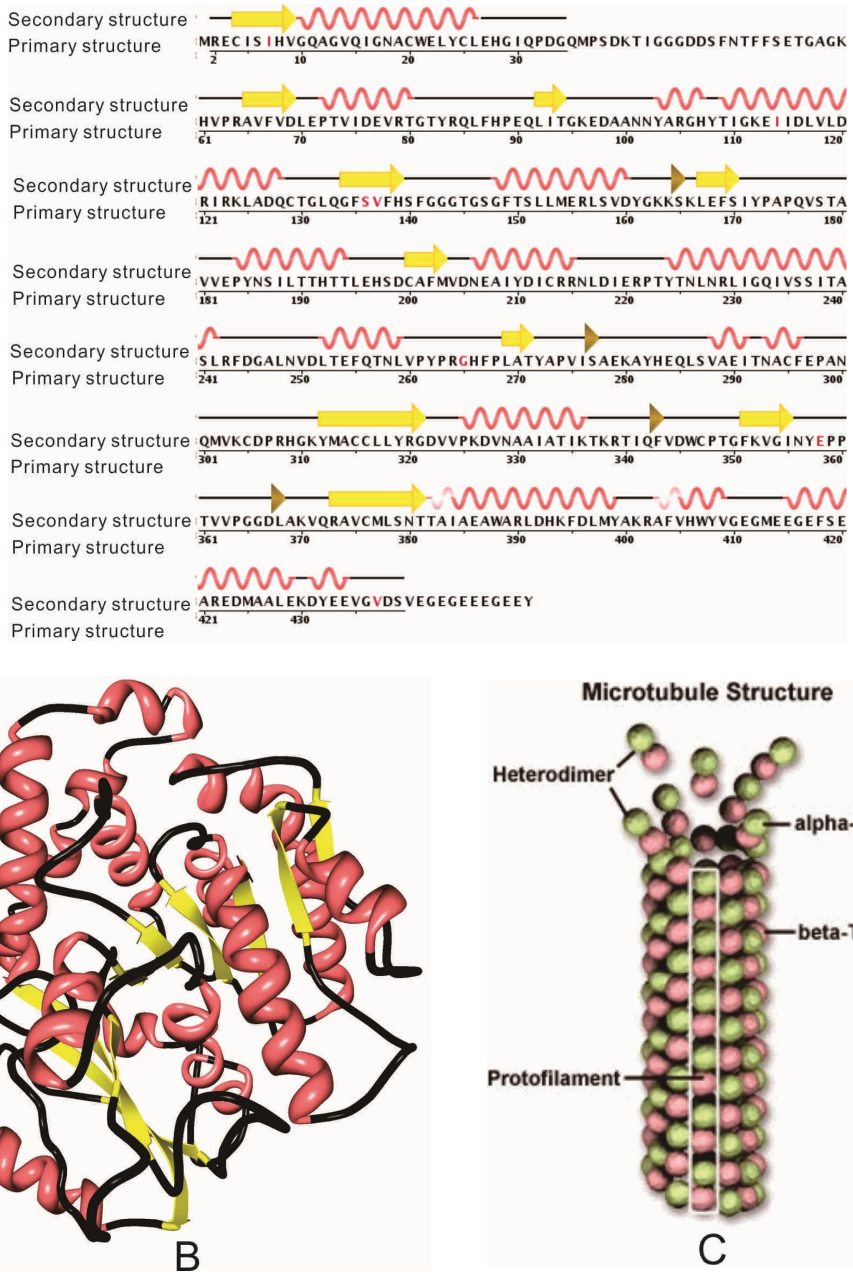


Figure 2-4: Protein structure hierarchy. A) The primary and secondary structures of α-tubulin protein, which is pivotal for maintaining the cell structure; B) The tertiary structure of α-tubulin protein; C) The quaternary structure of microtubule which is formed by α and β-tubulin proteins. On panels A and B, helix is colored red (dark grey), strand is colored yellow (light grey), coiled turn is colored purple and the remaining coils are colored black. On panel C, green (light grey) and red (dark grey) balls stand for α and β-tubulin proteins respectively.

2.1.3 Ligands

In biochemistry, a ligand refers to a substance that forms a complex with a biomolecule where the formation of the complex has an impact on certain biological processes. Though some compounds, e.g., glycol, frequently occur in PDB structure files, they are usually introduced by the procedures during protein expression, purification, and crystallization and are not relevant to biological processes. These biologically-irrelevant ligands are excluded from discussion in this dissertation. The biological-relevant ligands, which occur in PDB, mainly include metal ions, nucleotides, and carbohydrates.

- *Metal ion*s are essential mineral nutrients and are present in every cell type in every organism. They play pivotal roles in cellular signaling, enzyme activation, and catalysis (Que et al., 2008). It was estimated that metal ions are involved in the activities of one third of the enzymes (Silva and Williams, 1991).

- *Nucleotides* are multifunctional molecules that are essential for numerous biological processes. These molecules are structural units of nucleic acid chains (DNA and RNA), and they serve as sources for chemical energy, participate in the cellular signaling, and they are involved in the enzymatic reactions (Fields and Burnstock, 2006; Rich, 2003; Fredholm, 1994).

- *Carbohydrates* are organic compounds with the $C_m(H_2O)_n$ formula. They are components of coenzymes and the backbone of the DNA and RNA. Carbohydrates are essential for the storage of energy and play key roles in the immune system, fertilization, prevention of pathogenesis, blood clotting, and development of organisms (Maton et al., 1993).

This dissertation is focused on protein-small organic compound interactions with special emphasis on protein-nucleotide interactions. Therefore, the structure and functions of nucleotides are discussed in detail.

2.1.4 Structure and biological functions of nucleotides

A nucleotide is composed of a nucleobase, a five-carbon sugar, and between one and three phosphate groups. The phosphate groups form bonds with the 2, 3, or 5-carbon of the sugar, with the 5-carbon site being the most common. The structures of the common nucleotides are summarized in Figure 2-5.

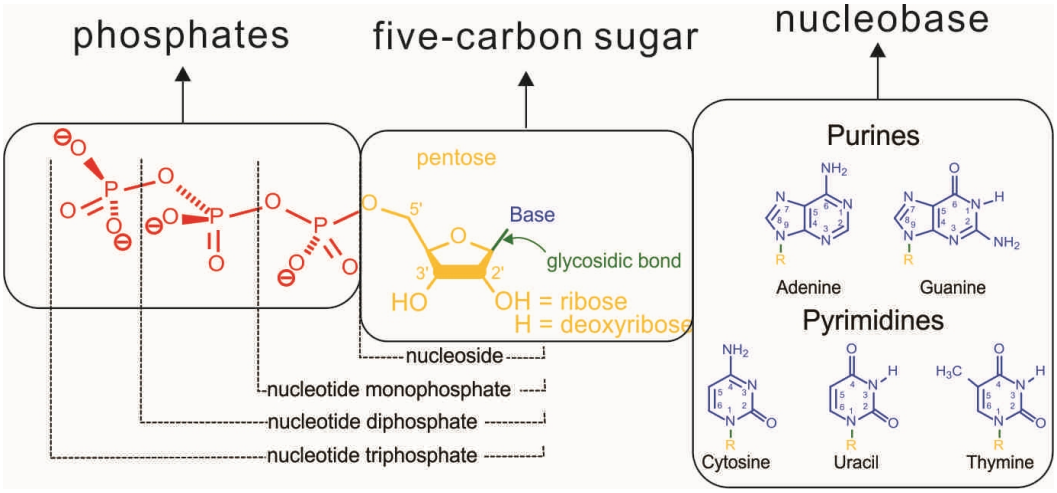


Figure 2-5: Structure elements of nucleotides (adapted from Wikipedia public domain image resource) (Wikimedia Foundation, 2006).

When the phosphate group is bound to two of the sugar's hydroxyl groups, a cyclic nucleotide is formed. Figure 2-6 presents the structure of adenosine monophosphate (panel A) and the structure of cyclic adenosine monophosphate (panel B).

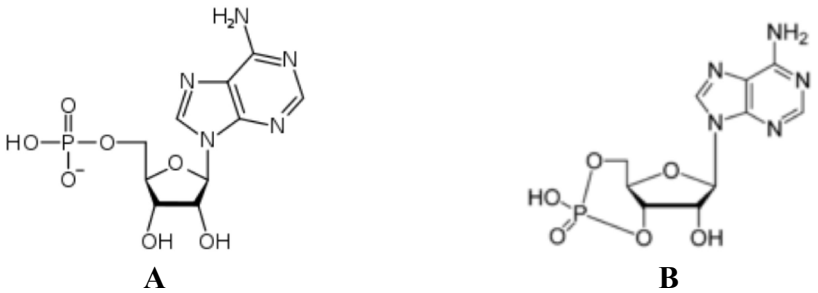


Figure 2-6: Structures of adenosine monophosphate (panel A) and cyclic adenosine monophosphate (panel B) (adapted from Wikipedia public domain image resource) (Wikimedia Foundation, 2006).

The biological functions of nucleotides mainly include:

- Production of energy for majority of the cellular activities. The supply of energy for cellular process involves the conversion between Adenosine triphosphate (ATP) and Adenosine diphosphate (ADP) or the conversion between Guanosine triphosphate (GTP) and Guanosine diphosphate (GDP). Plants use photosynthetic pathways to convert and store the energy from sunlight, via conversion of ADP and GDP to ATP and GTP (Rich, 2003). Animals use the energy released in the breakdown of glucose and other molecules to convert ADP and GDP to ATP and GTP, which can then be used to fuel necessary growth and cell maintenance (Rich, 2003).

- Maintenance of cell structure by facilitation of assembly and disassembly of certain elements of the cytoskeleton. For instance, ATP is required for the shortening of actin and myosin filament cross-bridges required for muscle contraction, which is essential for locomotion and respiration (Bárány et al., 2001). On the other hand, the binding of GTP prevents the depolymerization of microtubule (Mitchison and Kirschner, 1984).

- Signaling. ATP, ADP, GTP, cyclic AMP, cyclic GMP and some other nucleotides are involved in a variety of extracellular signaling processes and signal transduction, a process by which an extracellular signaling molecule activates a membrane receptor that in turn alters intracellular molecules creating a response (Fields and Burnstock, 2006; Fredholm, 1994).

- Building of nucleic acids. Nucleotides are also structure bases for the synthesis of DNA and RNA, the generic materials that define the development and functioning of all known living organisms.

2.1.5 Interactions between proteins and ligands

Proteins interact with ligands by the means of forming a complex where the ligand is attached to a certain patch on the protein surface. This patch is often

called a binding site or a binding pocket, and the amino acids in the protein that form bonds with the ligand (which interact with the ligand) are called binding residues. Consequently, prediction of protein-ligand interactions boils down to the prediction of the binding residues or binding pockets. The protein-ligand interactions are established through covalent, coordination, and hydrogen bonds, and the electrostatic and Van der Waals forces.

Covalent bond is a form of chemical bonding that is characterized by the sharing of pairs of electrons between atoms. The structure of the macromolecules, i.e., proteins and nucleic acids, are maintained by covalent bonds. Several types of covalent interactions, i.e., thioether bond and disulfide bond, are observed between proteins and ligands. The interaction between a non-hydrogen atom A_1 of a residue (a residue is an amino acid in a protein) and a non-hydrogen atom A_2 of a ligand is defined as the covalent bond if the residue and the ligand do not have the opposite charge that would result in electrostatic force of attraction and the distance d of these two atoms satisfies $d < \text{radius}(A_1) + \text{radius}(A_2) + 0.5\text{Å}$, where $\text{radius}(A_i)$ represents the radius of A_i. As discussed by Davis and colleagues (Davis et al., 2003), in a typical 3Å resolution structure, the uncertainty of the position of the individual atoms can be at 0.5Å or more. The marginal 0.5Å value used in the formula accommodates for the uncertainty of the positions of both atoms and for the variation of the length of covalent bonds, i.e., the length of a single bond between carbon atoms ranges between 1.2Å to 1.54Å.

Coordination bond is a kind of 2-centre, 2-electron covalent bond in which the two electrons derive from the same atom. The coordination bonds are frequently observed among protein-metal ion interactions. Metal ions usually do not contain electrons in their outer shell. Therefore, if a metal ion forms the covalent bond with another atom, the pair of electrons shared by the metal ion and the second atom should be provided by the other atom. The corresponding covalent bond is defined as the coordination bond.

Hydrogen bond is the attractive interaction of a hydrogen atom with an electronegative atom, such as nitrogen, oxygen or fluorine, which belongs to another molecule or chemical group. Hydrogen bond plays a crucial role in the formation and maintenance of protein secondary structures and the double helix structure of DNA. Hydrogen bond is also pivotal in protein-organic compound interactions. The hydrogen bonds were calculated with HBPLUS program (McDonald and Thornton, 1994). To identify hydrogen bonds, this program finds all proximal donor (D) and acceptor (A) atom pairs that satisfy specified geometrical criteria for the formation of the bond. Theoretical hydrogen atom (H) positions of both protein and ligand are calculated with REDUCE program (Word et al., 1999). Following the criteria used in a previous study by Luscombe et al. (2001), hydrogen bond is established if H–A distance < 2.7Å, D–A distance < 3.5Å, D–H–A angle > 90° and H–A–AA angle > 90°, where AA is the atom attached to the acceptor.

Electrostatic force represents the interaction between electrically charged particles. The strength of the interaction is calculated by the Coulomb's law. Among the 20 AAs, the electrostatic force concerns positively charged Arg, His, and Lys residues and negatively charged Asp and Glu residues. The charge of the ligand is annotated using PDB dictionary located at http://deposit.rcsb.org/public-component-erf.cif, which provides the charge of each atom of the ligand. An atom of the ligand and an AA in the protein are considered to exert electrostatic force with each other if they have opposite charges and at least one non-hydrogen atom of the AA is less than 3.5Å away from the charged atom of the ligand.

Van der Waals force is the sum of the attractive or repulsive forces between molecules (or between parts of the same molecule) other than those due to covalent bonds or to the electrostatic interaction of ions with one another or with neutral molecules. The Van der Waals force is relatively weak compared to covalent bonds or electrostatic interactions. However, Van der Waals interactions are the dominant forces in many protein-protein and protein-organic compound interactions. Following the definition by Ma and colleagues (Ma et al., 2003), A

non-hydrogen atom A_1 of a protein and a non-hydrogen atom A_2 of a ligand form van der Waals contact if the distance d between these two atoms satisfies $d <$ vdW(A_1) + vdW(A_2) + 0.5Å, where vdW(A_i) is the van der Waals radius of A_i and where these two atoms do not form covalent bond, coordination bond, hydrogen bond, and electrostatic force. Similar as for the covalent bond, the 0.5Å is used to accommodate for the uncertainty in the position of the atoms.

2.2 Background on computational methods

2.2.1 Classification and clustering

Computational analysis and prediction of protein-small ligand interactions involves the classification, clustering, and feature selection using protein datasets. In machine learning, *classification* refers to the task of assigning a set of instances in a given dataset into a (small) number of categories. The method or algorithm that performs the classification task is called classifier. To perform a classification task, a training set and a test set should be prepared. Each instance in the training set is represented by a feature vector that describes the characteristic of the instance and a label that indicates the category it belongs to. The instances of the test set contain only the feature vectors and the corresponding categories/labels are hidden. A classifier learns from the training set with instances that have been properly labeled and assigns/predicts labels for the instances from the test set. A number of classification algorithms have been developed in the past several decades. We use support vector machine (SVM) since this method provides accurate predictive models and is included among the top 10 data mining algorithms (Wu et al., 2007). The SVMs are introduced in (Cortes and Vapnik, 1995) and here we briefly introduce the main concepts behind this methodology. We define a training set $D=\{(x_i, y_i)|\ i=1,2,\ldots,n\}$, where x_i is the feature vector of i^{th} instance and y_i is either 1 or -1, indicating the class (category) of the i^{th} instance. The SVM classifier intends to find the maximum-margin hyperplane that separates the instances with $y_i=1$ (positive instances) from those with $y_i=-1$ (negative instances). A hyperplane could be written in the $w \cdot x - b = 0$ form,

where w is a normal vector. We use two parallel hyperplanes $w \cdot x - b = 1$ and $w \cdot x - b = -1$ to separate the instances. The distance between the two hyperplanes is $\frac{2}{\| w \|}$. Therefore, the objective is to minimize $\| w \|$ (in order to maximize the size of the margin) given the following constraints: the positive instances should satisfy $w \cdot x_i - b \geq 1$ and the negative instances should satisfy $w \cdot x_i - b \leq -1$. The corresponding optimization problem is:

$$\text{minimize} \quad \| w \|$$

$$\text{subject to} \quad y_i (w \cdot x_i - b) \geq 1 \quad \text{for } i=1,2,\ldots,n$$

However, if no hyperplane that separates the positive and negative instances can be found, the *Soft Margin* method, which chooses a hyperplane that separates the instances with as few violations of the constrains (misclassifications) as possible, is used. The soft margin allows for misclassification of instances but it still maximizes the margin between the "cleanly" split instances. The objective function is augmented by a function which penalizes the misclassification with non-zero constants ξ_i, and the optimization becomes a trade off between a large margin and a small error penalty. Given a linear penalty function, the optimization problem becomes:

$$\min_{w,\xi} \left\{ \frac{1}{2} \| w \| + C \sum_{i=1}^{n} \xi_i \right\}$$

$$\text{subject to} \quad y_i (w \cdot x_i - b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad \text{for } i=1,2,\ldots,n$$

The abovementioned linear hyperplane may not be able to cleanly separate positive and negative instance, while a transformation of the data into a new space could lead to a better separation. To this end, the kernel extension was proposed. By using kernel functions, the original data are mapped through non-linear function into a feature space where the instances are potentially linearly separable.

In this dissertation we apply two popular kernel functions, the polynomial and radial basis function (RBF) kernels. These two types of kernels are defined as

$$k(x_i, x_j) = (x_i \cdot x_j)^d$$
$$k(x_i, x_j) = e^{-\gamma \|x_i - x_j\|}$$

where d is the degree of the polynomial and γ is the spread of the Gaussian function. The setup (selection) of the values of these two parameters is discussed in subsequent chapters. In this dissertation, the SVM classifier is used for the sequence-based prediction of the nucleotide-binding residues where the instances (amino acids) are labeled as "binding" or "non-binding". Therefore, we do discuss the multi-class (for more than 2 labels) extension of the SVM.

Clustering is the assignment of a set of instances into subsets (called *clusters*) so that instances in the same cluster are similar. The clustering methods include the hierarchal algorithms, the partitional algorithms, and the subspace clustering algorithms. In this dissertation, we use the hierarchal clustering algorithm for the structure-based prediction of the nucleotide-binding sites. We use the agglomerative algorithm that begins with each instance as a separate cluster and merge them into successively larger clusters. The agglomerative algorithm is given as follows:

STEP 1. Each instance is initialized as a separate cluster.

STEP 2. Calculate distances between any pair of the clusters. The distance between cluster A and B is defined as the minimum distance between any instance of cluster A and any instance of cluster B.

STEP 3. Find the minimum distance d among the distances between the clusters. If $d \leqslant D$, where D is the cutoff distance to terminate the clustering procedure, merge the two clusters with distance d. Repeat the STEP 2 and 3 until the minimum distance d is above the cutoff distance D.

2.2.2 Feature representation and feature selection

Instances must be represented by feature vectors with the same number of dimensions when performing classification or clustering. To this end, the observed data, e,g,, the amino acids in a protein sequence, are converted into feature vectors of a fixed size. The feature representation that is used in this dissertation includes three types of features: features calculated from the primary sequence, features generated from the multiple sequence alignment, and features based on predicted structural descriptors, such as the predicted secondary structure.

The complete feature representation includes thousands of features and requires a tremendous amount of computations for classification and clustering. The exclusion of irreverent and redundant features not only reduces the computations but may also improve the predictive quality. The feature selection algorithms typically fall into two categories: feature ranking and subset selection. The methods based on feature ranking rank the features using a given metric and eliminate all features that do not achieve an adequate score. The subset selection methods search over the entire feature set for a subset that results in an "optimal" predictive quality. The subset selection methods iteratively evaluate a candidate subset of features (by removing features or adding additional features to the subset and evaluating whether the new subset results in an improvement of the predictive quality over the old subset). In this dissertation, we use a hybrid feature selection method that combines feature ranking and subset selection. Details are given in the subsequent chapters.

2.2.3 Architecture of the prediction pipeline

The classification and clustering of protein datasets include three steps, see Figure 2-7. First, the protein sequences/structures are converted into vectors of a fixed length; this step is named *feature representation*. Next, the irreverent and redundant features are removed from the feature set, and this step is called *feature selection*. Lastly, the selected features are fed into a classifier or a clustering

method and the outputs are generated. The feature selection step is optional and may not be implemented in some studies.

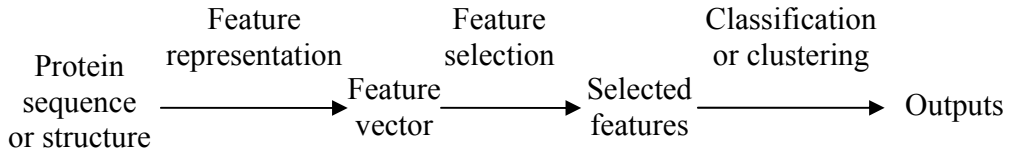


Figure 2-7: Architecture of classification and clustering of protein datasets.

2.2.4 Docking scoring function

Docking scoring functions are fast approximate mathematical methods used to estimate the strength of the non-covalent interaction between two molecules. In this work, docking scoring functions are used to estimate the binding affinity, which measures the strength of the interaction, between proteins and nucleotides. The calculated binding affinity is used to rank the predicted binding sites. There are three classes of scoring functions: the force field-based, empirical, and knowledge-based scoring functions. We utilize the AMBER force field (Cornell et al., 1995), a classical force field-based scoring function, which is used by a number of docking programs for energy calculation, i.e., AutoDock (Morris et al., 1998). The functional form of the AMBER force field is

$$V(r^N) = \sum_{bonds} \frac{1}{2} k_b (l - l_0)^2 + \sum_{angles} \frac{1}{2} k_a (\theta - \theta_0)^2 + \sum_{torsions} \frac{1}{2} V_n [(1 + \cos(n\omega - \gamma)] + \sum_{i<j} \left[\frac{A_{ij}}{(r_{ij})^{12}} - \frac{B_{ij}}{(r_{ij})^{6}} + \frac{q_i q_j}{\varepsilon r_{ij}} \right]$$

where the first term (sum over bonds) represents the energy between covalently bonded atoms, the second term (sum over angles) represents the energy due to the geometry of electron orbital involved in covalent bonding, the third term (sum over torsions) represents the energy for twisting a bond due to bond order and neighboring bonds or lone pairs of electrons, and the fourth term (sum over *i* and *j*)

represents the non-bonded energy between all atom pairs, which can be decomposed into the van der Waals and electrostatic energies. In our case, when using the AMBER force field, we assume that both the protein and the ligand structures are rigid. Therefore, the first three terms in the AMBER function do not change for a given protein-ligand pair and are omitted in the calculation. In the fourth term, i and j represent the number of atoms of protein and ligand respectively; r_{ij} represents the distance between the i^{th} atom of the protein and the j^{th} atom of the ligand; q_i and q_j represent the charge of the pair of atoms; and A_{ij}, B_{ij}, and ε are constants which are parameterized by AMBER GAFF.

2.2.5 Sequence alignment and BLAST

Sequence alignment is an arrangement of DNA, RNA, or protein sequences to identify similar regions (between a query chain and a given database of chains) that may be a consequence of functional, structural, or evolutionary relationships between these sequences. Aligned sequences are typically represented as rows of characters and gaps are inserted among the sequences so that identical or similar characters are aligned in successive columns. An example of multiple sequence alignment of protein sequences is given in Figure 2-8.

```
Query          … K R L E H G G G V A Y A I A K A C A G D A G L …
YP_002995377   … K Y L E H G G G V A Y A I A K A A S G D V R E …
YP_002958591   … K Y L E H G G G V A Y A I A K A A A G N V A E …
YP_003418650   … S Y L Q H G G G V A Y A I V K K G G - - - - - …
YP_002828572   … S Y L Q H G G G V A Y A I V K K G G - - - - - …
ZP_04861702    … G M L K H V G G V A A A I V K K G G - - - - - …
ZP_05391340    … G A L K H G G G A A A A I V K A G G - - - - - …
YP_003345806   … E Y L K H G G G V A G A I V R A G G - - - - - …
YP_003496764   … S H L K M G G G V A G A I R R A G G - - - - - …
```

Figure 2-8: An example of multiple sequence alignment. The first row shows the query chain and the subsequent rows show the eight aligned proteins. Each row contains the protein sequence ID (the first column) and the corresponding amino acid sequence (the third and subsequent columns), where "…" denotes continuation of the chain and "–" denotes a gap, which means that this part of the sequence could not be aligned.

BLAST is a package of programs that perform sequence alignment tasks (Altschul et al., 1997). Two programs in the BLAST package were used in this dissertation. First, the *psiblast* program was used to identify similar protein sequences from a given sequence database for a given query sequence. The *psiblast* program returns the *p*-value and sequence identity (which quantify similarity between the query and a given aligned sequence), and the alignment between the query sequence and the identified similar sequence form the sequence database. Second, the *blastpgp* program was used to generate the Position Specific Score Matrix (PSSM) for a given protein sequence. The program first identifies a list of related (similar) protein sequences from a given sequence database. Next, these sequences are combined into a general "profile", which summarizes similarity (and dissimilarity) across these sequences. Second, another query against the sequence database is run using the generated profile, and a larger group of sequences is found. This larger group of sequences is used to construct another profile, and the process is repeated. The PSSM profile was found useful for building various methods to predict protein structure, including prediction of the protein secondary structure (McGuffin *et al*., 2000) and solvent accessibility (Faraggi *et al*., 2009); the latter quantifies the ratio of the surface area of a given residue that is accessible to solvent. In this dissertation, the PSSM profile is used for the sequence-based prediction of the nucleotide-binding residues. An example of the PSSM profile is given in Figure 2-9.

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
									…	…	…	…	…	…							
									…	…	…	…	…	…							
1	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-3	5	-2	-3	-1	0	-1	-3	-2	-3
2	R	-2	2	-2	-3	-3	-1	-2	-3	1	-2	-1	0	-1	2	-3	-2	-2	2	7	-2
3	L	-2	-2	-4	-4	-1	-2	-3	-4	-3	2	4	-3	2	0	-3	-3	-1	-2	-1	1
4	E	-1	0	0	2	-4	2	5	-2	0	-4	-3	1	-2	-4	-1	0	-1	-3	-2	-3
5	H	-2	0	1	-1	-3	0	0	-2	8	-4	-3	-1	-2	-1	-2	-1	-2	-3	2	-3
6	G	0	-3	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-3	-3	-4
7	G	0	-3	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-3	-3	-4
8	G	0	-3	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-3	-3	-4
9	V	0	-3	-3	-4	-1	-2	-3	-4	-3	3	1	-3	1	-1	-3	-2	0	-3	-1	4
10	A	4	-2	-2	-2	0	-1	-1	0	-2	-1	-2	-1	-1	-2	-1	1	0	-3	-2	0
11	Y	-2	-2	-3	-4	-2	-2	-2	-4	1	0	1	-2	0	3	-3	-2	-2	2	6	-1
12	A	4	-2	-2	-2	0	-1	-1	0	-2	-1	-2	-1	-1	-2	-1	1	0	-3	-2	0
13	I	-1	-3	-4	-3	-1	-3	-4	-4	-4	5	2	-3	1	0	-3	-3	-1	-3	-1	3
14	A	4	-2	-2	-2	0	-1	-1	0	-2	-1	-2	-1	-1	-2	-1	1	0	-3	-2	0
15	K	-1	2	0	-1	-3	1	1	-2	-1	-3	-3	5	-2	-3	-1	0	-1	-3	-2	-3
16	A	4	-2	-2	-2	0	-1	-1	0	-2	-1	-2	-1	-1	-2	-1	1	0	-3	-2	0
17	C	2	-3	-3	-3	9	-2	-3	-2	-3	-1	-1	-2	-1	-3	-2	0	-1	-3	-2	-1
18	A	4	-1	-1	-1	-1	-1	-1	0	-2	-2	-2	-1	-1	-3	-1	2	0	-3	-2	-1
19	G	0	-3	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-3	-3	-4
20	D	-2	-2	1	6	-4	0	2	-1	-1	-3	-4	-1	-3	-4	-2	0	-1	-5	-3	-4
21	A	3	-2	-2	-2	-1	-1	-1	-1	-2	0	-1	-1	-1	-2	4	0	0	-3	-2	1
22	G	0	3	0	-1	-3	3	0	2	-1	-3	-3	1	-2	-3	-2	1	0	-3	-2	-3
23	L	-1	0	-1	0	-3	1	4	-2	-1	-1	1	2	0	-2	-2	-1	-1	-3	-2	-1
									…	…	…	…	…	…							
									…	…	…	…	…	…							

Figure 2-9: An example of the PSSM profile generated by the blastpgp program. The first and second columns are the residue number and type, respectively, in the input protein chain. The subsequent columns provide values of the multiple sequence alignment profile for a substitution to an amino acid type indicated in the first row. Initially, a matrix $\{p_{i,j}\}$, where $p_{i,j}$ indicates the probability that the j^{th} amino acid type (in columns) occurs at i^{th} position in the input chain (in rows), is generated. The position-specific scoring matrix $\{m_{i,j}\}$ is defined as $m_{i,j} = \log(p_{i,j} / b_j)$, where b_j is the background frequency of the j^{th} amino acid type.

2.2.6 Calculation of protein sequence conservation

Protein sequence conservation describes the phenomenon that similar or identical protein sequences are observed across species and in the same organism. Analysis of sequence conservation aids the detection of functionally important residues, i.e.,

residues that are involved in the interactions with ligands (Capra and Singh, 2007). Mapping the conservation information onto a protein tertiary structure helps with visualization of potential functional surfaces, i.e., parts of the protein surface that interact with ligands, of a given protein. The protein sequence conservation is calculated from the multiple sequence alignment profiles. Several formulas and algorithms, including the Shannon entropy, the derivatives of Shannon entropy that incorporate background frequency of amino acids (Capra and Singh, 2007; Wang and Samudrala, 2006; Pei and Grishin, 2001) and the Rate4Site algorithm (Pupko et al., 2002) were proposed for the calculation of the conservation scores. The ConSurf-DB database (Goldenberg et al., 2009) collects the pre-calculated conservation scores by Rate4Site algorithm for all protein sequences in PDB.

2.3 Performance evaluation

This section describes the concepts related to the preparation of datasets and the protocols that are used to evaluate the various methods presented in the following chapters.

2.3.1 Dataset preparation

The analysis and development of methods described in this dissertation require preparation of various datasets of protein sequences and structures. The sequences and structures are taken from the PDB (Berman et al., 2000) and the CulledPDB (Wang and Dunbrack, 2003). The PDB is a world-wide repository for the tertiary structural data of large biological molecules, which include proteins and nucleic acids. By the end of 2010, the PDB has included around 65,000 protein structures. The protein files in PDB include information concerning the protein sequence, the coordinates of all non-hydrogen atoms of a protein, the coordinates of the ligand atoms, the ligand type, and the method used to determine the protein structure, etc. Therefore, the PDB protein files contain sufficient information for the calculation and visualization of protein-ligand interactions and to establish ground truth to build and validate prediction methods. Since some protein families (proteins that share similar function) are overrepresented by PDB and these protein families

may contain hundreds of similar structures, we use CulledPDB to choose a list of representative proteins in PDB. The CulledPDB pre-calculates a number of protein lists according to the desired quality of the structure (quality with which the structure was experimentally determined) and sequence similarity. For instance, a user may download a list of protein chains that share a maximal sequence identity of 25% that have high quality structures (e.g., with the R-factor below 0.25 and a resolution below 2.0 Å). Such list includes proteins that are dissimilar to each other, which means that they uniformly cover the protein space, and which have high quality structures, i.e., for which the ground truth is accurate.

In general, similar protein sequences usually have similar structures and biological functions. The sequence similarity between the training and test datasets has a substantial impact on the predictive quality of a given method (Kurgan and Homaeian, 2006). To this end, similar protein sequences and structures are usually removed from the training and test datasets and the pairwise sequence identity within one dataset is usually reduced to between 25% and 40%. However, we note that in some cases the sequence similarity is insufficient for disclosing the relationship between two chains. For instance, many globins (a family of proteins that are responsible for oxygen transport) have similar tertiary structures and biological functions whereas they have less than 10% sequence similarity. To properly reflect the structural and functional relationships between proteins, the structural similarity between proteins are used. Currently, two databases, the SCOP (structural classification of proteins) database (Andreeva et al., 2008) and CATH Protein Structure Classification databases (Cuff et al., 2009) provide a hierarchical classification of protein domains with known structures. In this dissertation, we use the SCOP database for the annotation of protein family, superfamily, fold, and class, which are the 4 levels of the SCOP hierarchy.

2.3.2 Evaluation protocols

The optimization of a model on the training dataset may result in overfitting into the training data, i.e., a situation where the model memorizes the dataset,

including potential noise and incorrect and inconsistent values. The overfitting on the training data may lead to over-estimation of the performance of a given method on that dataset. Cross-validation is a technique that prevents the overfitting of the training data. Two types of cross-validation are used in this dissertation, the *n*-fold cross-validation and the jackknife test.

In the *n-fold cross-validation*, the original dataset is randomly partitioned into *n* equally-sized subsets. Each of the *n* subsets is used as a test dataset once and the remaining *n*-1 subsets are used as the corresponding training dataset. The cross-validation process is repeated *n* times (the *n* folds). The results from the *n* folds are averaged to produce a single estimate of the predictive quality. The advantages of the *n*-fold cross-validation are that all samples are used for both training and testing and that each sample is used for test exactly once.

Jackknife test (leave-one-out cross-validation) involves using a single sample from the original dataset as the test dataset, and the remaining samples as the training dataset. This is repeated such that each sample in the dataset is used as the test data once. This is the same as the *n*-fold cross-validation with *n* being equal to the number of samples in the original dataset. Leave-one-out cross-validation is usually computationally expensive due to the large amount of computations during the training process, i.e., larger number of bigger training dataset is used, when compared with the *n*-fold cross-validation that utilizes a smaller value of *n*.

2.3.3 Statistical tests

The Wilcoxon signed-rank test (Wilcoxon, 1945) is a non-parametric statistical test that is used to compare two related samples or repeated measurements on a single sample to assess whether their population means differ. This test is used as an alternative to the paired t-test when the considered populations do not follow normal distribution. We used the Shapiro-Wilk test (Shapiro and Wilk, 1965) to verify that our data are not normal. In this dissertation, we use the Wilcoxon signed-rank test to measure significance of the differences between the qualities

of the predictions generated by different methods on the same dataset. Suppose we have collected 2 sets of observations (predictive qualities for two methods), denoted as $\{A_i\}$ and $\{B_i\}$ (i=1,2,…,n). A_i and B_i refer to the first and second observations measured on the same subject (the same dataset), which means that A_i and B_i are paired. We have multiple pairs of the measurements generated over different folds in the n-fold cross-validation test. Let $Z_i = A_i - B_i$ (i=1,2,…,n), the procedures to perform the Wilcoxon signed-rank test are given as follows:

1. Exclude observations with $Z_i = 0$ and let m be the reduced sample size.

2. Sort the absolute values $|Z_1|$, $|Z_2|$,…,$|Z_m|$ in ascending order and assign the rank of 1, 2, …, m to the m absolute values.

3. The Wilcoxon signed-rank statistic W_+ is defined as $W_+ = \sum_{i=1}^{n} \varphi(Z_i) R(|Z_i|)$, where $\varphi(Z_i)$ equals 1 if Z_i >0 and equals 0 otherwise. $R(|Z_i|)$ is the rank of $|Z_i|$.

4. The Wilcoxon signed-rank statistic W_- is defined as $W_- = \sum_{i=1}^{n} \varphi(Z_i) R(|Z_i|)$, where $\varphi(Z_i)$ equals 1 if Z_i <0 and equals 0 otherwise. $R(|Z_i|)$ is the rank of $|Z_i|$.

5. Set S as the smaller of these two rank sums: $S = min(W_+, W_-)$.

6. Find the critical value for the given sample size m and compare S to the critical value, a value that a test statistic must exceed in order for the null hypothesis to be rejected. If S is equal or greater than the critical value, we conclude that $\{A_i\}$ and $\{B_i\}$ differ significantly.

CHAPTER 3 Organization of the thesis

This dissertation addresses computational characterization and prediction of protein-small ligand interactions. As of April 2011, the PDB included 11,865 distinct types of ligands, which are characterized by a diverse atomic composition (different number and different atom types) and a wide range of molecular weights and structures. We systematically analyze interactions between proteins and these 11,865 types of ligands to find whether they are governed by some generic patterns. These patterns summarize the interacting atoms on the corresponding protein and ligand and the corresponding interacting amino acid types on the protein. Initially, we investigate whether there are generic interaction patterns that occur in all protein-small ligand complexes. Since we could not find such generic patterns, we analyze patterns that occur in the interactions with specific types of ligands and for the specific types of bonds. This analysis, which is explained in Chapter 4, demonstrates that the interactions between proteins and small ligands are diverse and can be described only partially using several different patterns. These interaction patterns are specific to interactions with individual ligand types, and many of these interactions are too specific to form a pattern. Our results imply that it would be challenging, if not impossible, to build a well-performing model that can predict all protein-small ligand interactions, which is what current prediction methods attempt to do. However, it should be feasible to build a divide-and-conquer type of solutions that predict certain types of interactions, i.e., interactions with certain types of small ligands, and combine them together to establish a generic model.

To date, more than a dozen methods have been already proposed for the prediction of protein-small ligand interactions. Therefore, before attempting to build a new method, we investigate the predictive quality of the state-of-the-art methods. Such investigation allows us to find out potential weaknesses/strengths of the current methods, which can be then addressed/exploited in our solution. We note that the existing methods concentrate on the prediction of binding sites for

small organic compounds because such compounds are of substantial research and development interest, i.e., they constitute more than 80% of the drugs approved by the U.S. Food and Drug Administration (Wishart et al., 2008). The comparative survey on the existing predictors is given in Chapter 5. The result shows that the predictive quality of the existing methods was significantly improved during the past decade but there is still plenty of room for improvements, particularly when predicting for certain types of organic compounds. We observe that the predictive qualities of the top performing methods are quite different for different organic compound types, and thus we identify the need to develop predictors that address interactions with these specific types. Moreover, we also show that use of a consensus-based predictor results in improvements and that certain types of predictors (i.e., template-based methods) outperform other types of solutions. These conclusions provided us with useful feedback to develop new solutions.

Among the organic compounds, we focus our attention on nucleotides because they are highly abundant and ubiquitous, i.e., they interact with more than 10% of the proteins that are deposited in PDB and they play important roles in a number of biological processes, including metabolism, cellular signaling, maintenance of cell structure, and enzyme activation and catalysis (Barrell et al., 2009; Goto et al., 2002). The high abundance in PDB demonstrates high interest in these ligands. Building of a prediction method for nucleotides is relatively challenging due to the ubiquity of these ligands, i.e., ability to perform so many diverse functions and thus the ability to interact with so many diverse proteins. This means that the protein-nucleotide interactions are complex and would require the divide-and-conquer solution. To this end, we designed two methods that address the prediction of protein-nucleotide interactions from protein sequences and from protein structures, respectively. The sequence-based annotation of the nucleotide-binding residues is motivated by the fact that the tertiary structures of majority of the protein sequences are unknown (see Chapter 1). The biological functions of the protein sequences without the known tertiary structures can be

analyzed/predicted only using sequence-based methods. In Chapter 6, we propose a method that takes protein sequences as the input and predicts the nucleotide-binding residues. This method decomposes the problem by utilizing a combination of models that address prediction of the interactions with specific nucleotides, including ATP, ADP, AMP, GTP and GDP. We empirically show that the proposed method significantly outperforms existing methods and several baseline predictors. We also identify several sequence-derived hallmarks that are characteristic for the protein-nucleotide interactions.

Some proteins with the known tertiary structure lack annotations of (some of) their biological functions. These proteins are of particular interest because they find applications in molecular docking-based rational drug discovery (Meng et al., 2011; Brooijmans and Kuntz, 2003). This motivates the development of computational methods that annotate biological functions, i.e., nucleotide-binding, for proteins with the known structures. In Chapter 7, we propose a method that takes protein structures as the input and predicts nucleotide-binding sites on the protein surface. Based on the conclusions from Chapter 5, we design a new consensus-based approach, in which we perform search for local similarity (within the binding pocket) between the input structure and a library of known protein-nucleotide interactions. We demonstrate that this method significantly outperforms the existing structure-based binding site predictors, including the geometry-, energy-, and threading-based approaches. We also show that our method can accurately, when compared to the current methods, find distant functional relationships between proteins from different families, superfamilies, and folds.

CHAPTER 4 Investigation of atomic level patterns in protein-small ligand interactions

4.1 Introduction

The facts that the protein-small ligand interactions were never systematically studied and that the rules that govern these interactions are not yet fully disclosed motivate the work presented in this chapter. We analyze both covalent bonding (normal covalent bonds and coordination bonds) and non-covalent interactions (electrostatic force, hydrogen bonds, and van der Waals force) between proteins and small ligands (Chen and Kurgan, 2009). Our objective is to discover interaction patterns that describe frequently occurring regularities observed in the binding sites, i.e., favored residue types that interact with a certain atoms on the ligand and the spatial arrangement of the binding residues.

4.2 Related work

Although the protein-small ligand interactions were not systematically studied in the past, the interactions between proteins and macromolecules, i.e., protein-protein (Zhu et al., 2008; Rajamani et al., 2004; Ma et al., 2003; Jones and Thornton, 1996), protein-DNA (Luscombe et al., 2001), and protein-RNA interactions (Ellis et al., 2007), have been systematically investigated. Thornton's study compared the size, shape, residue interface propensities and hydrophobicity of the protein-protein interface for four different types of protein-protein complexes (Jones and Thornton, 1996). Ma's report shows that several structurally conserved residues could be used to distinguish between binding sites and general exposed surface; for instance, conservation of Trp, Phe, and Met residues on the protein surface was shown to be associated with a higher likelihood of formation of a binding site (Ma et al., 2003). Rajamani and colleagues show that the anchor residues in protein-protein interactions maintain similar conformations before and after the binding, which allows for a relatively smooth binding process (Rajamani

et al., 2004). Luscombe *et al.* studied the role of hydrogen bonds, van der Waals contacts, and water mediated bonds in protein-DNA interaction. They concluded that the majority of the amino acid-base interactions follow general principles that apply across all protein-DNA complexes (Luscombe et al., 2001).

There was also some work concerning the protein-small ligand interactions. The principles that govern protein-metal ion interaction were investigated by Dudev and Lim (Dudev and Lim, 2008a). They summarized several rules with respective to the coordination mode, coordination number, metal selectivity and coordination stereochemistry. In another study by Dudev and Lim, various factors governing metal binding affinity and selectivity were systematically analyzed (Dudev and Lim, 2008b). The structure and properties of the metal-binding sites were also discussed for specific metal ions like Ca^{2+} and Zn^{2+} (Gifford et al., 2007; Maret, 2005). However, studies that would consider a wider range of small ligands, including organic compounds, various inorganic ligands, and metal ions are missing.

4.3 Problem definition

We focus on the study of interactions between proteins and small ligands that exclude proteins and nucleic acids which were already investigated by other researchers. *Our aim is to find frequent atomic-level regularities (patterns) that could be used to summarize interactions between the protein and the considered ligands.* The term "atomic-level" refers to the fact that patterns concern interactions between individual atoms of the protein residues and the ligand. We discuss specific details of these interactions across different residue types and ligands, e.g., the number of residues and the residues types that are involved in the coordination bonds with specific metal ions, and we quantify their relative abundance, which can be used to assess their importance in protein-ligand interactions. The protein-ligand interactions are grouped into four categories including protein-organic compound, protein-metal ion, protein-inorganic anion and protein-inorganic cluster interactions.

4.4 Dataset preparation

The protein chains, which are selected using culledPDB list (Wang and Dunbrack, 2003), are characterized by the following criteria: 1) the chains share sequence identity of below 25%; 2) the resolution of the protein-ligand complex structure is below 2.0Å; and 3) the R-factor value is below 0.25. These criteria, which result in selection of 2320 chains, assure that the selected proteins share low sequence identity (they adequately sample the sequence space) and that the corresponding structures have sufficient quality. The protein and a ligand are assumed to interact with each other when at least one pair of non-hydrogen atoms, one from the protein and one from the ligand, can be found within a 3.9Å distance (Luscombe et al., 2001). If the same ligand binds a given protein in multiple pockets, all pocket-ligand complexes are included. Excluding the water molecule, all molecules annotated as "HET" in PDB, which includes organic compounds and ions, are taken as ligands. This excludes protein and nucleic acid chains. As a result, 7759 pockets which have at least one contact with the considered ligand are extracted from the 2320 chains.

Among the 7759 complexes, some of the ligands appear multiple times, some are similar and could be grouped together and the same/similar ligands bind to a variety of pockets. To facilitate analysis of the protein-ligand interactions we select only these ligands that occur frequently and we group them into several categories. The ligands that bind to at least 100 pockets cover 59.4% of the considered complexes. Among these ligands, EDO, NAG, and ACT are organic compounds, Ca^{2+}, Zn^{2+}, Na^{+}, Mg^{2+} and Cd^{2+} are metal ions, and SO_4^{2-}, PO_4^{3-}, Cl^{-}, Br^{-} and I^{-} are inorganic anions. Additionally, some inorganic clusters, i.e., Fe-S cluster, also bind to a relatively large number of pockets. Therefore, the considered ligands (including those that occur in less than 100 pockets) are grouped into four categories: organic compounds, metal ions, inorganic anions, and inorganic clusters. We analyze total of 3685 organic compounds (that include 560 distinct types), 1682 metal ions (25 types), 1837 inorganic anions (19 types),

and 54 inorganic clusters (9 types), which cover (3685+1682+1837+54)/7759 = 93.5% of all extracted pockets.

4.5 The atomic level patterns in protein-ligand interactions

4.5.1 Summary of the interaction patterns

The protein pocket-ligand interactions are summarized in Figure 4-1. The top layer divides the 7759 protein pocket-ligand complexes into 5 categories based on the ligand type. The second layer lists the major forces that are involved in formation of protein-ligand complexes for a given ligand type. For instance, protein-organic compound complexes are formed mainly by the means of covalent bonds, hydrogen bonds, and van der Waals contacts, which accommodate for 99.9% of the interactions. The remaining 0.1% of the contacts between a protein and the organic compound, which are omitted in the Figure 4-1, is based on the electrostatic force. The bottom layer provides significant patterns that are associated with interactions for a given type of the ligand and a given type of bond/force, which are discussed in detail in the following subsections. The patterns are shown in $X^R \ldots Y^L$ or $X^R - Y^L$ format where X denotes an atom type of residue R in the protein, Y denotes an atom type of the ligand L, strong interactions (covalent and coordination bonds) are depicted by "–", and weak interactions (hydrogen bond) are represented by "…".

The forces that are omitted in Figure 4-1 are less significant (less frequent or nonexistent) for a given type of the protein-ligand interaction. Our analysis concentrates on the forces that are characterized by frequently occurring patterns for a given ligand category, while omitting some forces which are listed in Figure 4-1 and for which we could not find strong regularities (patterns). For the protein-organic compound interactions, we focus on the hydrogen and covalent bonds since they exhibit more regular and frequent patterns than the van der Waals contacts. In the case of the protein-metal ion interactions, electrostatic force and coordination bonds, which cover 95% these interactions, are analyzed. The discussion of the protein-inorganic anion interactions concentrates on the

electrostatic force and hydrogen bonds; the van der Waals contacts are omitted due to lack of regular interaction patterns. Finally, our analysis of the protein-inorganic cluster interactions concerns only the coordination bonds since they constitute the main driving force for these interactions, i.e., they are involved in all considered protein-inorganic cluster complexes. Although we investigate all four interaction types, in our analysis we concentrate on the protein-organic compound and the protein-metal interactions since they occupy the largest fraction of the considered protein-ligand complexes and they are important for many biological processes (Zoltowski et al., 2007; Ma et al., 2006).

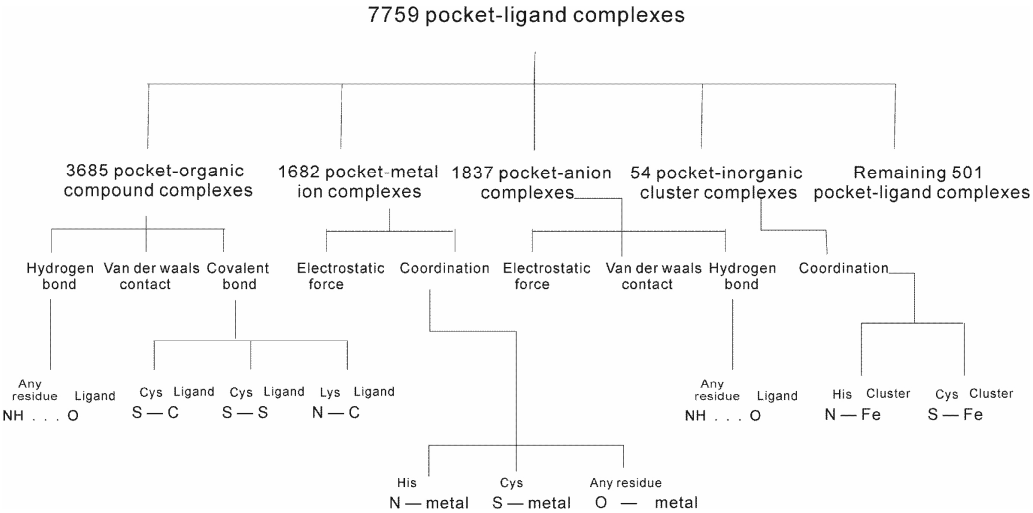


Figure 4-1: An overview of the protein pocket-ligand interactions. The top layer divides protein-ligand complexes into 5 major groups based on the type of the ligand. The second layer shows the major forces that are involved in formation of protein-ligand complexes for each type of the ligand. The bottom layer summarizes significant (frequently occurring) patterns for each force/bond type and each type of the ligand. The patterns are shown in $X^R \ldots Y^L$ or $X^R - Y^L$ format where X denotes an atom type of residue R in the protein, Y denotes an atom type of the ligand L, strong interactions (covalent and coordination bonds) are depicted by "–", and weak interactions (hydrogen bond) are represented by "…".

4.5.2 Interaction patterns in protein-organic compound complexes

Organic compounds bind to proteins mainly by the means of the van der Waals contacts and the hydrogen bonds. Total of 85771 contacts are observed between

an organic compound and a protein and they include 77554 van der Waals contacts, 7914 hydrogen bonds, and 246 covalent bonds. The remaining 0.1% of contacts is due to the electrostatic force. Among the 3685 protein pocket-organic compound complexes, 1067 complexes (29%) are based solely on the van der Waals contacts, 2309 (62.7%) involve both hydrogen bonds and van der Waals contacts, 107 (2.9%) incorporate covalent bonds and van der Waals contacts, and 135 (3.7%) include covalent bonds, hydrogen bonds, and van der Waals contacts, see Figure 4-2. We note that the number of hydrogen bonds is likely underestimated since REDUCE program could not supply complete coordinates for hydrogen atoms of some ligands and thus some potential hydrogen bonds could not be counted.

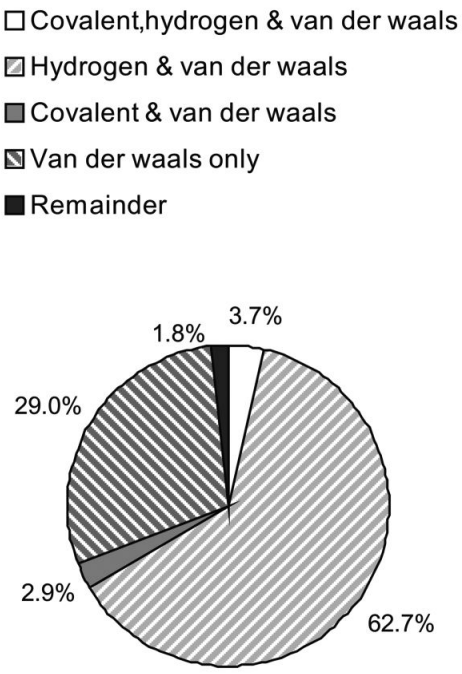


Figure 4-2: The summary of forces/bonds that are involved in formation of protein-organic compound complexes. The chart shows that most of the complexes involve multiple contact types with the most frequent contacts involving both van der Waals force and hydrogen bonds.

Covalent bonds

Majority of the 246 covalent bonds formed between organic compounds and proteins are summarized with four patterns: 1) 27 covalent contacts are formed between the thiol of Cys residue and the carbon atom of the organic compound (thioether bond); 2) 139 are formed between the nitrogen atom of Asn residue and the carbon atom of N-Acetyl-D-Glucosamine (NAG); 3) 28 concern the thiol of Cys residue and the sulfur atom of the organic compound (disulfide bond); and 4) 23 involve the nitrogen atom of Lys residue and the carbon atom of organic compound. We observe that the interaction between protein and NAG is established through the process of glycosylation and this interaction is not observed for other ligands. Therefore, this interaction is not included as a pattern for covalent bond. We denote the other three patterns as S^{cys}—C^{ligand}, S^{cys}—S^{ligand}, and N^{lys}—C^{ligand} respectively. They cover (27+28+23)/107 = 73% of all investigated covalent bonds between proteins and organic compounds; see summary in Table 4-1. Both the thiol of Cys and the nitrogen atom of Lys could interact with a variety of organic compounds. The result indicates that the covalent bonds could be formed only between a few specific atoms of some AAs and a few specific atoms of the organic compounds.

Table 4-1: A summary of interaction patterns concerning covalent bonds formed between a protein and an organic compound. The patterns are shown in X^R – Y^L format where X denotes an atom type of residue R in the protein and Y denotes an atom type of the ligand L.

Interaction pattern[1]	Average bond length (Å)	Occurrence	Ligands (organic compounds)
S^{cys}—C^{ligand}	1.83	27	3GC, 6NA, ACM, CYC, DBV, DKA, DPM, FAD, GOA, GVE, HC4, LBV, MKE, PEB, PLM, PYR, T10, XY2
S^{cys}—S^{ligand}	2.09	28	BME, DTT, SEO
N^{lys}—C^{ligand}	1.37	23	3PY, AZE, BGX, HPD, P3T, PBG, PLP, PYR, RET

Since some covalent bond patterns concern only a few dozens of complexes, we investigate whether they are specific to a certain protein family or whether they are more generic and associated with a variety of families. We note that in

contrast to the covalent bonds, in the case of the subsequently discussed coordination and hydrogen bonds, thousands of contacts between the proteins belonging to a wide range of families and the ligands are established. Based on SCOP classification system (Andreeva et al., 2008), the S^{cys}—C^{ligand} bonds are formed for proteins belonging to 15 families, which cover four major structural classes, i.e., all-α, all-β, α/β, and α+β. Similarly, the S^{cys}—S^{ligand} and N^{lys}—C^{ligand}, bonds concern proteins from 15 and 10 families and 4 and 3 structural classes, respectively. This shows that the above patterns span dozens of structurally different protein families, which in turn indicates that they are not specific to a certain protein family.

Hydrogen bonds

Hydrogen bonds are formed in 2466/3685 = 66.9% of the protein-organic compound complexes. Although all 20 AAs can establish hydrogen bonds with compounds, their ability to form hydrogen bonds varies. Table 4-2 shows the distribution of occurrence of the hydrogen bonds formed by each AA and the occurrence of the AAs in the 3685 pockets. The seven hydrophilic residues (based on the low values of their hydropathy index (Kyte and Doolittle, 1982)), including Arg, Lys, Asn, Thr, Ser, Gln, and His establish larger number of hydrogen bonds when compared their occurrence in the pockets. Moreover, six hydrophobic residues, i.e., Ala, Cys, Val, Ile, Met, and Leu, occupy 26.1% of the residues in the pockets and they form only 10.7% of the hydrogen bonds. This suggests that the hydrophilic residues form hydrogen bonds with the organic compounds more frequently when compared with the hydrophobic residues. Among the 7914 hydrogen bonds between proteins and organic compounds, AAs serve as donors for 6526 hydrogen bonds, and as acceptors for only 1371 hydrogen bonds; they serve as both donors and acceptors for the remaining bonds.

Table 4-2: A summary of hydrogen bonds formed between specific amino acids and organic compounds. [1]the hydropathy index values from reference (Kyte and Doolittle, 1982); the larger (smaller) the index values is, the more hydrophobic (hydrophilic) the amino acid. [2]the percentages of

hydrogen bonds between specific amino acids and DNA molecules were taken from Table 2 of reference (Luscombe et al., 2001).

Amino acid	% hydrogen bonds with organic compounds	% of occurrence in binding sites	% hydrogen bonds with DNA molecules[2]	# hydrogen bonds		Hydropathy index value[1]
				as acceptor	as donor	
Arg	20.0%	7.5%	33.6%	29	1555	-4.5
Lys	10.4%	5.1%	14.8%	18	802	-3.9
Ser	8.8%	6.2%	10.1%	63	631	-0.8
Thr	8.0%	5.6%	8.2%	68	566	-0.7
Asn	7.6%	5.1%	7.9%	106	497	-3.5
Gly	6.8%	8.8%	3.7%	50	488	-0.4
Tyr	5.2%	5.9%	3.5%	69	346	-1.3
His	5.1%	4.0%	3.6%	91	312	-3.2
Asp	4.8%	5.7%	1.0%	278	103	-3.5
Gln	4.5%	3.3%	6.3%	75	282	-3.5
Glu	4.5%	4.9%	1.6%	300	53	-3.5
Ala	3.0%	5.6%	1.8%	38	200	1.8
Leu	2.2%	7.3%	0.4%	40	137	3.8
Trp	2.1%	3.6%	0.3%	13	156	-0.9
Val	2.1%	5.1%	0.7%	36	128	4.2
Ile	1.8%	4.4%	1.3%	27	113	4.5
Phe	1.2%	5.2%	0.4%	21	72	2.8
Met	0.9%	2.2%	0.4%	11	57	1.9
Cys	0.7%	1.5%	0.4%	9	43	2.5
Pro	0.4%	3.0%	0.1%	29	2	-1.6

The distribution of occurrence of the hydrogen bonds with the organic compounds for the individual AAs is compared with the corresponding results obtained for protein-DNA interactions, which were derived based on 129 protein-DNA complexes (Luscombe et al., 2001), see Table 4-2. In both cases, the distributions are similar, i.e., Arg, Lys, Ser, Thr, and Asn establish the largest number of hydrogen bonds with both the organic compounds and the DNA molecules, while Phe, Met, Cys, and Pro establish the smallest number of hydrogen bonds with both types of ligands. The two AAs that establish the highest number of hydrogen bonds, Arg and Lys, are characterized by a larger number of bonds in the case of the binding with DNA, although we emphasize that the order of AAs in both cases is consistent. This suggests that the ability of AAs to establish hydrogen bonds could be an intrinsic characteristic of the AA itself, which is independent of the type of the ligand.

The negatively charged residues Asp and Glu did not exhibit strong affinity towards establishing hydrogen bonds in spite of having relatively high solvent accessibility and inclusion of two oxygen atoms in their side chains. We observe that Asp and Glu form the largest number of hydrogen bonds (278 and 300) when the AA serves as an acceptor. At the same time they form only 103 and 53 hydrogen bonds when they serve as donors, which is relatively small when contrasted with the number of hydrogen bonds formed by other hydrophilic residues, e.g., 1555 for Arg and 802 for Lys. This low affinity to form hydrogen bonds could be explained by considering that the carboxyl groups of Asp and Glu often lend their H^+ to solution, and as a result the two oxygen atoms on the carboxyl group are not bonded to hydrogen atom and cannot serve as a donor when forming the hydrogen bond.

The most frequently formed hydrogen bond is established between NH- group (as the donor) of an AA and the oxygen atom of an organic compound. This type of the hydrogen bond covers 5206/7914 = 65.8% of all hydrogen bonds. To compare, the NH- group of organic compound serving as the donor and the oxygen atom of AAs account for only 325 hydrogen bonds. The surface patch that is characteristic for NH- group has high potential to form hydrogen bonds with organic compounds. For instance, in the chain A of neuraminidase protein (PDB entry 1F8E) (Smith et al., 2001), the pocket that binds 4,9-AMINO-2,4-DEOXY-2,3-DEHYDRO-N-ACETYL-NEURAMINIC (abbreviated to 49A in PDB) includes 4 Arg residues, i.e., Arg118, Arg152, Arg292, and Arg371, see Figure 4-3. Three of them, Arg118, Arg292, and Arg371, are spatially adjacent and they form 5 hydrogen bonds with the oxygen atoms of 49A, while the other residues in the pocket establish only 2 hydrogen bonds. The cluster of the five hydrogen bonds is crucial for the interaction between the protein chain and the compound.

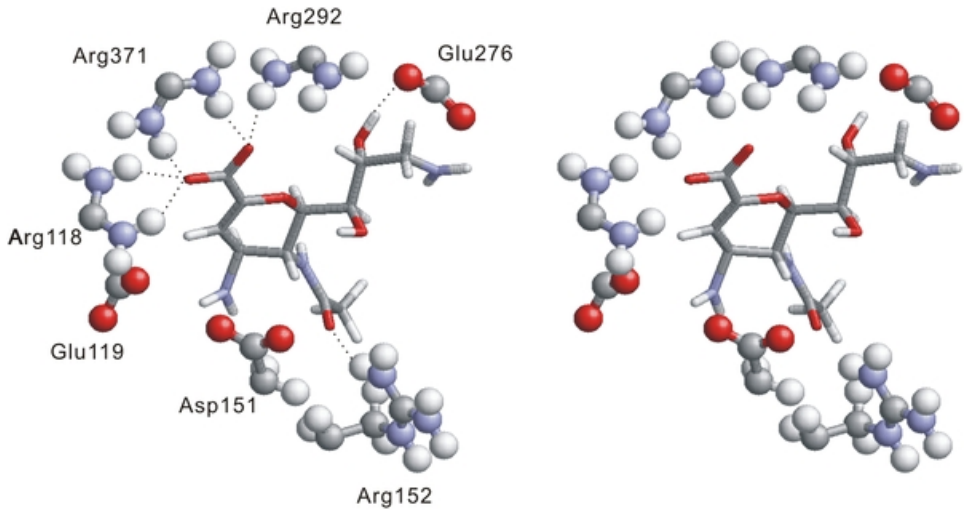


Figure 4-3: An example stereo diagram of hydrogen bonds formed between NH-group of a residue and oxygen atom of an organic compound. The oxygen atom is colored red, nitrogen atom is blue, carbon atom is gray, and hydrogen atom is white. The residues in the pocket are in ball and stick format while the ligand is in stick format. Hydrogen bonds are represented by "…". The structure is taken from chain A of neuraminidase protein (PDB entry 1F8E), which interacts with 49A. The binding pocket contains four Arg residues and each residue contains 2 NH- groups. Three Arg residues (Arg118, Arg292, Arg371) are spatially adjacent, and they form five hydrogen bonds with the oxygen atoms of the ligand.

Van der Waals contacts

Majority of the van der Waals contacts are formed between carbon, oxygen and nitrogen atoms. These three atoms result in nine potential combinations which cover 94.8% of all van der Waals contacts between proteins and organic compounds. The most common van der Waals contacts are established between a carbon atom of a residue and a carbon atom of a compound, and a carbon atom of a residue and an oxygen atom of a compound. Each of the abovementioned two cases accounts for more than 25% of all van der Waals contacts. In contrast to the covalent and hydrogen bonds, van der Waals contacts are irregular and lack frequently occurring patterns that would indicate involvement of particular residues.

4.5.3 Interaction patterns in protein-metal ion complexes

Among 1682 protein-metal ion complexes, 639 involve both coordination bonds and electrostatic force, 459 are based on electrostatic force but with no coordination bonds, and 499 incorporate coordination bonds with no electrostatic force. Overall, electrostatic force and coordination bonds are involved in (639+459+499) = 1597 complexes, which correspond to 1597/1682 = 94.9% of all protein-metal ion complexes.

Asp and Glu residues are negatively charged and could potentially form electrostatic contact with metal ions. Since the charge is not evenly distributed over the AAs, we analyzed which non-hydrogen atom of Asp/Glu is the closest to the metal ions. Among 1098 complexes which involved the electrostatic force, metal ions formed electrostatic interactions with Asp and Glu for 1511 times (in some complexes more than 1 electrostatic interaction is formed). In the case of 1385 out of the above 1511 interactions, the oxygen atoms of the carboxyl group of Asp and Glu are the closest to the metal ion. This suggests that these two oxygen atoms could be more negatively charged than other atoms in the side chains.

Metal ions are observed to form coordination bonds with up to 6 atoms of a given protein, i.e., in chain A of 4-chlorobenzoyl coenzyme A dehalogenase protein (PDB entry 1NZY) (Benning et al., 1996), the calcium ion is coordinated with oxygen atoms of Gly49, Leu202, Ala203, Ala205, Thr207 and Gln210. On the other hand, some metal ions form coordination bonds with just one atom, i.e., in the chain A of human sex hormone-binding globulin protein (PDB entry 1D2S) (Grishkovskaya et al., 2000), the calcium ion interacts only with His136. A total of 2345 coordination bonds are formed among the 1138 protein-metal ion complexes that involve this type of bond. The nitrogen atom in the side chain of His forms 787 bonds with the coordinating metal ions, sulfur atom of Cys forms 434 coordination bonds with metal ions, and oxygen atom (of any AA except Asp/Glu since interaction between metal ion and Asp/Glu is considered to be

based on the electrostatic force) forms 1039 coordination bonds. The bonds based on these three patterns correspond to (787+434+1039)/2345 = 96.4% of all coordination bonds. The strong affinity of the oxygen atom to form coordination bonds with metal ion suggests that the interaction between the negatively charged Asp and Glu residues and metal ions could be a combination of both the coordination and the electrostatic force. The interaction between metal ions and Asp/Glu has been considered as coordination in many other studies. For instance, Angkawidjaja and colleagues reported that Ca^{2+} is coordinated by the side chains of Asp153, Asp157, and Gln120, and the carbonyl oxygens of Thr118 and Ser144 (Angkawidjaja et al., 2007); similarly, Declercq and coworkers show interaction between Ca^{2+} and the coordinating oxygen atoms of Asp51, Asp53, Ser55, Phe57, Glu59 and Glu62 (Declercq et al., 1999). As a result, the interactions between metal ions and Asp/Glu should be regarded as both coordination and electrostatic contacts if the distance between the corresponding atoms satisfies the definition of the coordination bond and the electrostatic contact.

Although the generic principles that govern protein-metal ion interactions were discussed in prior works (Dudev and Lim, 2008a; Dudev and Lim, 2008b; Gifford et al., 2007; Maret, 2005), e.g., interactions concerning Cys-rich Zn^{2+}-binding sites and affinity of interaction between Mg^{2+} and Asp/Glu in protein cavities (Dudev and Lim, 2008a), we could not find a systematic study that investigates how many residues and what residues types are involved ("preferred") in the coordination bonds with specific metal ions, and that provides insights concerning similarities in the geometry of the coordination-based interactions with metal ions, which are discussed below.

Among the metal ions, Ag^{+}, Ca^{2+}, $Cu^{2+}(Cu^{+})$, Cd^{2+}, $Co^{2+}(Co^{+})$, $Fe^{3+}(Fe^{2+})$, Hg^{2+}, K^{+}, Mg^{2+}, Mn^{2+}, Na^{+}, Ni^{2+}, Pb^{2+}, Sm^{2+} and Zn^{2+} form coordination bonds with atoms of residues, see Table 4-3. Zn^{2+} forms coordination bonds in the largest number of pockets. This ion is coordinated by atoms of at most 4 residues in a given pocket and it favors to be coordinated by 3 or 4 residues. The second highest number of pockets that involve coordination bonds with a metal ion

concerns Ca^{2+}. These ions are coordinated by atoms of up to six residues in a pocket, and they prefer to form the coordination bonds with 4 or 5 residues. Coordination bonds with Mg^{2+} and Cd^{2+} ions involve 228 and 109 pockets, respectively. In contrast to Zn^{2+} and Ca^{2+}, Mg^{2+} and Cd^{2+} ions form most of these bonds with atoms of 1 or 2 residues in a given pocket. Na^{+} ions form coordination bonds in 150 pockets and it favors to be coordinated by atoms of 3 or fewer residues. These 5 ions form coordination bonds in (426+328+228+109+150) = 1241 pockets, which constitutes 1241/1542 = 80.5% of all relevant pockets. The above results suggest that different metal ions prefer to be coordinated by a different number of residues in a given protein pocket.

Table 4-3: A summary of the coordination bonds between metal ions and a given number of residues in a protein pocket that contributes at least one atom to form the bond.

Metal ion	# pockets in which a given metal ion forms coordination bond with atoms of *x* residues						# of pockets for a given metal ion
	x = 6	*x* = 5	*x* = 4	*x* = 3	*x* = 2	*x* = 1	
Zn^{2+}	0	0	120	123	74	109	426
Ca^{2+}	24	70	84	44	50	56	328
Mg^{2+}	1	0	14	59	73	81	228
Na^{+}	1	5	17	44	41	42	150
Cd^{2+}	0	1	5	7	26	70	109
Mn^{2+}	0	1	16	24	20	24	85
Fe^{3+}	1	4	15	28	6	5	59
K^{+}	1	7	13	11	11	3	46
Cu^{2+}	1	1	5	19	8	5	39
Ni^{2+}	0	0	4	14	10	7	35
Co^{2+}	0	0	4	9	5	0	18
Hg^{2+}	0	0	1	0	6	9	16
Ag^{+}	0	0	0	1	0	0	1
Sm^{2+}	0	0	0	0	1	0	1
Pb^{2+}	0	0	0	0	0	1	1

The residues which are coordinated by the same metal ion are grouped and we denote such groupings as the residue groups. We count the frequencies of the residue groups among different metal ions. For instance, given that Zn^{2+} forms coordination bonds with 4 Cys residues in 47 pockets, the corresponding frequency of $(Cys)_4$ residue group is 47. The residue groups that are coordinated by at least 10 metal ions are shown in Figure 4-4, Figure 4-5 and Figure 4-6. The

frequencies of residue groups that contain 5 or more residues are below 10 and thus they are not included in the above Figures. Total of 5 residue groups, i.e., $(Cys)_4$, $(Cys)_3(His)$, $(Cys)_2(His)_2$, $(Asp)_2(His)_2$, and $(Asp)(His)_3$, include 4 residues, see Figure 4-4. We observe that the $(Cys)_4$ group is coordinated by the largest number of metal ions (47 metal ions). There are 11 residue groups that incorporate 3 residues, see Figure 4-5. These groups include $(Cys)_3$, $(Cys)_1(His)_2$, $(Asp)_3$, $(Asp)_2(Glu)$, $(Asp)_2(His)$, $(Asp)(Glu)_2$, $(Asp)(Glu)(His)$, $(Asp)(His)_2$, $(Glu)_2(His)$, $(Glu)(His)_2$ and $(His)_3$. The $(Asp)(His)_2$ and $(His)_3$ groups are coordinated by the largest number of 44 and 38 metal ions, respectively. Finally, 6 residue groups, i.e., $(Asp)_2$, (Asp)(Glu), (Asp)(His), $(Glu)_2$, (Glu)(His) and $(His)_2$, that make contact with 2 residues, see Figure 4-6.

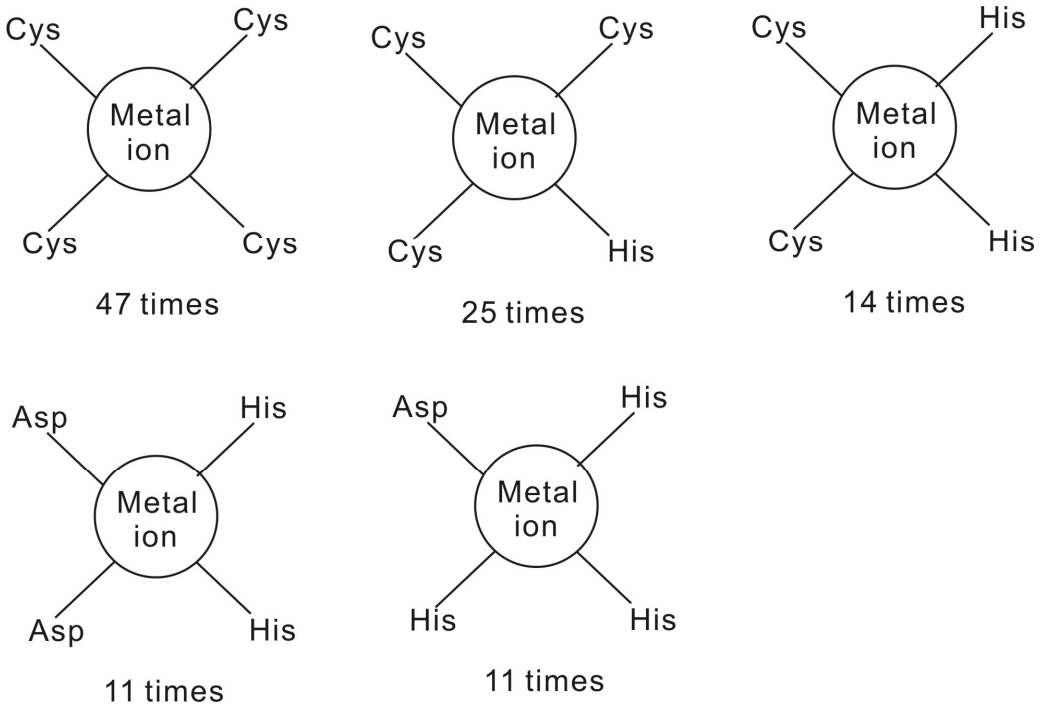


Figure 4-4: The residue groups that are coordinated by at least 10 metal ions and consist of 4 residues.

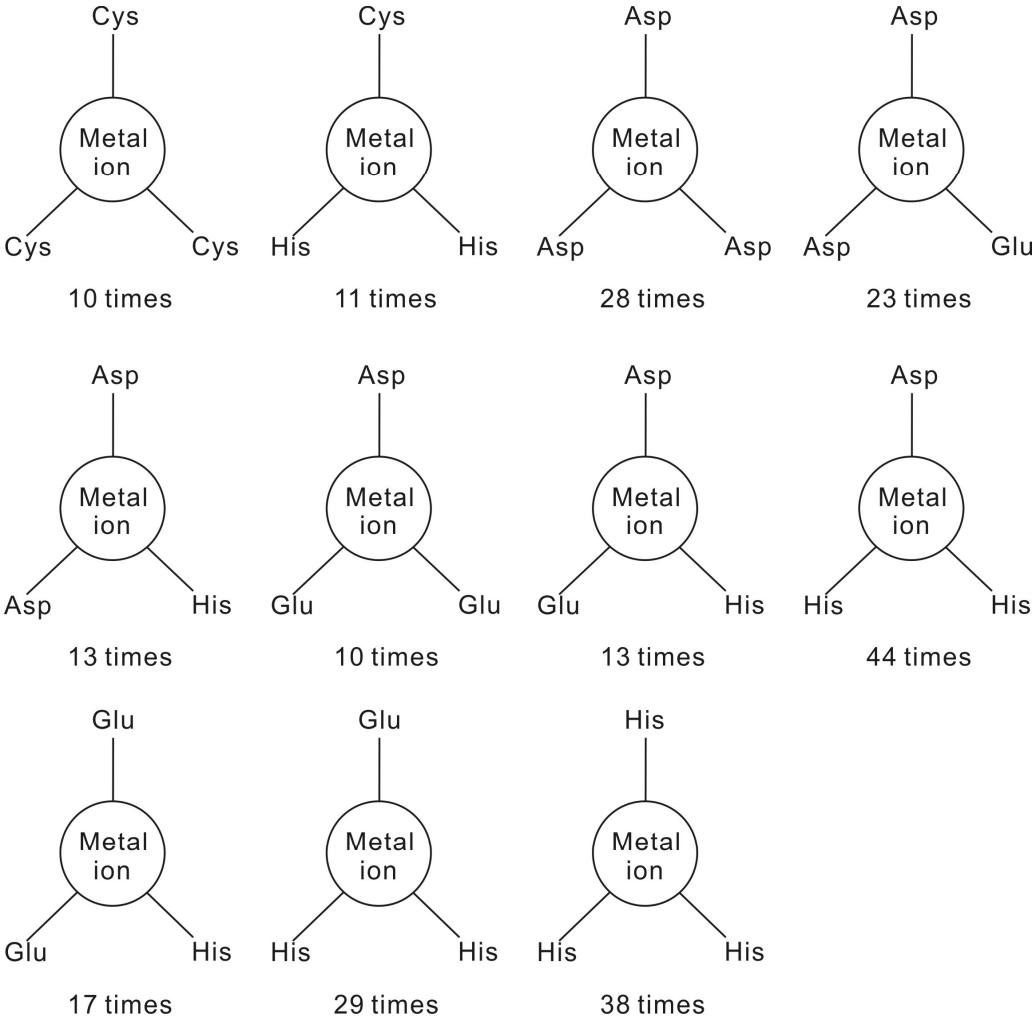


Figure 4-5: The residue groups that are coordinated by at least 10 metal ions and consist of 3 residues.

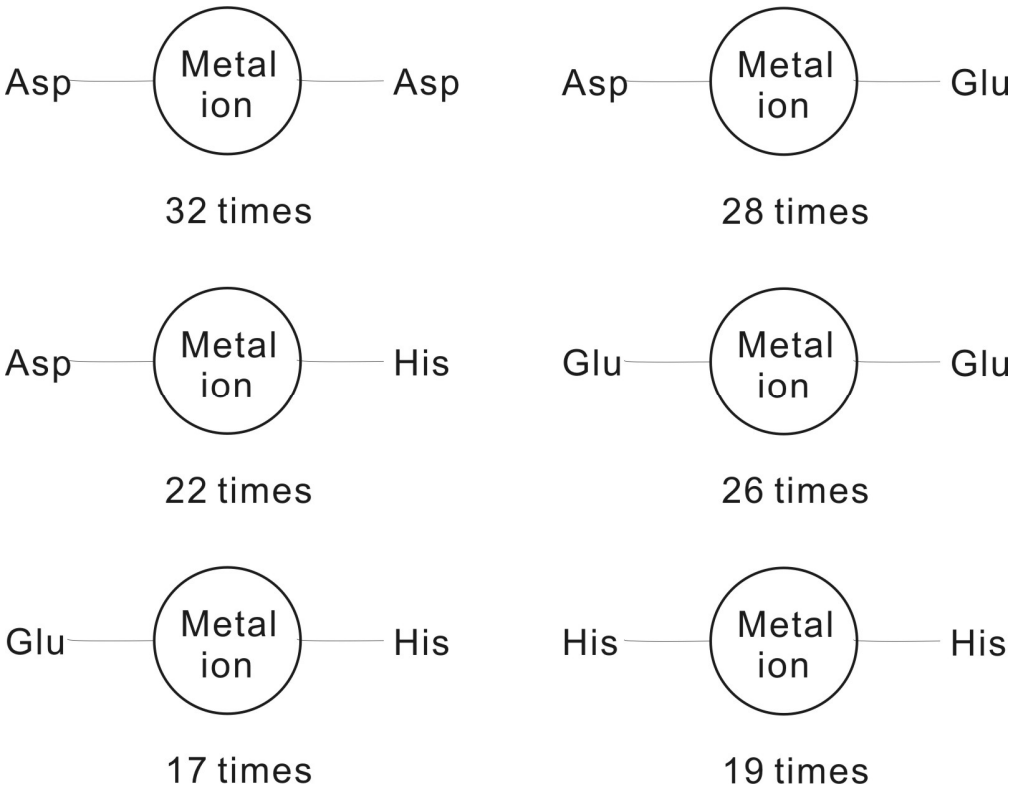


Figure 4-6: The residue groups that are coordinated by at least 10 metal ions and consist of 2 residues.

Cys and His are among the residues that have a strong ability to form coordination bonds with the metal ions. We observe that although the geometry of $(Cys)_4$–metal ion and $(His)_3$-metal ion interactions is different, each of these residue groups has similar geometry across the set of the corresponding pockets. The prevalent way to form the coordination bond between Cys and a metal ion involves four Cys residues arranged spatially close to each other to form a pocket; the metal ion is located in the center of this pocket. For example, in the chain A of PHD finger protein 21A (PDB entry 2PUY) (Lan et al., 2007), the zinc ion forms coordination bonds with Cys503, Cys506, Cys529, and Cys532. The distance between zinc ion and the sulfur atom of the four Cys residues varies between 2.26 Å and 2.41 Å. The four sulfur atoms form an approximate regular tetrahedron and the zinc ion is located in its center, see panel A in Figure 4-7. The length of the tetrahedron edges varies between 3.63 Å and 3.93 Å. On the other hand, the

coordination interaction between His and metal involves three His residues arranged to form a pocket with the metal ion located in approximately the same distance to the nitrogen atoms of these three residues. For example, in the chain A of Zn-dependent hydrolase protein (PDB entry 2R2D) (Liu et al., 2007), the zinc ion forms coordination bonds with nitrogen atoms of His111, His113, and His191. The distance between the zinc ion and the nitrogen atoms varies between 2.06 Å and 2.18 Å, see panel B in Figure 4-7. The three nitrogen atoms form an approximate equilateral triangle with the length of the sides that varies between 3.14 Å and 3.31Å.

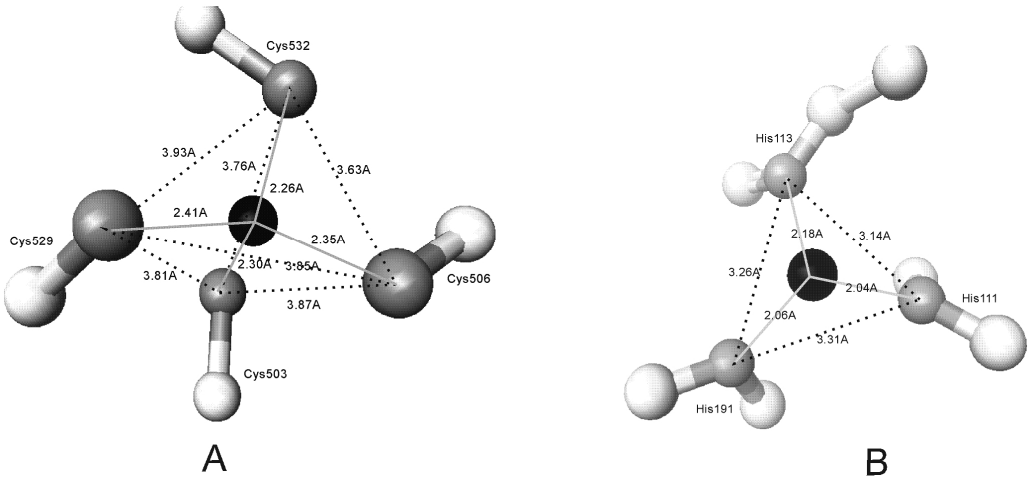


Figure 4-7: Examples of typical coordination bonds between metal ions and Cys and His residues. Coordination bonds are represented by solid lines; the dashed lines show the distance between atoms of different residues. Panel A shows the coordination bond between zinc ion and four Cys residues where sulfur atom is shown in gray, carbon atom in white, and zinc ion in black. The sulfur atoms of four Cys residues form an approximate regular tetrahedron and the zinc ion is located in its center. Panel B shows the coordination bond between zinc ion and three His residues. The nitrogen atoms are shown in gray, other atoms of the His side chain are in white, and zinc ion is colored black. The three nitrogen atoms form an approximate equilateral triangle with the length of the sides that varies between 3.14 Å and 3.31 Å. The zinc ion is not located on the triangle plane.

4.5.4 Interaction patterns in protein-inorganic anion complexes

Inorganic anions bind to proteins mainly through electrostatic force, hydrogen bonds and van der Waals contacts. Among the 1837 anions, 1188 interact with the

positively charged AAs such as Arg, His and Lys based on electrostatic interaction and 641 bind to the pocket by the means of hydrogen bonds and van der Waals contacts.

Similarly as in the case of metal ions, we studied which atoms of the positively charged residues are the closest to the inorganic anions. Among the 1188 protein-anion complexes that involve electrostatic force, 202 anions bind to His, 327 to Lys, and 659 to Arg. The nitrogen atom in the side chain of these three residues is the closest atom to the anion for 172 anion-His interactions, 222 anion-Lys interactions, and 565 anion-Arg interactions. These numbers suggest that the nitrogen atoms of positively charged residues may be closer to the center of the charge than other non-hydrogen atoms.

Among the anions that occur in PDB more than 100 times, 743 SO_4^{2-} (743/948 = 78.5%) and 109 PO_4^{3-} (109/148 = 73.6%) bind to positively charged residues, while some other anions less frequently bind with the charged residues. More specifically, 165 Cl^- (165/345 = 47.8%), 33 Br^- (33/126 = 26.2%), and 22 I^- (22/108 = 20.4%) bind to positively charged residues. This could be explained by the formation of hydrogen bonds between the oxygen atoms of SO_4^{2-} and PO_4^{3-} and the NH- group of positively charge residues. For instance, PO_4^{3-} forms 254 hydrogen bonds with positively charge residues (254/109 = 2.3 hydrogen bonds per pocket) and SO_4^{2-} forms 1394 hydrogen bonds with positively charge residues (1394/743 = 1.9 hydrogen bonds per pocket). The combination of electrostatic force and hydrogen bonds stabilizes the anion-positively charged residue interaction.

Similarly to the protein-organic compound complexes, the most frequent hydrogen bond incorporates the NH- group of a residue that serves as the donor and the oxygen atom of a ligand that serves as acceptor. This pattern concerns 2777 hydrogen bonds which converts into 2777/3190 = 87.1% of all hydrogen bonds between a protein pocket and an inorganic anion.

4.5.5 Interaction patterns in protein-inorganic cluster complexes

Amidst the nine types of inorganic cluster that could be found in PDB, FS4, FES, SF4, F3S, CLF, and FS3 are Fe-S clusters and contain only iron and sulfur atoms. The remaining three inorganic clusters, which include CFN, FSO, and NFS, also mainly contain iron and sulfur atoms.

We observe that coordination bonds are involved in all 54 protein-inorganic cluster complexes. These bonds are usually formed between the iron atom of the cluster and the sulfur atom of Cys residue, and the iron atom of the cluster and the nitrogen atom in the side chain of His residue. These two coordination bond patterns cover 201/204 = 98.5% of all coordination bonds between inorganic cluster and a protein pocket. Although FS4, SF4, F3S, and FS3 are positively charged and FSO is negatively charged, these clusters do not interact with charged residues. We did not find the electrostatic force based interactions between the inorganic clusters and proteins.

4.5.6 Overlap and coverage of the interaction patterns

The 10 patterns that concern covalent bonds, coordination bonds and hydrogen bonds, see the bottom layer in Figure 4-1, appear in 2013 protein-organic compound complexes, 1138 protein-metal ion complexes, 1115 protein-anion complexes, and 53 protein-inorganic cluster complexes, which corresponds to (2013+1138+1115+53)/7759 = 55.7% of all protein-ligand complexes. Significant majority of the above complexes incorporates just one of the discussed patterns. More specifically, except for 81 protein-organic compound complexes and 546 protein-metal ion complexes that incorporate two or more interaction patterns, the remaining 4238 protein-ligand complexes include one interaction pattern.

4.6 Conclusions

In this chapter, we investigated several frequently occurring interaction patterns that concern atomic-level protein-ligand interactions. The considered protein pocket-ligand complexes were grouped into four categories: protein-organic

compound, protein-metal ion, protein-anion, and protein-inorganic cluster complexes. These groups cover 93.5% of all protein-ligand complexes from PDB and we show that they are governed by different types of interaction forces. The protein-organic compound complexes are governed by the hydrogen bonds, van der Waals contacts and covalent bond. The protein-metal ion complexes are based on the electrostatic force and coordination bonds while the protein-anion complexes are governed by the electrostatic force, hydrogen bonds and van der Waals contacts. Finally, the protein-inorganic cluster complexes are established mostly due to the coordination bonds.

We present several frequently occurring interaction patterns, defined in terms of prevalent interactions between specific atoms of specific residue in the protein's pocket and specific atoms of the ligand, for the abovementioned four groups and for the specific types of interaction forces. We quantify relative abundance of specific interaction types and discuss their characteristic features such as commonly interacting amino acid types. Total of 10 interaction patterns that occur in 56% of all considered complexes are found. For example, we show that 66.9% of the protein-organic compound complexes involve hydrogen bonds and that 65.8% of these hydrogen bonds are formed between the NH- group of the protein's residue and the oxygen atom of the organic compound. As a result, we believe that the geometric and electrostatic complementary, which are used for molecular recognition, should be supplemented by implementation of hydrogen bond(s) in the case of the protein-organic compound complexes. As another example, only three interaction patterns are sufficient to summarize significant majority, i.e., 73%, of normal covalent bond interactions between proteins and ligands; they include the covalent bond between the thiol of Cys residue and the carbon atom of the ligand (thioether bond), the thiol of Cys residue and the sulfur atom of the ligand (disulfide bond), and the nitrogen atom of Lys residue and the carbon atom of the ligand. We also show that the AAs serve as donors for significant majority of these hydrogen bonds. We observe that most of the inorganic anions interact with positively charged AAs including Arg, His, and Lys.

We show that the organic compounds form hydrogen bonds more frequently with hydrophilic AAs when compared with hydrophobic AAs, which is consistent with the results obtained for the protein-DNA interactions (Luscombe et al., 2001). This suggests that the ability of AAs to establish hydrogen bonds could be an intrinsic characteristic of a given AA, which is independent of the ligand type.

To conclude, we show that for a given type (group) of ligands and a given type of the interaction force, majority of protein-ligand interactions are repetitive and could be summarized with several simple atomic-level patterns. These interaction patterns not only provide a comprehensive overview of protein-ligand interactions, but they also may have important implications for the development of binding site prediction methods. Our results suggest that one model cannot effectively predict all protein-small ligand interactions. The prediction method should be composed of several modules that target predictions for specific types of small ligands, for which interaction pattern can be found.

CHAPTER 5 Assessment on existing binding site predictors for small organic ligands

5.1 Introduction

In chapter 4, we have demonstrated that some interaction patterns are shared by a number of protein-ligand complexes. This result implies that (some) binding sites are predictable. They could be identified, for example, by comparing them to a library of known, similar binding sites or by using a model that describes a given pattern or a set of patterns. To date, more than a dozen methods have been already proposed for the prediction of binding sites of small organic compounds. Unfortunately, these methods were never systematically compared and analyzed. Some studies that introduced new binding site predictors have compared them to a few existing solutions, however, the benchmark dataset used for the comparison is characterized by a largely incomplete annotation of binding sites. Additionally, prior benchmark datasets include annotations of biologically "irrelevant" ligands, such as the glycol molecule that is introduced by the purification and crystallization procedures. To this end, we perform a comparative evaluation of the predictive quality of ten representative binding site predictors on a set of proteins that are annotated with multiple binding sites, which are confirmed to be biologically relevant (Chen et al., 2011). This chapter gives an overview of the state-of-the-art binding site predictors and finds several limitations of these methods. These limitations should be addressed by the future designers of the binding site predictors.

5.2 Related work

We selected 10 prediction methods that offer either a web-server or a standalone program to generate the predictions. Overall, the structure-based binding site predictors utilize three types of approaches including geometrical analysis, calculation of binding energy, and threading using structural templates;

additionally, one solution is based on a consensus of geometry- and energy-based approaches. The geometry-based methods find crevices/pockets on the protein surface and output them (after sorting them based on their volume and/or sequence conservation) as the predicted binding sites. This approach is based on the premise that small ligands usually bind to pockets on the protein surface. The energy-based methods find patches on the protein surface that offer energetically favorable conformation, as evaluated using one of the scoring functions discussed in Section 2.2.4. These patches include atoms that could potentially interact with small ligands. The threading-based methods utilize a library of known protein-small ligand complexes and perform predictions by finding the most similar complex for a given input protein. The considered methods include the geometry-based SURFNET (Laskowski, 1995), PocketFinder (Hendlich et al., 1997), PASS (Brady and Stouten, 2000), LIGSITE^{csc} (Huang and Schroeder, 2006), PocketPicker (Weisel et al., 2007), ConCavity (Capra et al., 2009), and Fpocket (Le et al., 2009), the energy-based Q-SiteFinder (Laurie and Jackson, 2005), the threading-based Findsite (Skolnick and Brylinski, 2008), and the consensus-based MetaPocket (Huang, 2009). We evaluate the predictive quality of the 10 methods using two criteria that were introduced in prior works and a new quality index that gives additional insights. For convenience, we use 'ligands' and 'binding sites' to refer to the small organic compounds and the sites on the protein structure where they bind, respectively.

5.3 Preparation of benchmark dataset

The benchmark dataset is designed to cover a wide range of non-homologous (functionally dissimilar) protein structures and to include structures with the largest number of annotated binding sites. We select a representative chain for each SCOP family and we map the binding sites of other similar structures into this chain. Prior work shows that two chains from different SCOP families have less than 1% chance to share more than 25% sequence similarity (Levitt, 2007). Since every chain in our dataset comes from a different protein family, the

included proteins should be dissimilar in both their tertiary structure and sequence. We download all available protein-ligand complexes from the PDB as of August 18th, 2009 and we annotate these proteins with their corresponding SCOP families. One chain for each SCOP family is selected using the following procedure. First, sequence similarity and structural similarity expressed with TM-score (Zhang and Skolnick, 2005) are calculated for every pair of chains within a given SCOP family. TM-score measures the structural similarity between a pair of protein structures and varies between 0 and 1 (Zhang and Skolnick, 2005); larger values indicate higher similarity. Next, the two similarity scores are used to perform clustering. Two chains are assigned to the same cluster if their sequence similarity is above 80% and their TM-score is above 0.5, as suggested in (Zhang and Skolnick, 2005). We assume that the chains of the same cluster are homologous and that they share the same binding sites. Finally, we count the number of types of ligands that interact with the chains of each cluster. The cluster with the largest number of the ligand types is selected and this cluster is represented by the protein with the largest number of bound ligands. The latter choice is made to maximize the number and accuracy of the annotations of the binding sites. The ligands in the other chains in the selected cluster are superimposed into the representative structure using Fr-TM-align (Pandit and Skolnick, 2008). If the superimposed ligand structure clashes with the representative protein structure, then this ligand is removed. This step results in a protein structure that includes a (large) number of bound ligands, where some of these ligands could be redundant. A single-linkage clustering was performed to remove the redundancy. The distance between two ligands is defined as the minimum distance between any atom of one ligand and any atom of the other ligand. The clustering is terminated using 5Å threshold to ensure that ligands from one cluster do not overlap with ligands from another cluster. The median structures are chosen for each cluster of ligands. These median structures form a set of non-redundant ligands that bind to the protein structure, which represents a given SCOP family.

The resulting dataset contains 314 protein structures. These structures are manually inspected to filter out biologically irrelevant ligands, such as the glycol molecule that is introduced by the purification and crystallization procedures. For the structures with a published reference, a ligand is considered as biologically relevant if it is mentioned in the title or the abstract of the reference or the interaction between the ligand and the target protein is discussed in the results section. The ligands that only appear in the materials section or are never mentioned in the reference are removed. In case of the structures with no published reference, we use rules that were recently suggested by Wodak and colleagues (Dessailly et al., 2008). A given ligand is considered as biologically relevant if 1) it includes at least 10 non-hydrogen atoms; 2) it establishes at least 70 inter-atomic contacts with the protein atoms; and 3) the interaction does not concern lipid and membrane proteins. Our benchmark dataset includes 251 proteins after removing the "irrelevant" ligands. These are ligand-bound (i.e., holo) structures. Since the protein-ligand interactions could lead to conformation changes, we also generate a dataset that consists of the matching ligand-unbound (i.e., apo) structures. For each protein in the benchmark dataset we searched for its corresponding apo structure in the PDB. An apo structure is assumed to correspond to a given holo structure if they belongs to the same SCOP family, they share more than 80% sequence similarity and their TM-score is above 0.5. We found 104 apo structures and we created two additional datasets, D_{Apo} dataset that includes the 104 apo structures and D_{Holo} dataset which is a subset of the corresponding 104 holo structures from the benchmark dataset.

The proteins in the benchmark dataset have diverse overall structural topology. Based on the annotation from the SCOP database, they cover 6 structural classes, 148 protein folds, 184 superfamilies, and 251 protein families. The maximal pairwise sequence similarity is between 11% and 24%. The 251 proteins are annotated with 475 binding sites which interact with 253 types of ligands. All datasets including the full benchmark dataset and the D_{Holo} and D_{Apo} datasets are available at http://biomine.ece.ualberta.ca/BindingSitesPredictors/main.htm.

5.4 Evaluation measures

We use three indices to evaluate predictions of the considered binding site predictors:

- D_{CA}, which is defined as the minimal *distance* between the *center* of the predicted binding site (pocket) and any *atom* of the ligand, was widely used to assess the prediction quality in several prior studies. For instance, authors of LIGSITE^{csc}, PASS, PocketPicker and Fpocket assume that a predicted site is correct if its center is no farther that 4Å to any atom of the ligand. Instead of using one arbitrary threshold, we compute the success rates using D_{CA} values for integer thresholds between 1Å and 20Å.

- D_{CC}, which is defined as the minimal *distance* from the *center* of the predicted binding site to the *center* of the ligand, was proposed by Skolnick and colleagues (Skolnick and Brylinski, 2008). When compared with D_{CA}, this measure compensates for the size of the ligand, i.e., D_{CA} gives higher success rates for larger ligands. D_{CC} was recently used to compare Findsite and LIGSITE^{csc} (Skolnick and Brylinski, 2008). The success rates are computed using integer thresholds between 1Å and 20Å.

- O_{PL}, which quantifies *overlap* between the predicted *binding site* and the *ligand*, is proposed in this dissertation. This measure is defined as the ratio between the volume of the intersection of the predicted site and ligand, and the volume of their union. In addition to being sensitive to the size of the ligand, this quality index improves over both D_{CA} and D_{CC} by compensating for the relative spatial orientation of the ligand and the binding site. It can be computed for the four methods that output the full set of grid points of the predicted site (Q-SiteFinder, PocketPicker, ConCavity, and PocketFinder) instead of just the center of the pocket that is predicted by the other considered predictors. To calculate this value, both the binding site (pocket) and ligand are represented using a set grid points in the same grid scale. A

grid point is assigned to the ligand/site if the distance between this point and ligand/site is smaller than half of the length of diagonal in the grid cube. The O_{PL} value is computed as the ratio between the number of grid points that are shared by the ligand and the binding site, and the number of grid points that belong to either the ligand or the site.

5.5 Assessment on binding site predictors

We evaluate the performance of the ten prediction methods on a non-redundant benchmark dataset of 251 proteins. These methods are also compared against a baseline predictor which randomly selects a surface patch on the target protein; the center of the patch is used as the prediction. Prior studies usually take top three or top five predictions and verify whether any of them are within a certain distance (which is used as a cutoff for calculation of prediction accuracy) to the actual binding site. If at least one of the top predictions is below the cutoff, then the binding site is assumed to be correctly predicted. Since the previously used benchmark datasets contain proteins that are annotated with one binding site, the number of correctly predicted sites equals to the number of proteins and predictions were assessed "per protein". In our case majority of the proteins are annotated with multiple binding sites, and thus our assessment is "per binding site". For a protein with n binding sites we take the top n predictions for every considered method (in the case that one method generates less number of predictions than the number of binding sites, all predictions are used for evaluation). A given binding site is correctly predicted if the minimal distance between this site and any of the n predictions from a given method is below a threshold D. The success rate is defined as the number of correctly predicted binding sites divided by the total number of sites.

5.5.1 Comparison of the overall prediction quality

The success rates of the ten methods and the random baseline predictor quantified using D_{CC}, which measures the distance from the center of the predicted site to the center of the ligand, are shown in Figure 5-1, where y-axis stands for the success

rates and the *x*-axis represents the cutoff distance *D*. Findsite outperforms all other considered predictors for thresholds *D* up to 10Å. The ConCavity achieves the "second-best" success rates and is chosen to represent the geometry-based approaches in the subsequent analysis. Several predictors, including Q-SiteFinder, MetaPocket and PocketPicker have comparable, "third-best" success rates. The SURFNET, which is the oldest method that was designed over a decade ago, has the lowest success rates but it still improves over the baseline.

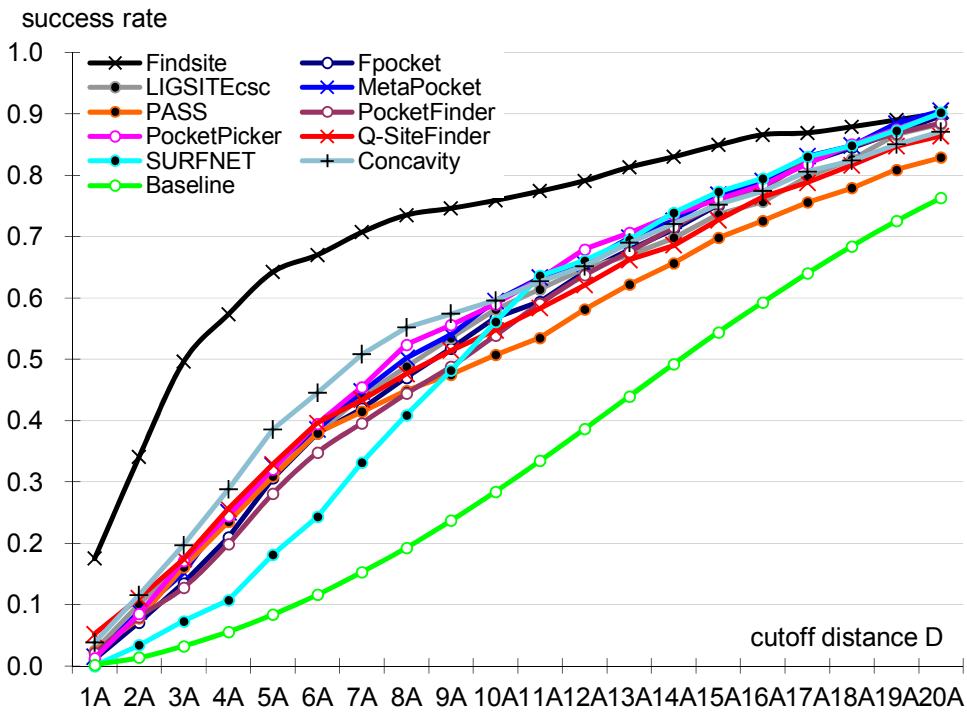


Figure 5-1: The success rates (*y*-axis) of the ten representative methods measured using D_{CC} (the minimal distance from the center of the predicted site to the center of the ligand) on the benchmark dataset. A given binding site is regarded as correctly predicted if the minimal distance between this site and the top *n* predictions is below the cutoff distance *D* (*x*-axis), where *n* is the number of binding sites of the protein that includes the evaluated binding site.

Figure 5-2 summarizes the success rates measured using D_{CA}, which is based on the distance from the center of the predicted site to any atom of the ligand. The results are similar to the results obtained using D_{CC}, except for *D* = 1Å where the Q-SiteFinder is the top-performing predictor. For the cutoff *D* = 4Å, which was

suggested by Skolnick and colleagues (Skolnick and Brylinski, 2008), the threading-based Findsite successfully predicts around 57% and 68% of the binding sites for the D_{CC} and D_{CA} distance definition, respectively, the geometry-based ConCavity identifies 28% and 51% of the binding sites, the energy-based Q-SiteFinder finds 26% and 44% of the binding sites, and the remaining methods cover 11-25% and 31-45% of the binding sites, respectively. To compare, the baseline random predictor correctly finds 5% and 9% of the binding sites when considering D_{CC} and D_{CA} distances, respectively.

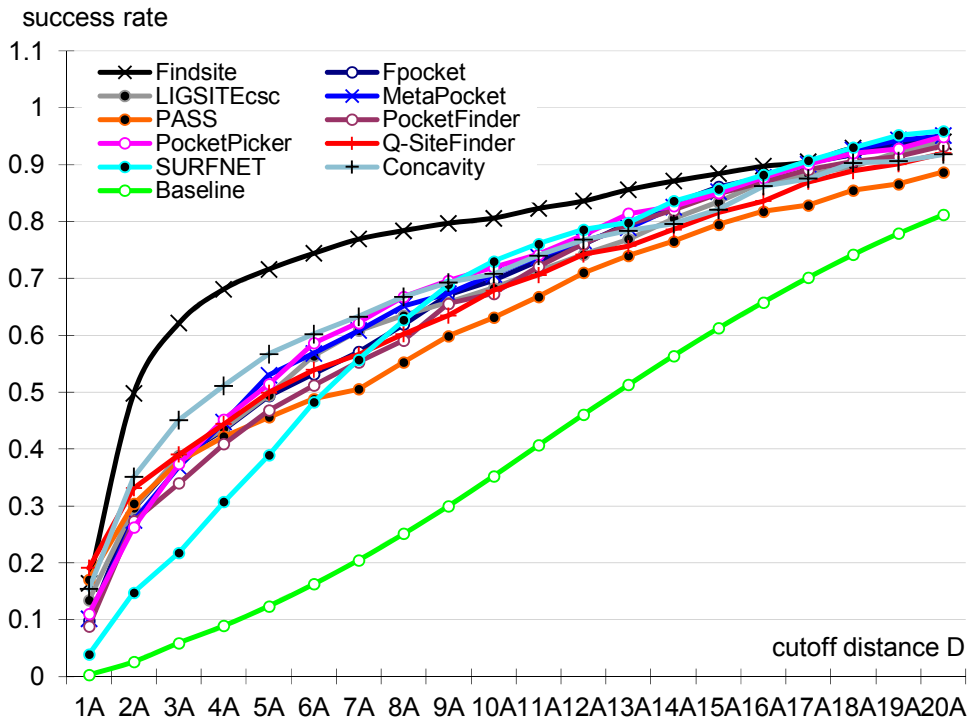


Figure 5-2: The success rates (y-axis) of the ten representative methods measured using D_{CA} (the minimal distance from the center of the predicted site to any atom of the ligand). A given binding site is regarded as correctly predicted if the minimal distance between this site and the top n predictions is below the cutoff distance D (x-axis), where n is the number of binding sites of the protein that includes the evaluated binding site.

The overlap index O_{PL}, which is defined as the ratio between the volume of the intersection of the predicted binding site and the ligand, and the union of the two volumes, expresses normalized spatial overlap between the predicted and the

actual location of the ligand. The O_{PL} could be calculated only for the ConCavity, Q-SiteFinder, PocketPicker and PocketFinder which generate a full set of grid points of the predicted pocket instead of just the center of the pocket that is outputted by the remaining predictors. We observe that about 60% of the binding sites predicted by ConCavity overlap with the predicted pocket while the coverage is only around 40% for Q-SiteFinder and PocketPicker; see Figure 5-3, where the *y*-axis shows the percentage of binding sites that have their O_{PL} values equal or greater than value on the *x*-axis. However, in most cases, the overlapping volume measured using O_{PL} is rather small; for instance, O_{PL} is above 20% only for about 16% of the binding sites.

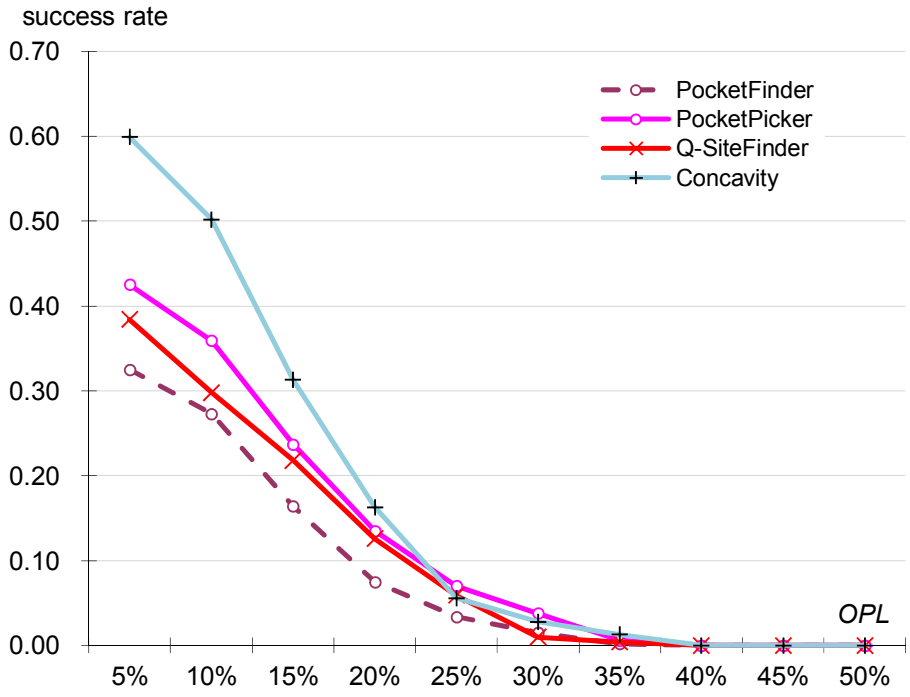


Figure 5-3: The success rates (*y*-axis) of PocketPicker, Q-SiteFinder, ConCavity and PocketFinder measured using O_{PL} (*overlap* between the predicted *pocket* and the *ligand*).

5.5.2 Statistical analysis of the predictions by the considered methods

We investigate significance of differences in the prediction quality measured with D_{CC} for all pairs of the considered prediction methods. For a protein with *n*

binding sites we take the top n predictions for every considered prediction method. We generate a set of minimal distances between each of the m binding sites (in the entire dataset) and the corresponding n predictions for each of the prediction methods. We assume that the predictions from different methods that are farther than 10Å away from the site are equally wrong, i.e., they are too far away to be meaningful, and thus we round them down to 10Å. The significance of the differences between a given pair of predictors was measured by evaluating the corresponding, for the same m, minimal distance values. Since the distances for the considered predictors are not normally distributed, per the Shapiro-Wilk test of normality at $p = 0.05$, we used the non-parametric Wilcoxon signed-rank test. We assume that the differences are significant if $p < 0.05$. Statistical analysis shows that the Findsite is significantly better than all other methods, see Table 5-1, where the "+"/"–" indicates that a method in a given column is significantly better/worse than a methods in a given row with $p < 0.05$ and "=" denotes that a given pair of methods is not significantly different. The ConCavity, Q-SiteFinder, MetaPocket and PocketPicker are second-best and not significantly different between each other (except for the ConCavity which significantly improves over the Q-SiteFinder), and this group is statistically better than LIGSITE^{csc}, SURFNET, PASS, PocketFinder and Fpocket.

Table 5-1: Statistical significance of the differences in distances measured using D_{CC} between the predicted and the actual location of the binding site for all pair of the considered ten prediction methods measured using Wilcoxon signed-rank test.

	PASS	SURFNET	Pocket-Finder	Fpocket	LIGSITE^{csc}	Q-SiteFinder	Pocket-Picker	Meta-Pocket	Con-Cavity	Findsite
PASS		=	+	+	+	+	+	+	+	+
SURFNET	=		+	+	+	+	+	+	+	+
PocketFinder	–	–		=	+	+	+	+	+	+
Fpocket	–	–	=		=	+	+	+	+	+
LIGSITE^{csc}	–	–	–	=		+	+	+	+	+
Q-SiteFinder	–	–	–	–	–		=	=	+	+
PocketPicker	–	–	–	–	–	=		=	=	+
MetaPocket	–	–	–	–	–	=	=		=	+
ConCavity	–	–	–	–	–	–	=	=		+
Findsite	–	–	–	–	–	–	–	–	–	

5.5.3 The impact of structural similarity between predicted protein and template library

Findsite is a threading-based method that utilizes a library of template structures. Its predictions are generated by clustering binding sites of the template structures and they rely on the availability of templates that are similar to the predicted protein. To study the impact of the availability of similar templates, we use a threshold to limit the structural similarity between the predicted proteins and the templates used for the prediction. Only the template structures with a similarity score below the threshold are utilized. The structural similarity is measured with TM-score, which varies between 0 and 1 (Zhang and Skolnick, 2005). We vary the threshold between 0.5 and 1 with step of 0.1. Findsite also utilizes a default cut-off TM-score = 0.4 below which a given template is rejected. In case if Findsite does not find a suitable template above the 0.4 cut-off, we lower it by 0.1 until a template is found.

We compare Findsite with the Q-SiteFinder, which is the only energy-based method, the ConCavity, a representative (best-performing) geometry-based method, and with the MetaPocket that represents the consensus-based approaches. The success rates of these four methods are quantified using D_{CC}. For Findsite, we generate six sets of predictions that correspond to the consecutive values, between 0.5 and 1, of the maximal similarity threshold. The MetaPocket, ConCavity, and Q-SiteFinder do not utilize templates thereby they have one set of predictions. Figure 5-4 reveals, as expected, that the predictive quality of Findsite improves with the increase of the similarity threshold. For the cutoff D = 4Å, its success rates equal 16%, 25%, 34%, 37%, 43% and 57% for the maximal TM-score threshold of 0.5, 0.6, 0.7, 0.8, 0.9, and 1, respectively. This indicates that the predictive quality of Findsite is largely dependent on the availability of structurally similar templates. We investigate significance of differences in the prediction quality measured with D_{CC} between the predictions generated by the four methods. Findsite is significantly better than the other three methods when the similarity threshold is 0.7 or higher, comparable to the other three methods

when the threshold is set to 0.6, and significantly inferior for the threshold equal 0.5. These results suggest that if Findsite identifies a template that shares a TM-score ≥ 0.7 with the query protein, then its predictions are expected to be better than the predictions of the ConCavity, MetaPocket and Q-SiteFinder. On the other hand, if the maximal TM-score between the Findsite's templates and the query protein ≤ 0.5, then the predictions generated by the ConCavity, MetaPocket or Q-SiteFinder are likely to be better.

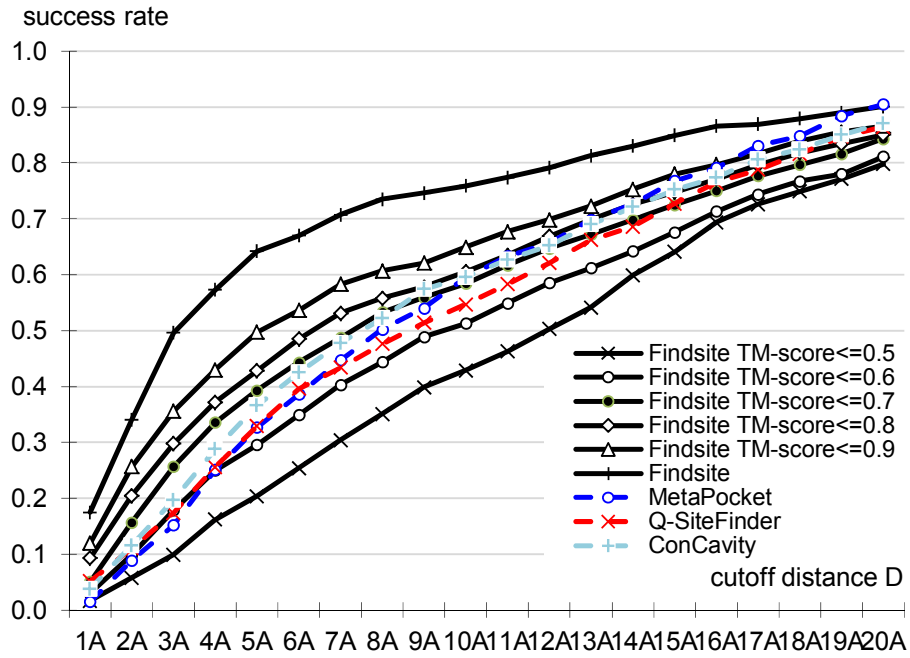


Figure 5-4: Comparison of the success rates (y-axis) of Findsite using its entire template library measured using D_{CC} for different cutoff distances D (x-axis) on the benchmark dataset with the predictions where the maximal structural similarly between a query protein and the templates limited to TM-score ≤ 0.9, ≤ 0.8, ≤ 0.7, ≤ 0.6, and ≤ 0.5. The figure also includes the success rates for Meta-pocket, ConCavity, and Q-SiteFinder.

5.5.4 Comparison of prediction quality between Apo and Holo structures

The benchmark dataset consists of holo structures, i.e., structures that are bound to ligands. Since the protein-ligand interactions may lead to conformational changes at the vicinity of the binding site, we also investigate the binding site

predictions performed on the apo structures, i.e., unbound-state proteins. We select a subset of proteins, for which both apo structures and holo structures are known, from the benchmark dataset. This results in two datasets, D_{Apo} that includes 104 of these apo structures and D_{Holo} that includes the corresponding set of the 104 holo structures (a subset of the benchmark dataset).

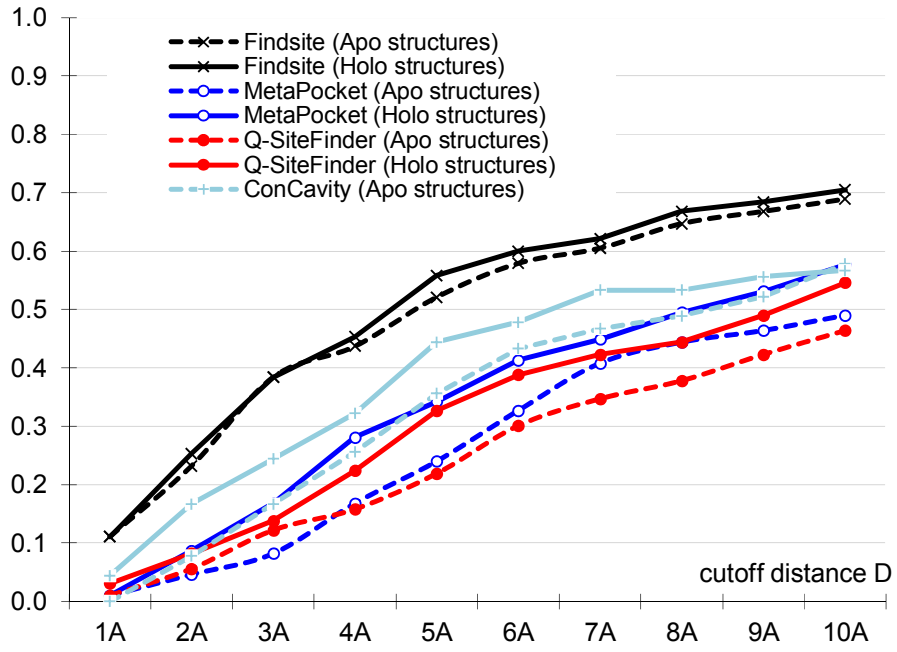


Figure 5-5: Comparison of success rates (y-axis) on the D_{Holo} and D_{Apo} datasets measured using D_{CC} for Findsite, MetaPocket, ConCavity, and Q-SiteFinder. The two datasets contains structures of the same set of proteins where D_{Holo} includes ligand-bound structures and D_{Apo} includes structures at the ligand-unbound state. The x-axis shows the cutoff distance D that is used to calculate the success rates.

We assess predictions generated by the four representative methods, the threading-based Findsite, the energy-based Q-SiteFinder, the consensus-based MetaPocket, and the best performing geometry-based ConCavity on both datasets; see Figure 5-5. Using the D_{CC} distance, the success rates of Findsite on the D_{Holo} dataset is on average (over different thresholds) about 1.6% higher than on the D_{Apo} dataset. For the MetaPocket, Q-SiteFinder, and ConCavity the success rates

on the D_{Holo} dataset are on average 6.7%, 6.2%, and 7.3% higher than on the ligand-unbound dataset. Similar trends are observed when using the D_{CA}, see Figure 5-6. Specifically, Findsite, MetaPocket, Q-SiteFinder, and ConCavity achieve 1.1%, 6.7%, 7.5%, and 6.9% better success rates on the D_{Holo} dataset, respectively. The significance of the differences in the predictive quality between the D_{Holo} and D_{Apo} datasets was calculated using the Wilcoxon signed-rank test. The test reveals that MetaPocket, ConCavity and Q-SiteFinder, achieve significantly better predictions with $p < 0.01$, $p < 0.01$ and $p < 0.05$, respectively, on the D_{Holo} dataset when compared with the D_{Apo} dataset, while Findsite achieves comparable results on both datasets. These results suggest that the geometry-, energy-, and consensus-based methods benefit from the usage of the holo structures, likely because the geometrical descriptors and the energy function can be calculated more accurately using these structures.

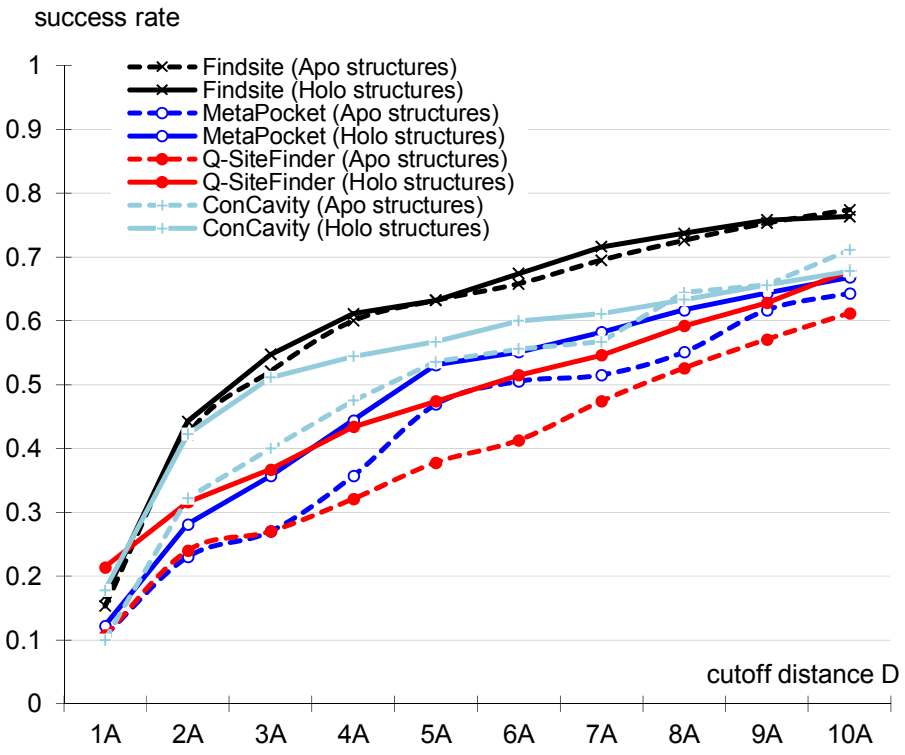


Figure 5-6: Comparison of the success rates (y-axis) on the D_{Holo} and D_{Apo} datasets measured using D_{CA} for the threading-based Findsite, the energy-based Q-

SiteFinder, the best performing geometry-based ConCavity, and the consensus-based MetaPocket. The two datasets include structures from the same set of proteins where D_{Holo} is composed of structures in the ligand-bound state and D_{Apo} in the ligand-unbound state. A given binding site is regarded as correctly predicted if the minimal distance between this site and the top n predictions is below the cutoff distance D (x-axis), where n is the number of binding sites of the protein that includes the evaluated binding site.

5.5.5 Impact of the size of the binding sites

We assessed the impact of the size of the binding sites on the predictive quality. The size is approximated by the number of interacting atoms in the binding site. A non-hydrogen atom of a residue is considered as an interacting atom if it is within 3.9Å to a non-hydrogen atom of the ligand (Luscombe et al., 2001). The binding sites that are sorted by their sizes are divided into five subsets with equal number of sites. The success rates of the four representative predictors are calculated using these five subsets. Using D_{CC}, we observe a consistent trend that higher success rates are achieved for the larger binding sites, see Figure 5-7. For instance, the average success rates for Findsite are 23%, 35%, 45%, 57% and 69% for the consecutive five subsets, respectively, when considering cutoff distances D between 1Å and 5Å. Similarly, the average success rates for Q-SiteFinder, MetaPocket, and ConCavity equal 3%, 4%, and 5%; 14%, 12%, and 11%; 22%, 18%, and 22%; 26%, 26%, and 24%; and 33%, 28%, and 34% on the five subsets, respectively. The Pearson correlations between the average success rates, over cutoff distances D between 1Å and 5Å, and the average size of the binding sites in each of the five subsets, see Figure 5-8, equal 0.98 for Q-SiteFinder and MetaPocket and 0.99 for Findsite and ConCavity. This shows that the predictive quality of these four methods is linearly correlated with the size of the binding sites. We measure the ratio between the solvent accessible area of the binding residues, computed with the DSSP program (Kabsch and Sander, 1983), and the protein surface, i.e., the sum of the solvent accessible area of all residues, for each protein. The average ratios in each the five subsets are similar and they vary between 0.085 and 0.105. This shows that the improved success rates are not due

to the larger binding areas, but rather due to inherent characteristics of these predictive models.

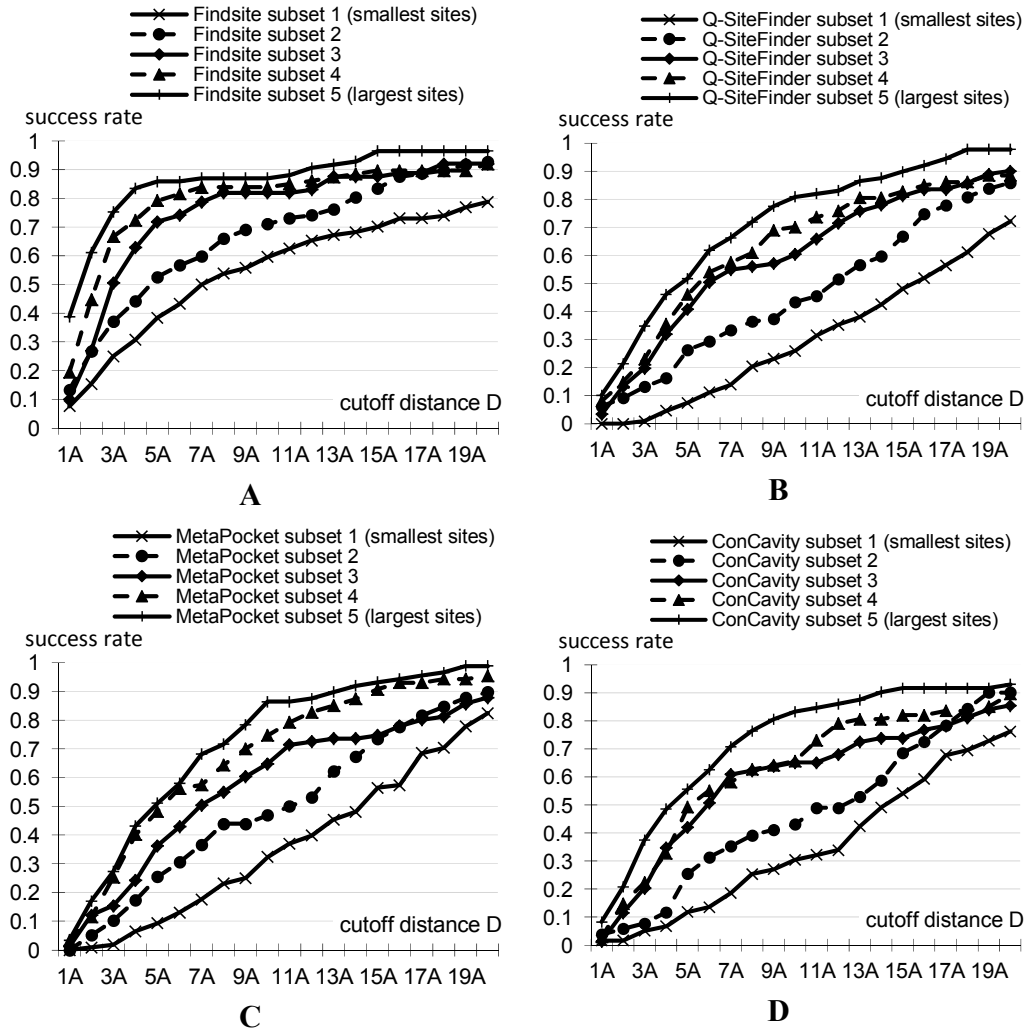


Figure 5-7: Success rates (*y*-axis) measured using D_{CC} for different cutoff distances *D* (*x*-axis) for A) Findsite, B) Q-SiteFinder, C) MetaPocket, and D) ConCavity as a function of the size of the binding site, which is approximated by the number of interacting atoms. The binding sites in the benchmark dataset are sorted by their sizes in the ascending order and they are binned into five equally sized subsets. Each line corresponds to the results on one of these subsets, where subset 1 includes the smallest sites and subset 5 the largest sites.

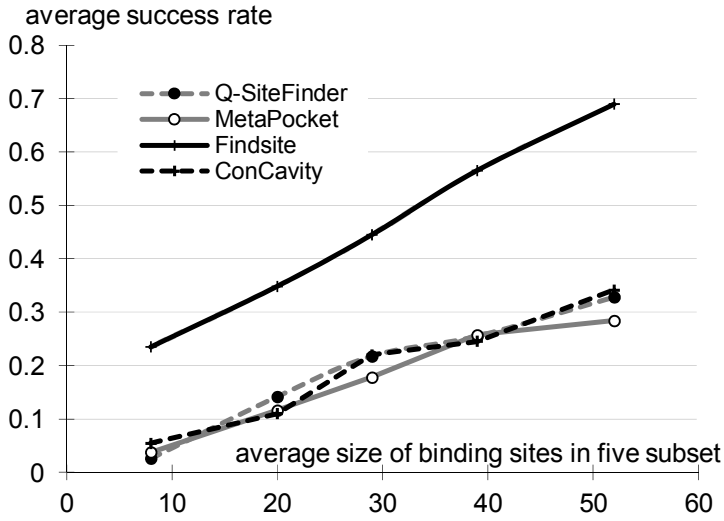


Figure 5-8: Relation between the average, over cutoff distances D between 1Å and 5Å, success rates (y-axis) measured using D_{CC} and the size of the binding sites for Findsite, Q-SiteFinder ConCavity, and MetaPocket. The binding sites in the benchmark dataset are sorted by their sizes, which are approximated by the number of interacting atoms, in the ascending order and they are binned into five equally sized subsets. The x-axis shows the average size of the binding sites for the five consecutive subsets.

5.5.6 Predictive quality for different ligand groups

The benchmark dataset includes 475 biologically relevant ligands that are categorized into 253 types. We manually inspected the ligands that occur in the dataset at least 3 times and we grouped them into four categories, including acids, carbohydrates, mononucleotides and cofactors (excluding mononucleotides). The breakdown of the ligand types in each category is given in Table 5-2. These ligands occur 219 times in the benchmark dataset and they cover 46% of all binding sites; see Figure 5-9. The remaining ligands are more unique and could not be clustered into sets that would allow for a statistically sound evaluation of the predictive quality.

Table 5-2: List of ligand types in the four considered major ligand categories.

Ligand category	3-letter abbreviation of ligand name	Formula
Acids	BEZ	$C_7 H_6 O_2$
	SAH	$C_{14} H_{20} N_6 O_5 S$
	ASP	$C_4 H_7 N O_4$
	EPE	$C_8 H_{18} N_2 O_4 S$
	GLU	$C_5 H_9 N O_4$
	MYR	$C_{14} H_{28} O_2$
	PEB	$C_{33} H_{40} N_4 O_6$
	TRP	$C_{11} H_{12} N_2 O_2$
Carbohydrates	BGC	$C_6 H_{12} O_6$
	GLC	$C_6 H_{12} O_6$
	FUL	$C_6 H_{12} O_5$
	GAL	$C_6 H_{12} O_6$
	GLA	$C_6 H_{12} O_6$
	XYP	$C_6 H_{10} O_5$
	MAN	$C_6 H_{12} O_6$
	FUC	$C_6 H_{12} O_5$
	NAG	$C_8 H_{15} N O_6$
Mononucleotides	2GP	$C_{10} H_{14} N_5 O_8 P$
	5GP	$C_{10} H_{14} N_5 O_8 P$
	A2P	$C_{10} H_{15} N_5 O_{10} P_2$
	A6P	$C_6 H_{13} O_9 P$
	ADP	$C_{10} H_{15} N_5 O_{10} P_2$
	AGP	$C_6 H_{16} N O_8 P$
	AMP	$C_{10} H_{14} N_5 O_7 P$
	AMZ	$C_9 H_{15} N_4 O_8 P$
	ANP	$C_{10} H_{17} N_6 O_{12} P_3$
	ATP	$C_{10} H_{16} N_5 O_{13} P_3$
	C5P	$C_9 H_{14} N_3 O_8 P$
	CMP	$C_{10} H_{12} N_5 O_6 P$
	G1P	$C_6 H_{13} O_9 P$
	GDP	$C_{10} H_{15} N_5 O_{11} P_2$
	GTP	$C_{10} H_{16} N_5 O_{14} P_3$
	NOS	$C_{10} H_{12} N_4 O_5$
	PAP	$C_{10} H_{16} N_5 O_{13} P_3$
	TMP	$C_{10} H_{15} N_2 O_8 P$
	TPP	$C_{12} H_{19} N_4 O_7 P_2 S$
	U5P	$C_9 H_{13} N_2 O_9 P$
	UDP	$C_9 H_{14} N_2 O_{12} P_2$
	UMP	$C_9 H_{13} N_2 O_8 P$
	UTP	$C_9 H_{15} N_2 O_{15} P_3$
Cofactors	ACO	$C_{23} H_{38} N_7 O_{17} P_3 S$
	COA	$C_{21} H_{36} N_7 O_{16} P_3 S$

	FAD	$C_{27}\ H_{33}\ N_9\ O_{15}\ P_2$
	FMN	$C_{17}\ H_{21}\ N_4\ O_9\ P$
	HEM	$C_{34}\ H_{32}\ Fe\ N_4\ O_4$
	NAD	$C_{21}\ H_{27}\ N_7\ O_{14}\ P_2$
	NAP	$C_{21}\ H_{28}\ N_7\ O_{17}\ P_3$
	PLP	$C_8\ H_{10}\ N\ O_6\ P$
	PQQ	$C_{14}\ H_6\ N_2\ O_8$
	SAM	$C_{15}\ H_{22}\ N_6\ O_5\ S$
	U2G	$C_{19}\ H_{24}\ N_7\ O_{13}\ P$

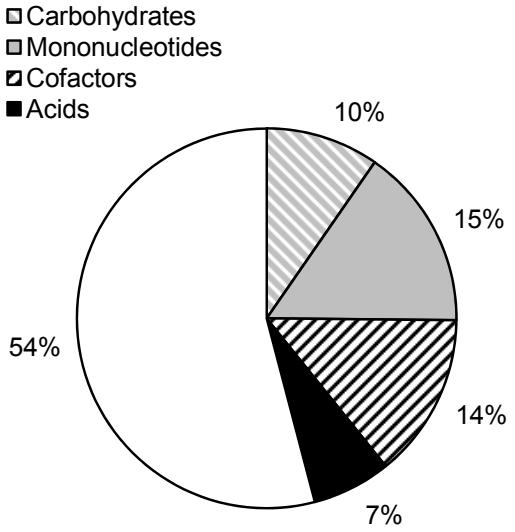
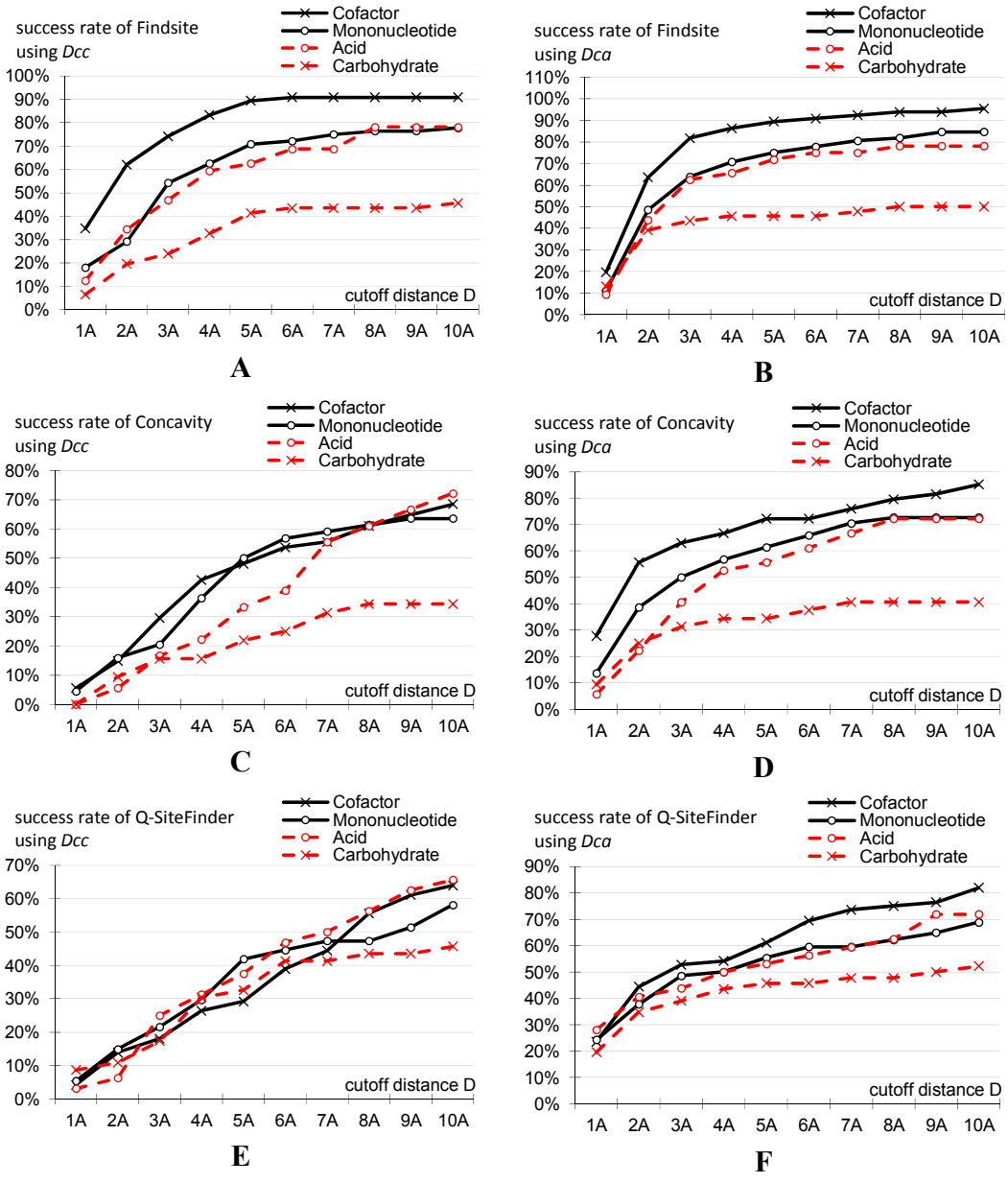


Figure 5-9: The rate of occurrence of the four major ligand groups, which include acids, carbohydrates, mononucleotides and cofactors (excluding mononucleotides) in the benchmark dataset. These four groups cover 46% of all ligands in the dataset.

We compare the success rates of the four representative prediction methods on the four ligand categories. Using the D_{CC} measure, the Findsite and ConCavity achieve the highest success rates for the cofactors, followed by the mononucleotides and acids, and the lower accuracies for the carbohydrates; see panels A and C in Figure 5-10. These differences are quite substantial, e.g. at D = 4 Å the success rates for cofactors and carbohydrates differ by 50%. In contrast, the differences between the success rates for different ligand groups for the Q-SiteFinder and MetaPocket are relatively minor; see panels E and G in Figure 5-10. Similar trends are observed when using D_{CA}; see panels B, D, F, and H in

Figure 5-10. The above suggests that the predictions generated by Q-SiteFinder and MetaPocket are not sensitive to the ligand types, while the predictive quality of Findsite and ConCavity varies relatively widely between different ligand groups.



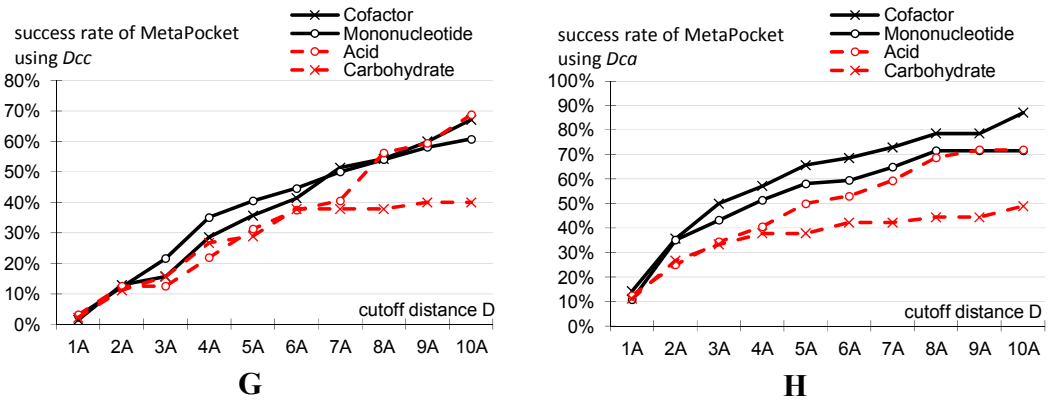


Figure 5-10: Comparison of the success rates (y-axis) for prediction of binding sites for four categories of ligands including acids, carbohydrates, mononucleotides and cofactors measured using D_{CC} (panels on the left) and D_{CA} (panels on the right). The x-axis shows the cutoff distance D used to calculate the success rates. A) results of Findsite measured using D_{CC}; B) results of Findsite measured using D_{CA}; C) results of ConCavity measured using D_{CC}; D) results of ConCavity measured using D_{CA}; E) Results of Q-SiteFinder measured using D_{CC}; F) results of Q-SiteFinder measured using D_{CA}; G) results of MetaPocket measured using D_{CC}; H) results of MetaPocket measured using D_{CA}.

5.5.7 Complementarity of predictors

The four representative methods are based on different approaches, i.e., Findsite uses threading, Q-SiteFinder is based on the energy calculations, ConCavity utilizes geometrical descriptors, and MetaPocket combines geometrical descriptors and energy calculations. We investigate whether these differences result in complementarity in their predictions. A given binding site is regarded as covered by a combination of several methods if it is correctly predicted by any of these methods. Figure 5-11 demonstrates that combining predictions of the best performing Findsite with the other three methods results in a larger coverage. For the thresholds D between 1Å and 10Å, the coverage when using the four methods together increases between 4% and 10% when compared with the predictions of the Findsite. For the cutoff distance D = 4Å, 7% of binding sites that are not captured by the Findsite are successfully predicted by the Q-SiteFinder and 10% of the sites that are missed by the Findsite are correctly predicted by one of the

three other methods. This shows that the four methods are complementary, which implies that they could be combined to build a consensus-based method.

We developed a simple consensus predictor by re-ranking the predictions generated by Findsite using the predictions from Q-SiteFinder, ConCavity, and MetaPocket. This solution, in contrast to a straightforward merging of the predictions from the three methods, is motivated by overall high predictive quality of Findsite, when compared to the runner-up approaches. Moreover, we observe that for a protein with n binding sites, Findsite sometimes generates more than n predictions and some of the correct predictions are not ranked among the top n outputs. Predictions generated by Findsite are scored by comparing them to the predictions generated by the other three methods to improve the ranking. A Findsite's prediction receives score of 3 if it is within 4Å to the predictions from Q-SiteFinder, MetaPocket, and ConCavity. The score equals 2 if the Findsite's prediction is within 4Å to the predictions of the two other methods. The score of 1 corresponds to the case when the Findsite's prediction is within 4Å to a prediction from one of the other three methods, and the score equals 0 if the other three methods did not generate predictions within the 4Å radius. The predictions are sorted in the descending order by their scores, and ties are resolved by using the original order of the predictions from Findsite. The solid line in Figure 5-11 reveals that the re-ranking improves the success rates of the original Findsite. When considering the cutoff distances D between 1Å and 5Å, the re-ranked predictions improve over the original Findsite on average by 2%. Although the magnitude of these improvements is relatively small, the Wilcoxon signed-rank test at the 0.05 significance level shows that they are statistically significant. This means that the distances between the native and the predicted positions of the ligand are consistently smaller when using our consensus approach. These preliminary results suggest that these four methods generate complementary predictions, and they motivate further research on the ensemble-based predictors.

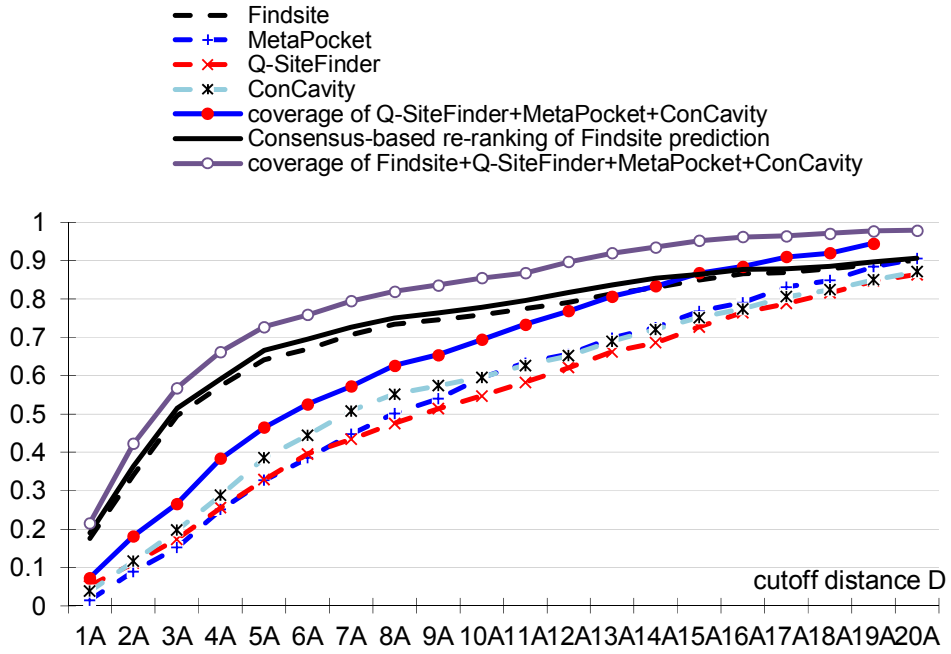


Figure 5-11: The success rates (*y*-axis) of the Findsite, Q-SiteFinder, ConCavity, MetaPocket and a consensus-based method measured using D_{CC} on the benchmark dataset compared to the coverage of the binding sites predicted by combination of the four methods. The *x*-axis shows the cutoff distance *D* that is used to calculate the success rates and the dashed line shows the success rates of the consensus-based re-ranking of the Findsite predictions.

5.5.8 Case studies

We use the chain A of Bcr-Abl protein (PDB code: 3K5V) (Zhang *et al*, 2010) and M2 proton channel of influenza A virus (PDB code: 2RLF) (Schnell and Chou, 2008) to demonstrate the utility of the four representative binding site predictors. These proteins are not included in our benchmark dataset and are subject to recent studies to reveal the atomic-level insights into their binding interactions (Zhang *et al*, 2010; Schnell and Chou, 2008). We superimpose the above two structures with other Bcr-Abl and M2 proton channels structures in the PDB, respectively, using Fr-TM-align (Pandit and Skolnick, 2008). This is performed to assure a complete (to date) annotation of the native binding sites. As a result, both proteins are annotated with two binding sites. We use the web servers of Findsite, MetaPocket, ConCavity, and Q-SiteFinder to generate the

predictions. The two structures with the ligands shown in black and the predictions from Findsite, ConCavity, MetaPocket and Q-SiteFinder that are colored green, pink, red and blue, respectively, are given in Figure 5-12.

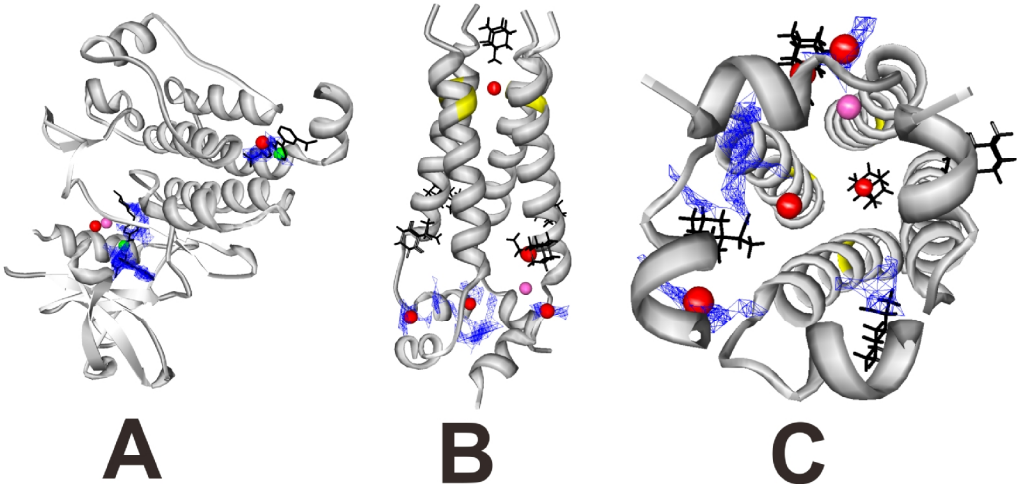


Figure 5-12: The binding sites predicted by Findsite, MetaPocket, ConCavity, and Q-SiteFinder for chain A of the Bcr-Abl protein (panel A) and the M2 proton channel (panels B and C show the side and the top views, respectively). The predictions by Findsite, MetaPocket, ConCavity, and Q-SiteFinder are denoted with green, red, and pink spheres and blue mesh, respectively. The Q-SiteFinder predicted grid points of the pocket are shown using the mesh. The ligands are in the stick format and are colored in black. The M2 proton channel consists of 4 chains and has 5 binding sites. Each of the 4 chains is annotated with 2 sites, where the site at the center of the channel is common to all of them. The other 4 sites are symmetrically distributed at the lipid-facing side of the four chains. The key interacting residues for the central binding site, Ser31, on these four chains are colored in yellow in panels B and C.

For the Bcr-Abl protein, we evaluated the top 2 predictions from each predictor since this protein has two binding sites. Both of these sites are predicted correctly by Findsite and Q-SiteFinder, see panel A in Figure 5-12. The distances between the predicted site and the center of the ligand are 1Å and 2Å for the Findsite and 1Å and 3Å for the Q-SiteFinder. The Q-SiteFinder predicts the grid points of the binding sites, which have more than 40% overlap, measured using O_{PL}, with the ligands. The predictions by the MetaPocket are less accurate; its D_{CC} for the two binding sites equals 6Å and 2Å. ConCavity generates one prediction for this

structure with the D_{CC} equal 5Å. The pocket identified by ConCavity is not shown in panel A of Figure 5-12 because it would obstruct predictions from the other methods; this pocket is visualized in the panel A of Figure 5-13. We note that these two sites are biologically relevant; a recent study has shown that inhibitors that bind to these two sites lead to the inhibition of Bcr–Abl activity (Zhang *et al*, 2010).

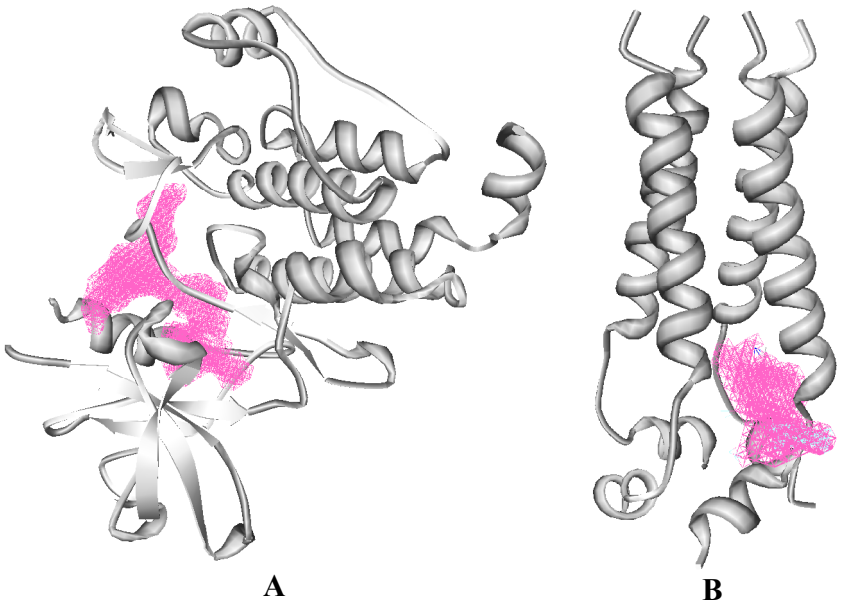


Figure 5-13: The pockets identified by the ConCavity, denoted by a pink mesh, for A) chain A of the Bcr-Abl protein; B) M2 proton channel.

The binding sites on the M2 proton channel of influenza A virus have recently attracted significant attention since a class of antiviral drugs, such as adamantane M2 inhibitors, interacts with this channel. The structure of the M2 proton channel in complex with inhibitors was solved in 2008 by two groups which proposed two distinct binding sites (Schnell and Chou, 2008; Stouffer, et al., 2008). A recent study confirmed that Adamantane and its derivatives are capable of interacting with both binding sites (Rosenberg and Casarotto, 2010). The sites on the M2 proton channel are difficult to predict due to two facts: 1) the channel is formed by 4 protein chains while some predictors, including Findsite, are designed to predict using a single chain; and 2) the binding sites are located in the transmembrane regions (Rosenberg and Casarotto, 2010) while most of the

complexes used to develop the binding site predictors concern globular proteins. Each of the four chains has two sites. The site located at the center of the channel is common to all four chains and the other sites are symmetrically distributed at the lipid-facing side of the four chains. As a result, this protein complex has total of five binding sites and thus we evaluated the top five predictions generated by each of the four prediction methods. The predicted binding sites and the ligands are shown in panel B (side view) and C (top view) in Figure 5-12. The binding site at the pore of the channel is predicted only by the MetaPocket. Although the distance between the predicted site and center of the ligand is around 6Å, the predicted site is at the center of four key binding residues (Ser31 on the four chains), which are depicted in yellow in panels B and C in Figure 5-12. The other sites, which are targeted by Rimantadine, are located at the base of the transmembrane helix on each of the chains. Only one of these sites is correctly predicted by the MetaPocket, and none of the top five predictions by Q-SiteFinder is close to the ligand ($D_{CC} >$ 8Å). The ConCavity predicts one pocket, which is shown in panel B of Figure 5-13, and this prediction relatively far from the actual site ($D_{CC} >$ 8Å). We note that Findsite did not generate predictions for the M2 proton channel due to the unavailability of suitable templates. Overall, we conclude that majority of the considered binding sites are found by at least one of the top four methods, which suggests that they provide useful inputs for the atomic-level discovery of protein-ligand interactions.

5.6 Conclusions

In this chapter, we empirically compare ten structure-based binding site predictors which were developed in the past decade and which offer either a web-server or a standalone program to generate predictions. The more recent methods including Findsite, Q-SiteFinder, ConCavity, and MetaPocket are shown to provide significant improvements over the older solutions. This indicates that progress was made over the last several years. However, a considerable fraction of the binding sites is not identified by any of the considered methods, which motivates

further research in this area. For instance, at a cutoff of 4Å and using the D_{CC} measure, about 33% of the binding sites are missed by the four best-performing methods. We demonstrate that the quality of the predictions is strongly positively correlated with the size of the binding sites. We also show that although Findsite is significantly more accurate than the other considered predictors and is more robust when performing predictions using the apo structures, this method is largely dependent on the completeness of its template library. When the maximal TM-score between the predicted protein and the best template identified by Findsite is below 0.5, then certain energy-, structure- and consensus-based predictors are shown to provide more accurate predictions. We developed a simple consensus-based approach that uses four complementary predictors, the threading-based Findsite, the energy-based Q-SiteFinder, the geometry-based ConCavity, and consensus-based MetaPocket. This method is shown to provide success rates improved by 2% when compared with the best performing Findsite. The important conclusions from our survey, which we use in Chapter 7, are that consensus-based predictions are useful and that threading-based approach is promising.

Since the threading-based method works by identifying a known similar fold for a given query protein, the templates that are used in the prediction are restricted to one structural fold. However, a recent study shows that conserved sugar-binding and aromatic-group binding fragments are found across multiple protein folds (Petrey et al., 2009). The phosphate-binding fragment that occurs in dozens of protein families was discovered already two decades ago (Saraste et al., 1990). This means that the approach taken by the Findsite may not work in these and related cases and it motivates further research. One of the potential solutions would be to develop a measure of similarity between surface patches on the query protein and the surfaces of the known binding sites. By comparing relevant "sub-structures" (fragments of the fold concerning the binding sites), the above approach could overcome the constraint on the similarity of the overall fold

between the query and the template structures. This approach is explored in Chapter 7.

CHAPTER 6 Prediction of nucleotide binding residues from protein sequence

6.1 Introduction

We surveyed the performance of existing structure-based binding site predictors in the previous chapter. We also evaluated the predictive quality of these methods on different compound groups, i.e., acids, carbohydrates, nucleotides, and cofactors. Among these compound groups, we focus on the nucleotides due to the following two reasons. First, nucleotides play important roles in a number of biological processes, as we discusses in Section 2.1.4. For instance, they are structural units of nucleic acid chains, they serve as sources for chemical energy, participate in the cellular signaling, and they are involved in the enzymatic reactions. Second, nucleotides interact with a wide range of proteins. As of June 2010, 5293 proteins in the PDB are annotated as "nucleotide binding", which means that nucleotides constitute about 15% of biologically relevant ligands included in this database (Dessailly *et al*, 2008; Goto *et al*., 2002). This demonstrates the ubiquity and the substantial interest in the protein-nucleotide interactions. Although a substantial effort in identification and characterization of the nucleotide binding sites was observed in the past two decades, most of these approaches are based on the analysis of known nucleotide-binding sequences and structures, which were used to identify conserved motifs in protein sequences and structures. For instance the Walker A and B sequence motifs were identified for the adenine nucleotide binding proteins (Walker *et al*., 1982). A fuzzy recognition template was proposed for the characterization of the adenylate-protein interactions (Moodie *et al*., 1996). The Johnson motif was reported to cover one-third of the adenine mononucleotide-binding proteins (Denessiouk and Johnson, 2000). The abovementioned studies characterize the sequence and structural motifs for a relatively narrow range of the nucleotide-protein interactions, usually only for a selected interaction mode for a single nucleotide type, or they require tertiary protein structures as the input, which substantially limits their utility. On

the other hand, the large number of protein chains with uncharacterized tertiary structures motivates the development of computational tools for high-throughput sequence-based annotation of the nucleotide-binding residues for a wide range of the nucleotides. To this end, in this chapter we introduce a method that predicts the nucleotide-binding residues from protein sequence.

6.2 Related work

Currently there is no method that predicts the binding residues from the protein sequence for a comprehensive set of nucleotides, and only recently two methods that predict the ATP-binding and GTP-binding residues were proposed (Chauhan *et al*., 2010; Chauhan *et al*., 2009). These two methods input information extracted from the sequence and the corresponding sequence profile using a window centered on the predicted residue into a machine learning classifier that predicts propensity of this residue to interact with the ATP or GDP. The two methods generate a real value that quantifies the probability of binding to ATP or GTP for each residue. They were implemented as ATPint and GTPbinder web-servers and are available at www.imtech.res.in/raghava/atpint/ and www.imtech.res.in/raghava/gtpbinder/, respectively.

6.3 Problem definition

Our method predicts whether a given amino acid in the input protein sequence is involved in the interaction with a given nucleotide type. Similar to the annotation of the DNA-binding and small ligand-binding residues (Chen and Kurgan, 2009; Luscombe *et al*., 2001), a given residue is annotated as "nucleotide binding" if at least one of its non-hydrogen atom is less than 3.9Å away from a non-hydrogen atom of the nucleotide. As suggested in (Luscombe *et al*., 2001), atoms within 3.9Å are considered to interact through the van der Waals contacts. For each residue, our method generates two levels of predictions, (i) the binary value that defines whether a given residue does or does not bind to a given nucleotide type;

and (ii) the real value that quantifies the probability of binding to certain nucleotide type.

6.4 Dataset preparation

The nucleotides that are considered in this study contain at least one of the five nucleobases, a 5-carbon sugar, and 1 to 3 phosphates. We extracted all complexes from PDB that included these nucleotides; we need these structures to obtain annotation of the binding residues to build and evaluate our predictor. The maximal pairwise sequence identity of the resulting protein chains for each of the nucleotides is reduced to 40% with CD-hit (Li and Godzik, 2006). We include the nucleotides with at least 50 chains in the corresponding set. The relatively low identity assures that these nucleotides bind a wide range of protein chains, which makes it challenging to find the binding residues using the sequence alignment. The availability of at least 50 chains provides us with a sufficient amount of annotated binding residues to build and evaluate a well-performing predictor. Three sets of protein sequences are generated for different evaluation purposes:

Dataset 1 includes 227, 321, 140, 56 and 105 chains that were released in PDB before March 10th 2010 and that bind to ATP, ADP, AMP, GTP and GDP, respectively. Dataset 1 includes 4688 ADP-binding, 3393 ATP-binding, 1756 AMP-binding, 853 GTP-binding, and 1577 GDP-binding residues, and 121158, 80409, 44009, 18888, and 36561 non-binding residues, respectively.

Dataset 2 consists of nucleotide-binding chains that were released after March 10th 2010. The maximal pairwise sequence identity in dataset 2 was reduced to 40%. Moreover, if a given chain in Dataset 2 shares above 40% identity to a chain in Dataset 1 and both chains interact with the same nucleotide, then we remove the chain from Dataset 2. This assures that the Dataset 2 is independent of the Dataset 1 and can be used to test models developed using Dataset 1. Consequently, Dataset 2 includes 17, 25, 18, 6, and 9 chains that bind to ATP, ADP, AMP, GTP, and GDP, respectively.

Dataset 3 consists of chains that do not interact with nucleotides, and is used to evaluate whether our method would "overpredict" nucleotide-binding residues. We use the pre-culled list of 1853 PDB chains generated by the PISCES server (Wang and Dunbrack, 2003) at 20% sequence identity, which correspond to high-quality structures with maximal resolution of 1.6Å and maximal R-factor of 0.25. Next, among this set of representative chains, we remove all chains that (potentially) interact with nucleotides. Any chain that shares > 40% identity to any chain in Dataset 1 (which is used to build our predictive model), or which is annotated as nucleotide-binding in the Gene Ontology database (Ashburner *et al*., 2000), or which binds to nucleotides among the depositions in the PDB is removed. As a result, we extracted 1372 chains that do not interact with the nucleotides.

Dataset 1 is used to build and evaluate the prediction models and the results on dataset 1 are based on 5-fold cross validation. Dataset 2 and 3 are used as independent datasets to assess the prediction models that are built utilizing Dataset 1. Dataset 2 is used to assess the performance of the prediction models on unobserved data while dataset 3 is used to evaluate whether the prediction models would "overpredict" nucleotide-binding residues. The datasets are available at http://biomine.ece.ualberta.ca/nSITEpred/.

6.5 Evaluation protocol

6.5.1 Evaluation measures

The binary predictions are assessed using 5 measures:

Precision (PREC) = $TP/(TP+FP)$
Recall (REC) = $TP/(TP+FN)$
Specificity (SPEC) = $TN/(FP+TN)$
Accuracy (ACC) = $(TP+TN)/(TP+FP+TN+FN)$
MCC = $(TP*TN–FP*FN) / \text{sqrt}[(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)]$

where *TP* (true positives) and *TN* (true negatives) are the counts of correctly predicted binding and non-binding residues, respectively, *FP* (false positives) are

non-binding residues that were predicted as binding, and *FN* (false negatives) are binding residues that were predicted as non-binding. The precision, recall, and specificity evaluate quality of predictions for the predicted binding residues, native binding residues, and native non-binding residues, respectively. The Matthews correlation coefficient (MCC) evaluates the overall predictive quality. MCC values are between -1 and 1 with higher values for better predictions; 0 means that all residues are predicted as binding (or non-binding).

The receiver operating characteristic (ROC) curves are used to examine the predicted probabilities. For each value of probability p achieved by a given method (between 0 and 1), all residues with probability $\geq p$ are set as the binding residue and all other residues are set as the non-binding residue. Next, the *TP-rate* = $TP / (TP + FN)$ and the *FP-rate* = $FP / (FP + TN)$ are calculated to draw the ROC curve and we use the area under the curve (AUC) to quantify the predictive quality. Unlike the measures that assess the binary predictions, which depend on the cutoff threshold to define binding/non-binding residues, the AUC value considers all possible thresholds and thus it provides a more comprehensive evaluation.

We analyzed statistical significance of the differences in the MCC and AUC values between predictions generated by our method (for convenience, our method is named as NsitePred in the following text) and the other considered methods. The MCC values are available for all considered methods, while the AUC values cannot be calculated for an alignment-based predictor (described in Section 6.5.2) which provides only a binary annotation. The MCC and AUC values are calculated per sequence (using the cross-validated predictions) for each method and we compare them using a paired test. Since these values are not normal, as tested using Shapiro-Wilk test at the 0.05 significance, we use the non-parametric Wilcoxon rank sum test to measure the differences between the paired MCC (AUC) values calculated for two predictors. We annotate the difference as significant when the p-value < 0.01.

6.5.2 Considered baseline predictors

The NsitePred is compared with the current predictors for the ATP and GDP, ATPint (Chauhan *et al*., 2009) and GTPbinder (Chauhan *et al*., 2010), and 3 baseline predictors based on the residue conservation, sequence alignment, and a simple classifier similar to the methods in (Chauhan *et al*., 2010; Chauhan *et al*., 2009) that uses the evolutionary profile:

- *Rate4site* program (Pupko *et al*., 2002) predicts functional sites by finding conserved residues. We first run psiblast program with the query sequence against the NCBI non-redundant database (Pruitt et al., 2009). For chains with at least 3 significant matches, we created alignments of the best 50 sequences, which is the default for the web version of the rate4site called consurf (Ashkenazy *et al*., 2010), using ClustalW (Larkin *et al*., 2007) and we inputted them into the rate4site. The rate4site generates conservation score for each residue, and the residues with the lower scores, which indicate higher conservation, have a higher probability to be binding residues. We use these scores to compute ROC curves and the corresponding AUC values. We threshold these scores by maximizing the MCC value on the entire dataset to obtain binary predictions. Both the AUC and MCC values are computed separately for each of the five nucleotide types.

- *Sequence alignment using BLAST* identifies similar sequences or segments from an annotated (with the nucleotide binding residues) dataset for a given query sequence. This approach predicts the binding residues by using the nucleotide binding annotations from the best aligned sequence, i.e., sequence with lowest E-value. We execute the BLAST-based alignment between a query sequence and all other sequences (except the query sequence itself) in the dataset for a given nucleotide type. The residues in the query sequence that are aligned with the binding residues on the best aligned chain are predicted as the binding residues.

- *PSSM profile* is widely used in related sequence-based predictors, including ATPint (Chauhan *et al*., 2009) and GTPbinder (Chauhan *et al*., 2010). To validate the effectiveness of the sequence representation proposed in this work, we build a simple predictor that uses SVM (with the same parameters as the corresponding SVM in our NsitePred) and takes PSSM profile as the only input. This allows estimation of the improvement provided by the new features that are used by our method and which exploit conservation scores and predicted secondary structure, relative solvent accessibility, and dihedral angles.

6.6 Proposed solution

6.6.1 Architecture

For a given protein sequence we use PSIPRED (McGuffin *et al*., 2000) to predict the secondary structure, REAL Spine3 (Faraggi *et al*., 2009) to predict the relative solvent accessibility (RSA) and dihedral angles (angles between consecutive amino acids in the protein chain), and BLAST (Altschul *et al*., 1997) to generates the PSSM profile. These inputs together with the sequence are processed using a sliding window to compute a set of numeric features that describe the residue in the center of the window; the features are inputted into a SVM classifier, which outputs probability of nucleotide binding for this residue. This machine learning-based approach is named as SVMPred. Moreover, we run the BLAST-based alignment between the predicted sequence and sequences in the training dataset for a given nucleotide type. The residues in the predicted sequence that are aligned with the binding residues in the best aligned training chain are predicted as the binding residues, i.e., they are assigned with probability equal 1 while the other residues are assigned with probability equal 0. The proposed NsitePred method implements a consensus of SVMPred and the alignment-based predictor by averaging the probabilities generated by SVMPred and the alignment-based predictor.

6.6.2 Feature-based sequence representation

The SVMPred utilizes both sequence and predicted structural descriptors, including the secondary structure, dihedral angles, and RSA, to generate features. We utilize a sliding window of size 17 centered on the predicted residue to extract the features, which include

- *Predicted secondary structure* generated by PSIPRED (McGuffin *et al*., 2000). We use probabilities of the 3 secondary structure states (helix/strand/coil) for each residue in the window (total of 3×17 = 51 features).

- *Predicted relative solvent accessibility* generated by Real-SPINE3 (Faraggi *et al*., 2009). We use the real values, which quantify the fraction of the surface area of a given residue that is accessible to the solvent, for each residue in the window (total of 17 features).

- *Predicted dihedral angles* generated by Real-SPINE3 (Faraggi *et al*., 2009). We utilize two real values, which represent *phi* (involving the backbone atoms C'-N-C^{α}-C') and *psi* (involving the backbone atoms N-C^{α}-C'-N) angles, for each residue in the window (total of 2×17 = 34 features).

- *PSSM profile* generated by blastpgp program (Altschul *et al*., 1997) with default parameters using the NCBI non-redundant database. We normalize these inputs with $1/(1+2^{-x})$, where x is the raw value from the PSSM profile; this transformation is commonly used in the secondary structure prediction. For a window centered at R_i residue at i^{th} position, we calculate 17×20 features $f_{i+k,j}$ where k = −8, −7,…, 7, 8 is the index of the position in the window and j = 1, 2,…, 20 is the index of the PSSM column. We average the values to the left and to the right of the central residue $g_{i+z,j} = (f_{i+z,j} + f_{i-z,j})/2$ where z = 0, 1,…, 8. As a result, the original 17×20 values are transformed to 9×20 values (total of 9×20 = 180 features).

- *AA groups* including hydrophobic residues (Ala, Cys, Ile, Leu, Met and Val), negatively charged (Asp and Glu), positively charged (His, Lys, Arg) and carboxamide-containing amino acids (Asn and Gln), are used to aggregate the normalized and averaged 9×20 PSSM values. The PSSM values for the AA types from a given group for a given position $z = 0, 1,\ldots, 8$ are averaged (total of 9×4 = 36 features).

- *Terminus indicator* is set to 1 for the first and the last 3 residues in the sequence, and it equals to 0 for the other positions (total of 17 features).

- *Secondary structure segment indicators* for helix/strand/coil predictions from PSIPRED on both sides of the window are calculated. If at least 4 / 3 consecutive residues on the left / right side of the window are predicted as helix (strand), then we set the helix (strand) indicator as 1 for the left/right side. If helix and strand indicators equal 0, then the coil indicator is set as 1 (total of 3(helix/strand/coil)×2(left/right) = 6 features).

- *Residue conservation scores* are calculated using the PSSM values for each position based on the Shannon entropy, and based on using two formulas proposed in (Capra and Singh, 2007; Wang and Samudrala, 2006) which incorporate the background frequency of the amino acids (total of 3×17 = 51 features).

- *Collocation of AA pairs* is calculated for the residues in the window. This is motivated by results for the membrane proteins where certain amino acid pairs are over-represented (Senes et al., 2000). Similarly, several sequence motifs occur frequently in the nucleotide-binding sites. To accommodate for mutations in these motifs, we use collocated AA pairs (pairs with gaps) to characterize these motifs. We only consider pairs formed between the central residue in the window and another residue up to 5 positions away. This results in 20×20×10=4000 frequencies (for 20 AA types and 10 positions; 5 on each side). The same as in (Senes et al., 2000), we calculated *p*-values

that indicate the significance of the association between a given amino acid pair and the nucleotide binding annotations. A low *p*-value indicates a low probability that the association between the corresponding amino acid pair and the nucleotide binding annotations is a coincidence. When analyzing 4000 randomly distributed variables, we expect to observe by chance one instance of a difference from expected value with significance $p < 0.00025$ (1/4000). We exclude the amino acid pairs with $p \geq 10^{-6}$, since based on the Engelman's study (Senes et al., 2000) their association with nucleotide-binding event would be random.

We note that the abovementioned features, except for the PSSM profile, were never before used for the prediction of the nucleotide-binding residues.

6.6.3 Feature selection and parameterization

The same features, except for the collocated amino acids pairs are considered to predict binding residues for each of the five nucleotides. Some of these features may not be relevant to the prediction of the nucleotide binding residues and they could be also redundant (correlated) with each other. Therefore, we perform feature selection to remove the irrelevant and redundant features. The selection is performed using the 5-fold cross validation separately for each of the five nucleotide types. First, the biserial correlation (Tate, 1954) between each of the features and the binary annotation of the binding residues is calculated for each of the 5 training sets. The averaged, over the 5 training sets, correlation values are used to rank the features. We use a wrapper-based feature selection with the forward best first search. More specifically, for a given list of feature $F = [f_i$ where $i = 1, 2, \ldots, n]$ sorted in the descending order by their average biserial correlation and an empty list *S* that stores the selected features, we add the top-ranked feature from *F* to *S* and run a linear SVM (Fan *et al*., 2008; Fan *et al*., 2005) with default parameters (i.e., linear kernel and complexity constant $C = 1$) using feature set *S* in the cross validation regime. If the addition of the top-ranked feature improves the average AUC value over the 5 test folds, then this feature is retained in *S*;

otherwise it is removed. We repeat that until F is empty, i.e., we scan the entire feature set once. Next, the SVM classifier is parameterized on the selected feature set. We considered the polynomial and the Radial Basis Function (RBF) kernels. For the polynomial kernel, the complexity constant C is initially fixed at 1 and the degree of the polynomial is adjusted between 0.5 and 5 with step = 0.5. The degree that results in the highest cross-validated AUC value is selected, and next we adjust C using consecutive powers of 2 between 2^{-3} and 2^{5}. Similarly for the RBF kernel, the *gamma* parameter is first optimized using the 2^{-7} to 2^{3} range when C is fixed at 1, and next C is adjusted using the 2^{-3} to 2^{5} range. We select the parameters that maximize the cross-validated AUC and we perform a separate parameterization for each of the five nucleotide types.

6.7 Results

6.7.1 Comparison between NsitePred and existing methods

Table 6-1 compares the NsitePred with the ATPint, GTPbinder and the three baseline predictors based on the alignment, conservation scoring, and evolutionary profiles. We use the tripeptide-based GTPbinder, which outperforms the single-residue and dipeptide-based versions (Chauhan *et al*., 2010), in two configurations including the GTPbinder_PSSM which utilizes PSSM profiles and the GTPbinder_seq which is based solely on the protein sequence.

Table 6-1: Comparison of the quality of the sequence-based prediction of the ATP, ADP, AMP, GDT, and GTP-binding residues between the NsitePred and the related predictors of nucleotide binding residues, including ATPint and GTPbinder that predict ATP-binding and GTP-binding residues, respectively, and predictors based on the alignment (utilizing BLAST), conservation scoring (utilizing Rate4site), evolutionary profiles (utilizing PSSM and SVM classifier), and SVMPred (utilizing the same feature representation as NsitePred and SVM classifier) on Dataset 1. We report the average values over the 5-folds cross validation. The highest values for each ligand type and each quality index, including AUC, precision (PREC), recall (REC), specificity (SPEC), accuracy (ACC), and MCC, are set in bold. The significance of the differences between NsitePred and the other methods are measured for the AUC and MCC and they are given in the "sig."

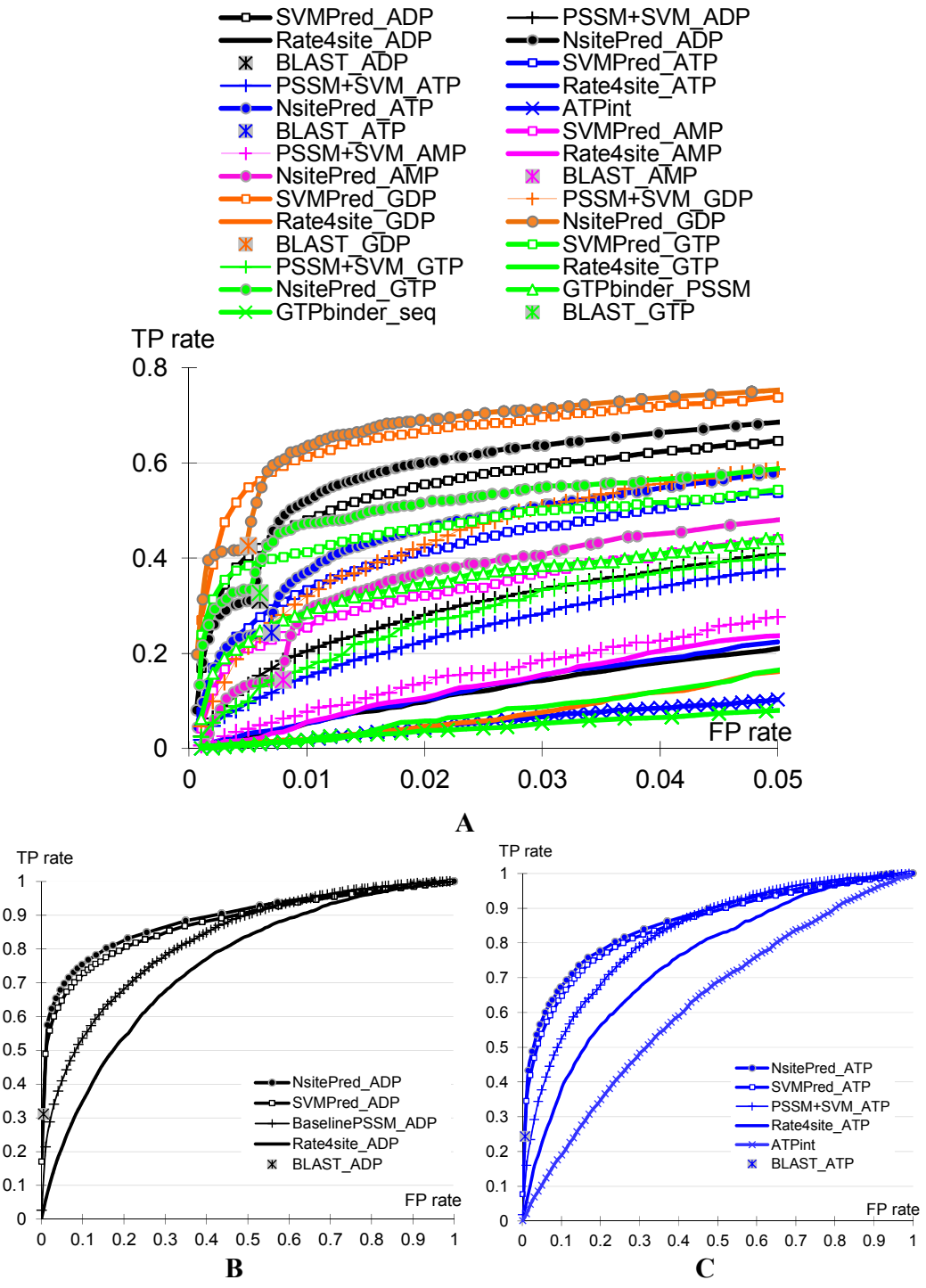
columns. The significance tests compare paired per-sequence prediction quality over a given benchmark dataset. The + and – mean that the NsitePred is statistically significantly better / worse with p <0.01, and = means that results are not significantly different. The "NA" means that the corresponding value could not be computed, i.e., BLAST generates only the binary predictions.

Type	Predictor	AUC		MCC		PREC	REC	SPEC	ACC
		value	sig.	value	sig.				
ATP	NsitePred	**0.861**		**0.46**		0.519	0.444	0.982	0.96
	SVMPred	0.854	+	0.433	+	0.564	0.361	0.988	**0.962**
	Rate4site	0.749	+	0.182	+	0.132	0.446	0.87	0.852
	PSSM+SVM	0.824	+	0.27	+	0.262	0.354	0.957	0.933
	BLAST	NA	NA	0.359	+	**0.578**	0.243	**0.993**	**0.962**
	ATPint	0.627	+	0.078	+	0.061	**0.539**	0.651	0.648
ADP	NsitePred	**0.893**		**0.572**		0.633	**0.544**	0.988	0.971
	SVMPred	0.885	+	0.555	+	**0.704**	0.458	0.993	**0.973**
	Rate4site	0.749	+	0.161	+	0.106	0.472	0.844	0.83
	PSSM+SVM	0.826	+	0.296	+	0.344	0.298	0.978	0.953
	BLAST	NA	NA	0.439	+	0.658	0.311	**0.994**	0.969
AMP	NsitePred	**0.829**		**0.377**		0.511	0.304	0.988	0.962
	SVMPred	0.82	+	0.36	+	**0.667**	0.208	**0.996**	**0.966**
	Rate4site	0.755	+	0.174	+	0.107	**0.562**	0.799	0.79
	PSSM+SVM	0.788	+	0.203	+	0.142	0.46	0.889	0.873
	BLAST	NA	NA	0.222	+	0.395	0.145	0.992	0.959
GDP	NsitePred	**0.91**		**0.675**		0.734	**0.646**	0.991	0.976
	SVMPred	0.905	+	0.655	+	**0.716**	0.623	0.989	**0.977**
	Rate4site	0.733	+	0.17	+	0.11	0.516	0.823	0.811
	PSSM+SVM	0.879	+	0.442	+	0.433	0.502	0.972	0.952
	BLAST	NA	NA	0.564	+	0.780	0.426	**0.995**	0.972
GTP	NsitePred	**0.844**		**0.562**		0.706	0.473	0.991	0.968
	SVMPred	0.836	+	0.551	+	**0.848**	0.373	**0.997**	**0.97**
	Rate4site	0.748	+	0.18	+	0.108	**0.569**	0.806	0.796
	PSSM+SVM	0.801	+	0.308	+	0.331	0.346	0.968	0.941
	BLAST	NA	NA	0.461	+	0.689	0.327	0.994	0.968
	GTPbinder_seq	0.548	+	0.03	+	0.055	0.177	0.876	0.849
	GTPbinder_PSSM	0.802	+	0.388	+	0.655	0.246	0.995	0.965

For Dataset 1, across predictions for the five nucleotide types, the NsitePred obtains AUC ≥ 0.83, MCC ≥ 0.38, and accuracy > 0.96. Our method outperforms the other approaches by a statistically significant margin for both AUC and MCC measures. Although some other approaches provide higher precision, recall or specificity, the NsitePred provides favorable balance between these three measures. Based on the MCC, which provides an overall estimate of the quality of the binary predictions, the NsitePred is superior to SVMPred and the BLAST-based predictor, followed by the PSSM profile-based predictor, and the Rate4site. We note that the consensus-based NsitePred achieves higher AUC and MCC

values and a better balance between the precision and recall than SVMPred. This suggests that the predictions from SVMPred are complementary to the alignment-based predictions. Since the Rate4site only considers the residue conservation, its relatively low predictive performance could be explained by the fact that the conserved residues could also include binding residues for other types of ligand such as the metal ions, carbohydrates, peptides, etc. This explanation is supported by the relatively high recall (i.e., high fraction of the correctly predicted native binding residues) coupled with the low specificity (which indicates an over-prediction of the binding residues) which are achieved by the Rate4site.

The AUC, MCC, precision, specificity and accuracy of ATPint are lower than the values achieved by NsitePred and the 3 baseline predictors, see Table 6-1, and they are also lower than it was originally reported (Chauhan *et al*., 2009). The likely reason for that is the fact that the authors of ATPint used a balanced number of binding and non-binding residues to design and evaluate their method, which resulted in the lower predictive quality when applied here to the full chains (where the number of binding residues is substantially lower than the number of non-binding residues). Our results indicate that the ATPint over-predicts the ATP-binding residues, which is evidenced by the low specificity and precision, i.e., a high number of false positives. We show that, as expected and as shown in (Chauhan *et al*., 2010), GTPbinder that utilizes the evolutionary profile (GTPbinder_PSSM) outperforms the version that does not use this information (GTPbinder_seq). The PSSM-based GTP_binder achieves AUC = 0.8 and MCC = 0.39 that are lower than the values achieved by the NsitePred (by the statistically significant margin) and the BLAST-based predictor, and higher than the values achieved by the PSSM profile-based predictor and Rate4site.


SVMPred_ADP
Rate4site_ADP
BLAST_ADP
PSSM+SVM_ATP
NsitePred_ATP
BLAST_ATP
PSSM+SVM_AMP
NsitePred_AMP
SVMPred_GDP
Rate4site_GDP
BLAST_GDP
PSSM+SVM_GTP
NsitePred_GTP
GTPbinder_seq
PSSM+SVM_ADP
NsitePred_ADP
SVMPred_ATP
Rate4site_ATP
ATPint
SVMPred_AMP
Rate4site_AMP
BLAST_AMP
PSSM+SVM_GDP
NsitePred_GDP
SVMPred_GTP
Rate4site_GTP
GTPbinder_PSSM
BLAST_GTP
TP rate
FP rate
0
0.2
0.4
0.6
0.8
0.01
0.02
0.03
0.04
0.05
A
TP rate
FP rate
NsitePred_ADP
SVMPred_ADP
BaselinePSSM_ADP
Rate4site_ADP
BLAST_ADP
B
TP rate
FP rate
NsitePred_ATP
SVMPred_ATP
PSSM+SVM_ATP
Rate4site_ATP
ATPint
BLAST_ATP
C

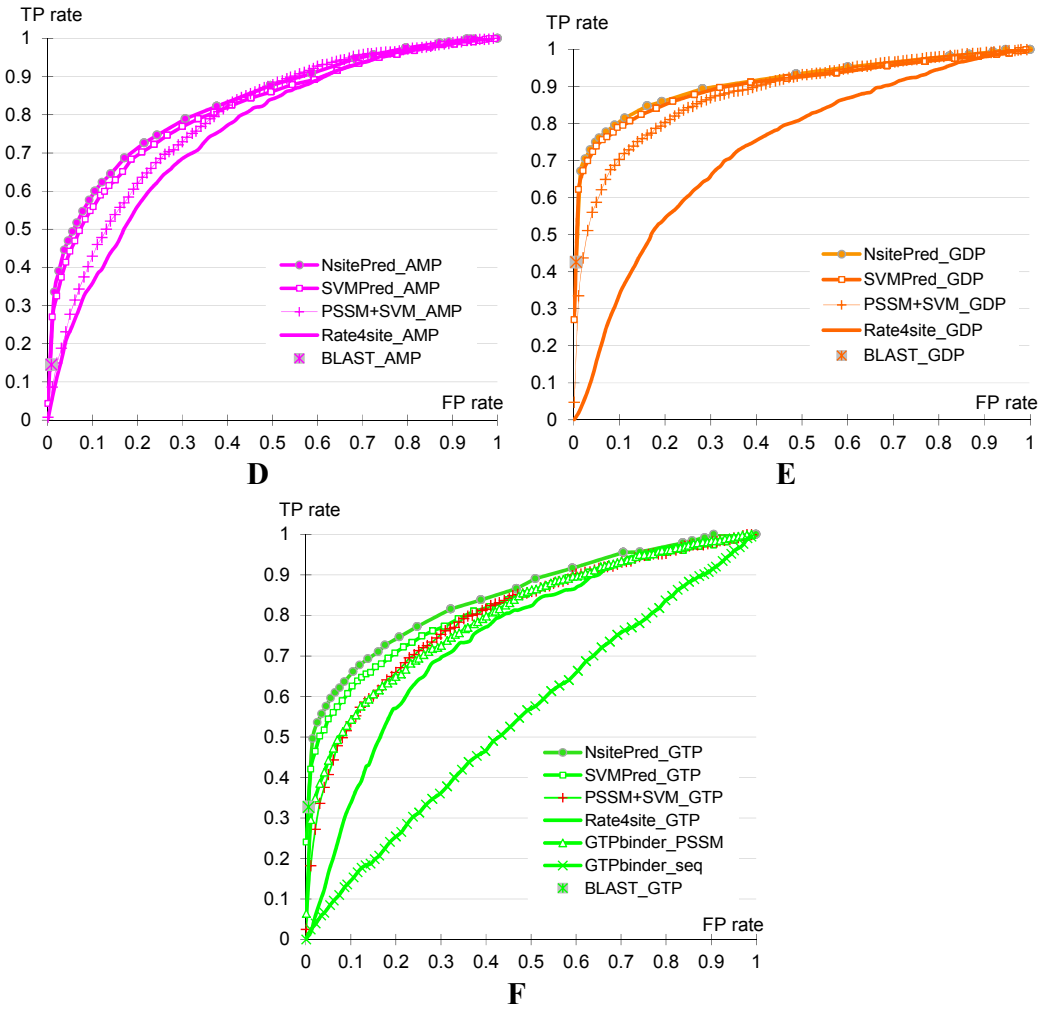


Figure 6-1: The ROC curves for the NsitePred (denoted using thick solid lines with filled circle markers), SVMPred (denoted using thick solid lines with hollow square markers), ATPint (thick solid line with x markers), GTPbinder (thick solid lines using cross and hollow triangle markers), Rate4site (thick solid line without markers), and the predictor based on the PSSM with the SVM classifier (thin solid line with cross makers) for predictions on Dataset 1. The BLAST-based solution is shown using a single point (star marker on grey background) that corresponds to the binary predictions. The results are based on 5-folds cross validation. A) The FP-rate is constrained to [0, 0.05] range for all 5 types of nucleotides; B) The full ROC curve for ADP; C) The full ROC curve for ATP; D) The full ROC curve for AMP; E) The full ROC curve for GDP; F) The full ROC curve for GTP.

The ROC curves based on the predictions on Dataset 1 are shown in Figure 6-1. Panel A focuses on the FP rates < 0.05 since only about 4% of residues bind to nucleotides; the full ROC curves are given in panels B, C, D, E and F. The

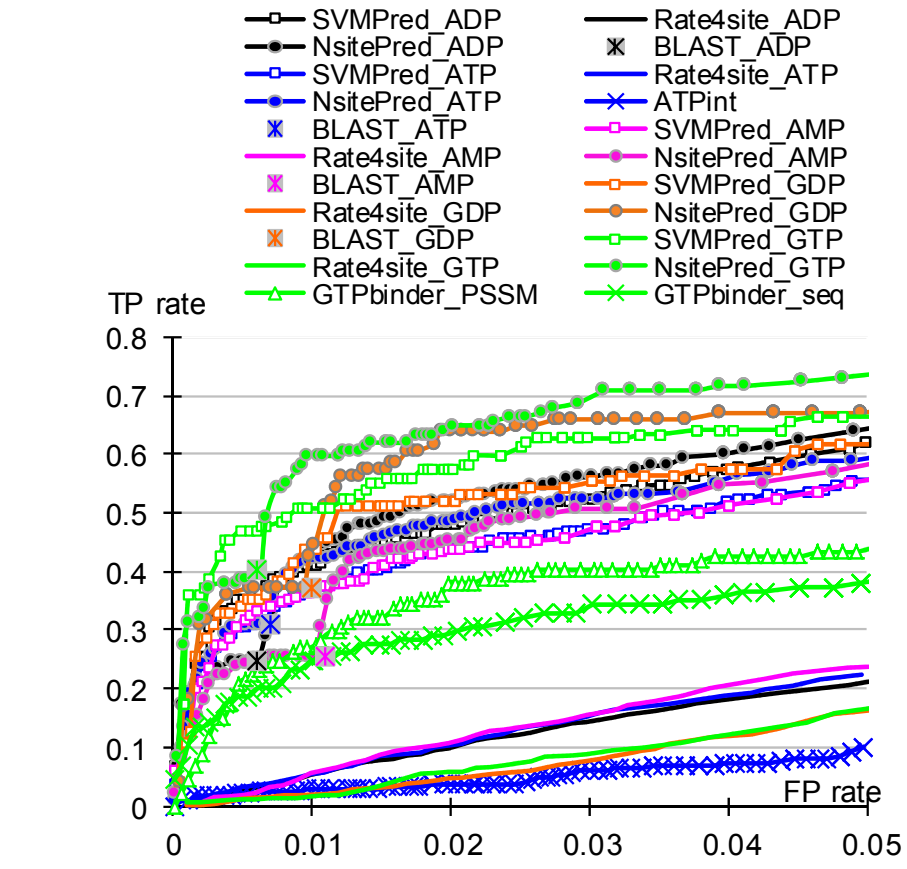
BLAST-based predictor does not provide the probabilities, and thus we include a single point that corresponds to its binary predictions. The ROC curves reveal that NsitePred provides higher TP rates for the FP values between 0.01 and 0.05 when compared with the other methods for each of the five types of the ligands.

Table 6-2 compares NsitePred with the existing methods and baseline predictors on Dataset 2, which consists of chains that were released after the NsitePred was designed and which are dissimilar to chains in the Dataset 1 that was used to build the predictive models. Similar to the results on Dataset 1, NsitePred achieves significantly higher AUC and MCC values when compared with the other methods for all 5 nucleotide types. NsitePred improves by 0.01 to 0.02 in AUC and 0.01 to 0.04 in MCC, depending on a nucleotide type, over the predictions from SVMPred by implementing the consensus of SVMPred and the BLAST-based alignment. These improvements are shown to be statistically significant. The ROC curves of NsitePred and the other methods on Dataset 2 are given in Figure 6-2 (panel A for the FP rates < 0.05 and panels B, C, D, E and F for entire range of FP rates). The ROC curves reveal that NsitePred provides higher TP rates for the FP values between 0.012 and 0.05 when compared with the other methods for each of the five types of the nucleotides. We also evaluated the predictive quality of the considered methods for prediction of all nucleotide-binding residues. In this case, a residue is defined as a "nucleotide-binding" if it interacts with any of the five nucleotides, and a residue is predicted as a "nucleotide-binding" when a given method predicts that this residue interacts with any of the five nucleotides; see the "All" rows in Table 6-2. Similarly as for the prediction of individual nucleotide types, NsitePred achieves higher AUC, MCC, and recall than the remaining methods, while the BLAST-based predictor achieves higher precision, specificity and accuracy, see Table 6-2.

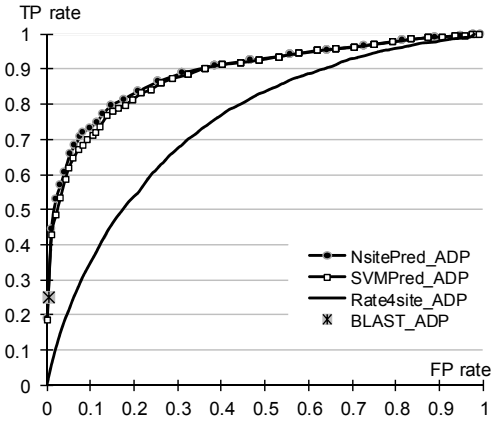
Table 6-2: Comparison of the quality of the sequence-based prediction of the ATP, ADP, AMP, GDT, GTP, and nucleotide-binding (indicated by all) residues between the NsitePred and the related predictors of nucleotide binding residues, including ATPint and GTPbinder that predict ATP-binding and GTP-binding residues, respectively, and predictors based

on the alignment (utilizing BLAST), conservation scoring (utilizing Rate4site), and SVMPred (utilizing the same feature representation as NsitePred and SVM classifier) on Dataset 2. In the evaluation of nucleotide-binding residue, a residue is defined as a "nucleotide-binding" if it interacts with any of the five nucleotides, and a residue is predicted as a "nucleotide-binding" when a given method predicts that this residue interacts with any of the five nucleotides. The highest values for each ligand type and each quality index, including AUC, precision (PREC), recall (REC), specificity (SPEC), accuracy (ACC), and MCC, are set in bold. The significance of the differences between NsitePred and the other methods are measured for the AUC and MCC and they are given in the "sig." columns. The significance tests compare paired per-sequence prediction quality over a given benchmark dataset. The + and – mean that the NsitePred is statistically significantly better / worse with p <0.01, and = means that results are not significantly different. The "NA" means that the corresponding value could not be computed, i.e., BLAST generates only the binary predictions.

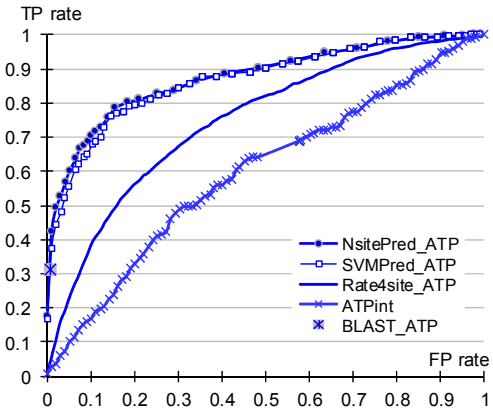
Type	Predictor	AUC		MCC		PREC	REC	SPEC	ACC
		value	sig.	value	sig.				
ATP	NsitePred	**0.875**		**0.476**		0.528	0.46	0.985	0.967
	SVMPred	0.868	+	0.451	+	0.587	0.367	0.991	0.969
	Rate4site	0.741	+	0.167	+	0.107	**0.464**	0.862	0.849
	BLAST	NA	+	0.422	+	**0.611**	0.31	**0.993**	**0.97**
	ATPint	0.606	+	0.066	+	0.051	0.512	0.66	0.655
ADP	NsitePred	**0.893**		**0.512**		0.589	0.474	0.987	0.968
	SVMPred	0.886	+	0.5	+	0.68	0.388	0.993	**0.971**
	Rate4site	0.735	+	0.166	+	0.102	**0.521**	0.823	0.812
	BLAST	NA	+	0.376	+	**0.608**	0.249	**0.994**	0.966
AMP	NsitePred	**0.876**		**0.501**		0.606	0.423	0.987	**0.969**
	SVMPred	0.87	+	0.478	+	**0.721**	0.335	**0.994**	0.967
	Rate4site	0.752	+	0.175	+	0.114	**0.52**	0.824	0.811
	BLAST	NA	+	0.339	+	0.504	0.255	0.989	0.959
GDP	NsitePred	**0.867**		**0.576**		0.598	**0.585**	0.985	0.97
	SVMPred	0.855	+	0.553	+	**0.632**	0.511	0.988	**0.971**
	Rate4site	0.748	+	0.173	+	0.116	0.545	0.793	0.781
	BLAST	NA	+	0.454	+	0.593	0.372	**0.99**	0.967
GTP	NsitePred	**0.909**		**0.64**		0.711	**0.604**	0.988	**0.969**
	SVMPred	0.887	+	0.602	+	0.783	0.485	0.993	**0.969**
	Rate4site	0.745	+	0.168	+	0.103	0.531	0.817	0.806
	BLAST	NA	+	0.539	+	**0.761**	0.403	**0.994**	0.966
	GTPbinder_seq	0.742	+	0.276	+	0.544	0.276	0.988	0.954
	GTPbinder_PSSM	0.822	+	0.418	+	0.597	0.321	0.989	0.957
All	NsitePred	**0.905**		**0.48**		0.386	**0.663**	0.957	0.946
	SVMPred	0.899	+	0.455	+	0.374	0.62	0.958	0.945
	Rate4site	0.741	+	0.17	+	0.107	0.512	0.83	0.818
	BLAST	NA	+	0.423	+	**0.426**	0.468	**0.974**	**0.955**


SVMPred_ADP
NsitePred_ADP
SVMPred_ATP
NsitePred_ATP
BLAST_ATP
Rate4site_AMP
BLAST_AMP
Rate4site_GDP
BLAST_GDP
Rate4site_GTP
GTPbinder_PSSM
Rate4site_ADP
BLAST_ADP
Rate4site_ATP
ATPint
SVMPred_AMP
NsitePred_AMP
SVMPred_GDP
NsitePred_GDP
SVMPred_GTP
NsitePred_GTP
GTPbinder_seq
TP rate
FP rate
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.01
0.02
0.03
0.04
0.05

A


TP rate
FP rate
NsitePred_ADP
SVMPred_ADP
Rate4site_ADP
BLAST_ADP
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

B


TP rate
FP rate
NsitePred_ATP
SVMPred_ATP
Rate4site_ATP
ATPint
BLAST_ATP
0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
0.8
0.9
1

C

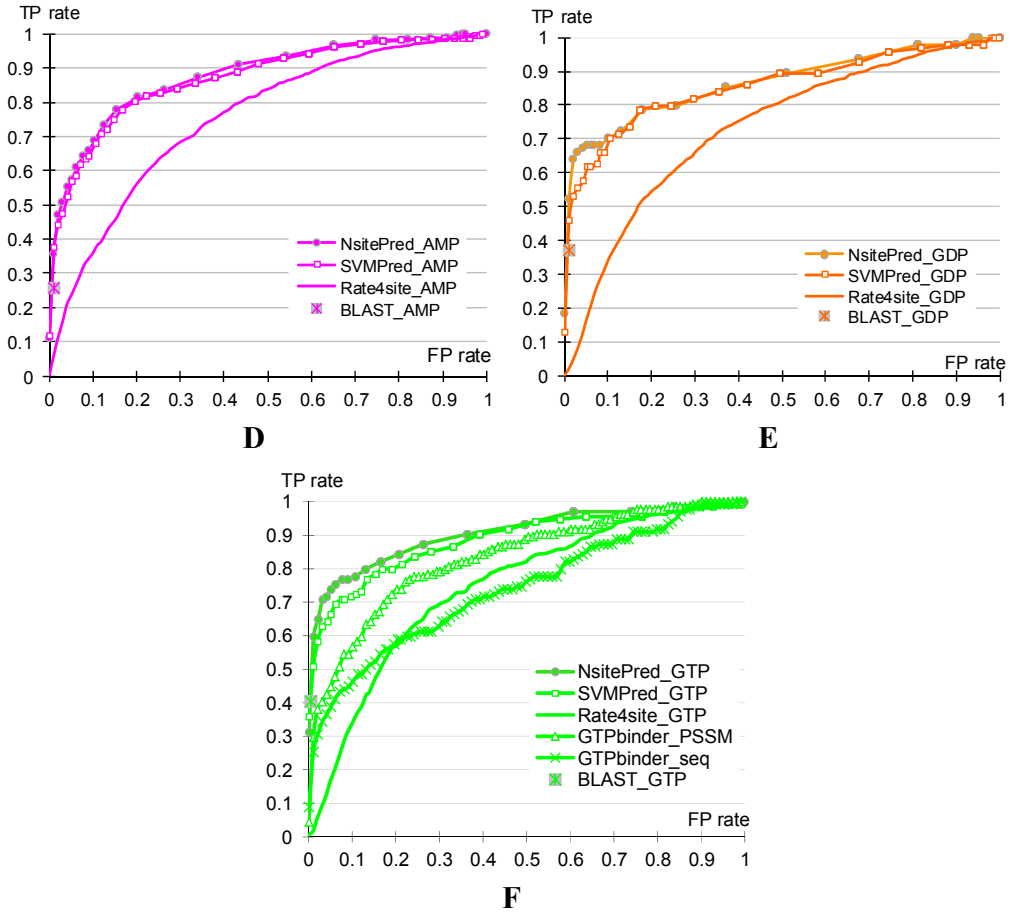


Figure 6-2: The ROC curves for the NsitePred (denoted using thick solid lines with filled circle markers), SVMPred (denoted using thick solid lines with hollow square markers), ATPint (thick solid line with x markers), GTPbinder (thick solid lines using cross and hollow triangle markers) and Rate4site (thick solid line without markers) for predictions on Dataset 2. The BLAST-based solution is shown using a single point (star marker on grey background) that corresponds to the binary predictions. Dataset 2 consists of chains that were released after the NsitePred was designed and which are dissimilar to chains in the Dataset 1 that was used to build the predictive models. A) The FP-rate is constrained to [0, 0.05] range for all 5 types of nucleotides; B) The full ROC curve for ADP; C) The full ROC curve for ATP; D) The full ROC curve for AMP; E) The full ROC curve for GDP; F) The full ROC curve for GTP.

6.7.2 Performance on non-binding chains

We also assess the predictive quality of NsitePred and the other methods on protein sequences that do not interact with nucleotides (Dataset 3). We measure

the error rate, which is defined as ratio between the number of false positives (FPs) and the total number of residues, for all considered methods; we note that there are no positive (nucleotide-binding) residues in this dataset. The error rates of NsitePred are 0.48%, 1.15%, 0.76%, 0.93%, and 0.67% for ATP, ADP, AMP, GTP, and GDP respectively, see Table 6-3. The error rates of NsitePred are slightly higher than the error rates of BLAST-based method and SVMPred, but lower than the error rates of ATPint and GTPbinder. We note that NsitePred predicts 3.6% and 3.1%, 3.2% and 3.1%, 2.4% and 2.2%, 2.6% and 4.0%, and 3.3% and 4.3% of the residues in Dataset 1 and Dataset 2 as ATP-, ADP-, AMP-, GTP- and GDP-binding residues, respectively. These results demonstrate that NsitePred predicts fewer nucleotide-binding residues for the non-binding chains than for the nucleotide-binding chains.

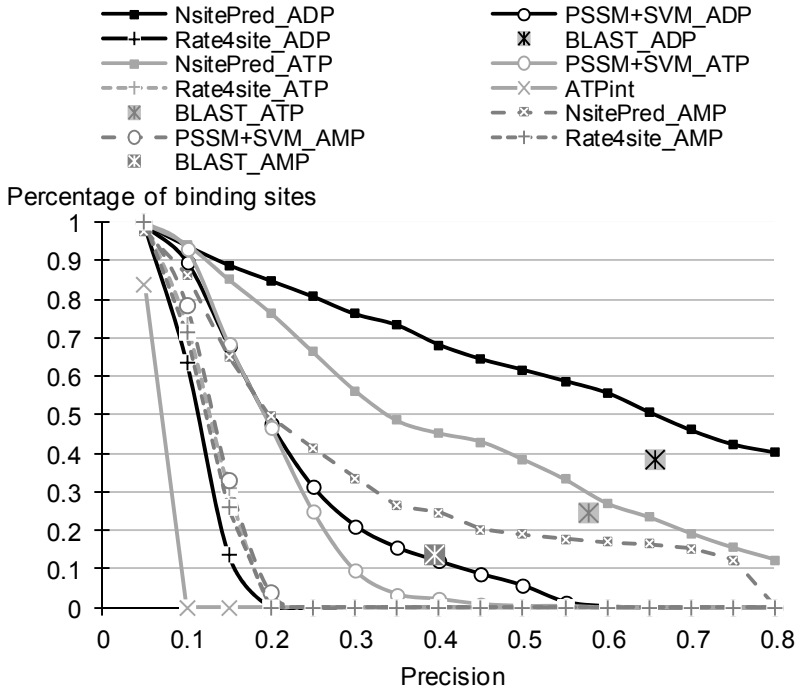
Table 6-3: The error rates of NsitePred and the other considered predictors on protein chains from Dataset 3 that do not interact with nucleotides. The error rate is defined as the ratio between the number of false positives and the total number of instances.

Predictor	Error rate				
	ATP	ADP	AMP	GTP	GDP
NsitePred	0.48%	1.15%	0.76%	0.93%	0.67%
SVMPred	0.36%	0.86%	0.58%	0.75%	0.53%
BLAST	0.28%	0.59%	0.43%	0.51%	0.34%
ATPint	20.1%	NA	NA	NA	NA
GTPbinder_seq	NA	NA	NA	3.47%	NA
GTPbinder_PSSM	NA	NA	NA	2.94%	NA

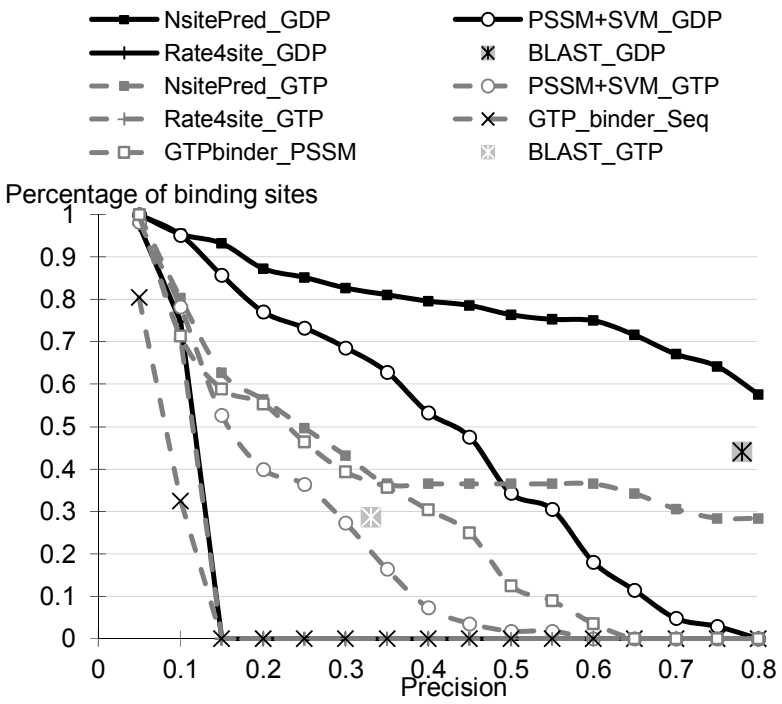
6.7.3 Evaluation per binding site

Besides the evaluation at the residue level, we investigate the quality of the predictions at the binding site level. A given binding site, which is made of residues that interact with the same molecule, is assumed to be correctly predicted if at least 50% of its residues are correctly predicted. We vary the per-residue precision between 0.05 and 0.8 (the number of correctly predicted binding sites is approximately 0 when the precision > 0.8) with 0.05 step to control the number of false positives. This is performed by thresholding the predicted probabilities (we vary the threshold to obtain the binary predictions) for all methods except for the

BLAST-based predictor, which is represented with one point that corresponds to its binary prediction. Figure 6-3 shows that NsitePred correctly predicts around 62% of the ADP-binding sites, 38% of the ATP-binding sites, 19% of the AMP-binding sites, 76% of the GDP-binding sites, and 37% of the GTP-binding sites at the precision equal 0.5, i.e., when half of the predicted binding residues are correct. To compare, the PSSM-profile based predictor correctly predicts only 6%, 0%, 0%, 34% and 2% of the binding sites for the ADP, ATP, AMP, GDP and GTP, respectively, when considering the same precision. The Rate4site predictor cannot achieve such high precision for any of the five types of nucleotides, and thus we assume that its success rate equals 0. The ATPint and GTPbinder_seq also cannot correctly predict any sites at precision of 0.5, while the GTPbinder_PSSM correctly predicts about 13% of the GTP-binding sites. When compared with the BLAST-based predictor at the same precision, the NsitePred correctly finds 5%-15% more binding sites. Overall, the results indicate that the NsitePred captures more binding sites than the other predictors, especially at the higher precision rates.



A



B

Figure 6-3: Evaluation of the predictions per binding site on Dataset 1 (based on 5-folds cross validation) for the NsitePred (denoted using square markers), ATPint (solid line with x markers), GTPbinder (dashed lines using x and square markers), Rate4site (cross markers), and the predictor based on the PSSM with the SVM classifier (hollow circle makers); the BLAST-based solution is shown using a single point (star marker on grey background) that corresponds to the binary predictions. A given binding site is assumed to be correctly predicted if at least 50% of its residues are correctly predicted. The y-axis shows the percentage of the correctly predicted binding sites. We vary the per-residue precision (x-axis) between 0.05 and 0.8 with 0.05 step to control the number of false positives.

6.7.4 Impact of the degree of spread of the binding residues in the protein chain

Some nucleotide binding sites consist of a single segment in the protein chain, e.g., the p-loop motif GXXXXGKS(T)T, while other sites are composed of binding residues that are sparsely distributed over the sequence. We hypothesize that the difficulty of the sequence-based prediction of the nucleotide binding sites depends on the degree to which the corresponding binding residue are spread over the sequence. The sites that are composed of a single segment of consecutive binding

residues should be easier to predict than the sites for which the binding residues are sparsely distributed over a large, relative to the total number of binding residues, stretch of the sequence. We study the relation between this degree of the spread of the binding residues in the chain and the predictive quality. We quantify this spread/clustering of the binding residues using a spread index that reflects the average number of non-binding residues between the consecutive binding residues in the chain, and which equals zero when a given site consists of a single segment of the consecutive binding residues. The spread index is defined as follows

$$spread = \frac{\sum_{i=1}^{n-1} gap(R_i, R_{i+1})}{n-1}$$

where R_i and R_{i+1} are the i^{th} and $(i+1)^{th}$ binding residue, respectively, in a given binding site that consists of n residues. Given that the binding residues are sorted by their residue number in the protein sequence, the $gap(R_i, R_{i+1})$ is defined as

$$gap(R_i, R_{i+1}) = \begin{cases} N_{i+1} - N_i - 1 & (\text{if } N_{i+1} - N_i < 13) \\ 12 & (\text{otherwise}) \end{cases}$$

where N_i and N_{i+1} are the residue numbers in the protein chain of the R_i and R_{i+1} binding residues, respectively The "$N_{i+1} - N_i - 1$" quantifies the number of the non-binding residues between R_i and R_{i+1}. Moreover, we assume that a given pair of consecutive binding residues that are separated by 12 or more non-binding residues is not likely to form local interactions in the tertiary structure and thus the corresponding $gap(R_i, R_{i+1})$ value is rounded down to 12. This cut-off threshold is based on the definition of the long-range interactions, which are defined as contacts formed between residues that are at least 12 positions away in sequence (Tegge et al., 2009).

We sort all binding sites for a given nucleotide type in the ascending order according to their spread index values, and we divide them into 5 equally sized subsets where the first subset contains 20% of sites with the lowest spread. The

average spread values for each subset and the corresponding predictive quality for NsitePred calculated based on the 5-folds cross validation on Dataset 1 are shown in Figure 6-4. Panel A shows the average precision (fraction of correct prediction among the predicted binding residues) at the recall equal 0.5, while panel B gives the average recall (fraction of correctly predicted native binding residues) at the precision equal 0.5. The results show that both precision and recall decline with the increase of spread value, and that this trend is independent of the nucleotide type. The NsitePred performs very well for compact sites, i.e., sites that include residues that are clustered close in the sequence, and its quality declines when the binding residues are spread over a longer fragment of the protein chain. Moreover, this relation also explains the differences in the predictive quality for different nucleotide types. The average spread values for the binding sites of GDP, ADP, ATP, GTP, and AMP are 2.53, 2.93, 3.25, 3.35, and 3.53, respectively. The Pearson correlation coefficients between these spread values and the corresponding AUC and MCC values achieved by NsitePred, see Table 6-1, equal -0.96 and -0.98, respectively.

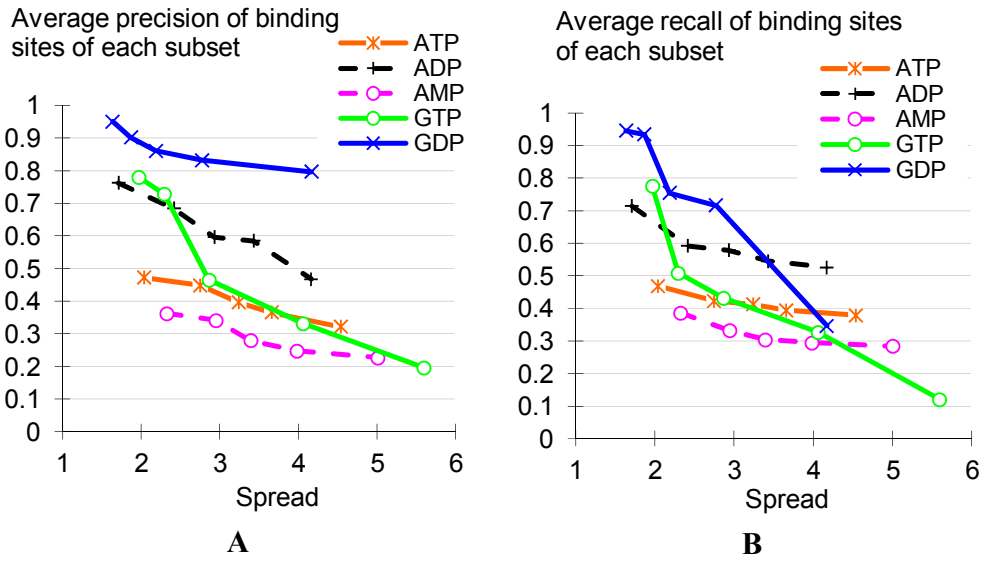


Figure 6-4: Relation between the predictive quality (*y*-axis) and the spread index values (*x*-axis). The binding sites for a given nucleotide type, which are sorted in the ascending order based on their spread index values, are divided into 5 equally sized subsets where the first subset (the left-most point) contains 20% of sites with the lowest spread, and the fifth subset (the right-

most point) with the 20% of sites with the highest values. Panel A shows the average precision (over the sites in a given subset) at the recall = 0.5. Panel B shows the average recall at the precision = 0.5.

6.7.5 Sequence-derived hallmarks of nucleotide-binding residues

The significant improvements in the quality of the prediction of the binding residues for the five considered nucleotides between the NsitePred and the PSSM profile-based predictions (denoted as PSSM+SVM), see Table 6-1, suggest that the increased quality stems from the use of the novel input features proposed in this work. This means that the nucleotide-binding residues could be characterized using the information concerning their conservation and the predicted secondary structure, relative solvent accessibility, and dihedral angles. We analyze the features used by the NsitePred to find the corresponding sequence-derived markers of the nucleotide-binding residues.

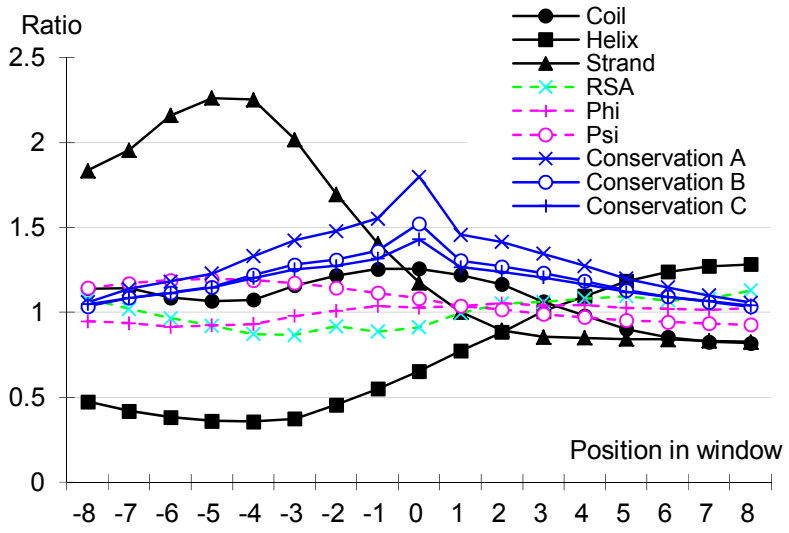


Figure 6-5: The ratios, which are calculated as the average of values of a given feature for the nucleotide-binding residues divided by the average for the non-binding residues, at the 17 positions in the sliding window used by NsitePred. The ratios are calculated for the predicted secondary structures (helix, strand, and coil), RSA, dihedral angles (*phi* and *psi*), and the three conservation scores based on the Shannon entropy (conservation A) and formulas proposed in (Wang and Samudrala, 2006) (conservation B), and in (Capra and Singh, 2007) (conservation C). The *x*-axis shows the positions in the sequence relative to the predicted residues, which is at 0.

We focus on the features that are selected for at least three nucleotide types. We observe that the selected features that are based on the predicted secondary structure, RSA, and *psi* angles are biased to positions in the sequence that are towards the N-terminus from the predicted residue. We investigate this asymmetry (the lack of use of the positions towards the C-terminus at the same position (relative to the predicted residues) computed for all native nucleotide binding residues and the non-binding residues, respectively, i.e., ratio = average value of a given feature for the nucleotide binding residues divided by the average for the non-binding residues. The value close to 1 indicates that similar average values are observed for the binding and the non-binding residues, and thus the corresponding feature at this position does not differentiate between these two types of residues. The ratios along the 17 positions of the sliding window used by the NsitePred for the probabilities of secondary structures, RSA values, dihedral angles, and the three conservation scores are shown in Figure 6-5. As our feature selection suggests, the plots for the secondary structure, dihedral angles and RSA are asymmetric, which is in contrast to the conservation scores that are symmetric. We note the particularly high ratios for the predicted probabilities of strands at the positions that are 4 to 5 residues towards the N-terminus. These ratios show that the nucleotide-binding residues are characterized by over twice higher probabilities of the predicted strand residues for these positions when compared with the non-binding residues. Moreover, positions towards the C-terminus show ratios relatively close to 1, i.e. between 0.826 and 1.001. We also observe that the helix is less likely to occur towards the N-terminus when compared with positions towards the C-terminus, which coincides with the above preference towards the strands. Similarly (as expected), the ratios for the *phi* and *psi* angles follow the pattern of the secondary structures, although they vary in a smaller range, e.g. ratios for the *psi* angles vary between 0.926 and 1.196, with the larger values only towards the N-terminus. The RSA values are smaller towards the N-terminus and close to 1 towards the C-terminus, which explains the bias towards the former positions among the selected features. The plot indicates that residues located near by and towards the N-terminus from the nucleotide-binding residues are less

likely to be solvent exposed. The ratios for the 3 conservation scores are symmetrically distributed around the central residue. Their largest values are at the central position, which indicates that the nucleotide binding residues are more conserved than the non-binding residues. These 3 plots also reveal that the residues at the adjacent positions have smaller ratios, which means that the nucleotide-binding residues are flanked by residues with a smaller degree of conservation.

We also note that several features extracted from the PSSM profile (including the aggregations using certain amino acid groups) and certain collocated amino acid pairs are included among the selected features, and thus they can be used to formulate sequence-derived hallmarks of the nucleotide binding residues. The PSSM profile-based features are likely correlated with the formation of certain secondary structure types, e.g., the scores aggregated for the hydrophobic residues is associated with the formation of strands and coils. Three amino acid pairs, GXXXS, GXG and GXS, where 'X' indicate a wild card residue (any amino acid type) and where the right-most residue is located at the center of a sliding window, are found to be strong markers for the nucleotide binding residues. These collocated pairs are related to the p-loop motif, GXXXXGKS(T), which is characteristic to the interactions with ATP (Saraste et al., 1990).

6.8 Conclusions

In this chapter, we described the NsitePred method, which is a collection of five accurate sequence-based predictors that identify binding residues for the five most populated nucleotides in the PDB, including ATP, ADP, AMP, GTP, and GDP. Empirical results demonstrate that NsitePred significantly outperforms the existing ATPint and GTPbinder methods, as well as solutions based on the sequence alignment and residue conservation scoring. The favorable predictive quality stems from the usage of novel custom-designed input features that are based on the sequence, the sequence-derived evolutionary profiles, the sequence-predicted structural descriptors, and the BLAST-based alignment. Our study

shows that NsitePred performs particularly well for the binding sites in which the binding residues are clustered close together in the sequence. Analysis of the features used in the predictive model reveals several interesting hallmarks of the nucleotide-binding residues, which are related to the arrangement of secondary structures, dihedral angles, and certain amino acid pairs in the specific neighboring positions in the sequence.

The NsitePred is implemented as a publicly-available web server at http://biomine.ece.ualberta.ca/nSITEpred/.

CHAPTER 7 Prediction of nucleotide-binding sites from protein structure

7.1 Introduction

In chapter 6, we proposed a method that predicts the nucleotide-binding residues from protein sequence. In this chapter, we propose a novel predictor, called NSiteMatch, which utilizes tertiary structure as its input. The structure-based prediction is expected to provide better performance than the sequence-based methods since it uses more information (structure instead of sequence). Although the application of the structure-based method is limited to protein for which the structure is already solved, we note that many of these proteins lack full annotation of their biological functions. Moreover, availability of the structure means that these proteins can be utilized to perform molecular docking-based rational drug discovery. The NSiteMatch method is similar to the methods that were surveyed in Chapter 5. The main difference is that NSiteMatch only predicts binding sites of nucleotides while the methods discussed in Chapter 5 identify binding sites of all types of small organic compounds. We focused on the binding sites of nucleotides because nucleotides play pivotal roles in a number of biological processes and they constitute about 15% of biologically relevant ligands included in PDB (Dessailly et al., 2008; Shin and Cho, 2005). Although the NSiteMatch method is designed for prediction of binding sites of nucleotides, it also serves as a platform that could be extended to predict interactions with other small ligands.

7.2 Related work

Substantial efforts were observed in the identification and characterization of the nucleotide binding sites in the past two decades. Most of these approaches analyzed the known nucleotide-binding protein sequences and structures to identify conserved motifs. For instance, the Walker A sequence motif was found

in a variety of nucleotide-binding proteins that include the alpha and beta subunits of ATP synthase, myosin, transducin, helicases, kinases and RecA (Walker et al., 1982). Moodie and colleagues proposed a fuzzy recognition template for the characterization of the adenylate-protein interactions (Moodie et al., 1996). However, the abovementioned studies characterize the sequence and structural motifs for a relatively narrow range of the nucleotide-protein interactions and none of these studies concerns prediction of nucleotide-binding sites. Although there are no methods that specifically focus on the prediction of the nucleotide-binding sites, over a dozen methods were proposed for the structure-based prediction of binding sites for small organic compounds. These methods were surveyed in Chapter 5. They include SURFNET (Laskowski, 1995), PocketFinder (Hendlich et al., 1997), PASS (Brady and Stouten, 2000), LIGSITE^{csc} (Huang and Schroeder, 2006), PocketPicker (Weisel et al., 2007), ConCavity (Capra et al., 2009), Fpocket (Le et al., 2009), Q-SiteFinder (Laurie and Jackson, 2005), Findsite (Skolnick and Brylinski, 2008) and MetaPocket (Huang, 2009). In Chapter 5, we demonstrated that Findsite outperforms all other methods while MetaPocket and Q-SiteFinder are among the second-best existing methods. Therefore, the proposed NSiteMatch method is compared with Findsite, MetaPocket, and Q-SiteFinder to investigate its predictive performance.

7.3 Problem definition

The input of NSiteMatch method is a protein structure with coordinates of all non-hydrogen atoms while the outputs are the coordinates which represent the centers of the predicted binding sites. For instance, if NSiteMatch identifies three binding sites for a given protein structure, the output would be the centers of the three binding pockets: (x_1, y_1, z_1), (x_2, y_2, z_2), and (x_3, y_3, z_3).

7.4 Dataset preparation

The benchmark dataset is designed to cover a wide range of nucleotide-binding proteins. Similarly as in Chapter 6, the nucleotides that are considered in this

study contain at least one of the five nucleobases, a 5-carbon sugar, and 1 to 3 phosphates. We extracted all complexes from PDB that include these nucleotides. Next, the maximal pairwise sequence identity of the resulting protein chains for each of the nucleotides was reduced to 40% with CD-hit (Li and Godzik, 2006). We include the nucleotides that bind to at least 50 chains where these chains belong to at least 20 different superfamilies based on the SCOP classification. The availability of at least 50 chains provides us with a sufficient number of annotated binding sites to build and evaluate a well-performing predictor. While the availability of at least 20 superfamilies assures that these nucleotides bind to a wide range of proteins that are diverse in their structure and sequence; the latter based on the 40% sequence similarity filtration. This also allows us to investigate the prediction of the distant functional relationships, i.e., binding of the same nucleotides to structurally different proteins. We extracted a total of 227, 321, and 140 chains that bind to ATP, ADP, and AMP, respectively. The lists of the protein chains that interact with the three nucleotides are given in Appendix A. The other nucleotide types were excluded due to the small sample size. Since the NSiteMatch and Findsite predictors utilize a template library which could contain structures that are (too) similar to the predicted protein, we created a reduced version of the dataset that contains protein chains annotated using the SCOP labels for each of the three nucleotides. In other words, we excluded the chains that were not included in the SCOP database. Using these annotations we could control the similarity/homology levels between the template library and the predicted protein. This allows us to assess the predictive quality of the NSiteMatch and Findsite when using templates that are dissimilar, at a given homology level, to the predicted protein. As a result, we extracted total of 114, 158, and 66 chains that are annotated with SCOP labels for ATP, ADP, and AMP, respectively.

7.5 Proposed solution

7.5.1 Preparation of the template library of NSiteMatch

The template library of NSiteMatch consists of the structures of the nucleotide-binding sites. For a given protein-nucleotide complex, a non-hydrogen atom of the protein is considered as an interacting atom if it is within 3.9Å to a non-hydrogen atom of the nucleotide (Chen and Kurgan, 2009; Luscombe *et al*., 2001). A binding site is defined as a collection of the interacting atoms that bind to the same nucleotide molecule. The 3D-coordinates, the atom types, and the residue types of the interacting atoms of each binding site were stored in the template library.

7.5.2 The NSiteMatch algorithm

The novelty of our approach is two-fold. First, based on the complementarity of the existing methods and the resulting improvements offered by the consensus-based approach, which was demonstrated in Chapter 5, the NSiteMatch combines the geometrical, energy-based, and threading approaches. Second, drawing from the observation that use of templates with the sufficient structure similarity leads to high quality predictions, which were also shown in Chapter 5, we use a template database to perform the predictions. However, unlike the only exiting threading-based Findsite that relies on the overall similarity of the entire protein fold, we use local similarity of the structure in the binding region to find the most suitable templates. This allows us to identify a larger number of potentially useful templates and to predict distant functional relationships, i.e., NSiteMatch utilizes the templates that share similarity in the binding region but which may share low homology with the predicted protein.

The NSiteMatch method includes two major phases. The first phase (steps 1 to 8) performs fitting of templates into the structure of the predicted protein based on a common substructure defined by the interacting atoms (steps 1 to 8); this is repeated for each template and each potential position of the center of the ligand.

The second phase (steps 9 and 10) processes the predictions, which are filtered using a docking energy function (see Section 2.2.4 for details), and next they are clustered and ranked. The method outputs the ranked list of the predicted centers of the ligands and the corresponding ranked list of the binding residues for each center. The above overview demonstrates that our method combines the geometrical approach, which is implemented in the procedures to define and score the templates, energy-based approach, by utilizing energy function to filter initial predictions, and threading, by using the template database.

The overall flow of the NSiteMatch algorithm is given in Figure 7-1. Given the predicted protein structure and a template library with the nucleotide-binding sites, the NSiteMatch algorithm is implemented with the following 10 steps:

Step 1. *Set the 3-dimensional grid space for the predicted protein*. We use grid with a step size of 2Å and a given grid point is retained if it is within 10Å to a non-hydrogen atom of the protein. A grid point is marked as protein and removed from the grid space if it is within 1.6 Å to a non-hydrogen atom of the protein; otherwise, the grid point is kept and annotated as solvent.

Step 2. *Select a binding site from the template library*. The binding site contains both the coordinates of the interacting atoms of the protein and the coordinates of the nucleotide atoms. We calculate the geometrical center of the nucleotide and the distances between the center and all interacting atoms of the protein. We use these distances to set values of two parameters. Among these distances, the maximal distance R represents the radius to cover all interacting atoms while the minimal distance r represents the distance between the center and the protein surface. The two parameters R and r are used in the subsequent steps.

Step 3. *Scan the grid space to assess which grid points fit the geometrical center of the binding site.* In step 3A, we first choose a grid point from the grid

space. Next, we assess whether the chosen grid point fits the geometircal center of the binding site, which is perfromed in step 3B. In step 3B, we first calculate the distances between this grid point and all atoms of the protein. Among these distances, the minimal distance is denoted as r_1. Our first premise is that if a grid point fits the geometrical center of a nucleotide, the distance between the grid point and the protein surface should be similar to the distance r between the center of the nucleotide and the protein surface. Therefore, a given grid point is retained only if $| r - r_1| \leq 2$Å. The 2Å margin is used to accommodate the step size of the grid space. Our second premise is that if a grid point fits the center of a nucleotide, the spatial arrangements of interacting atoms (atoms that participate in the protein-ligand interaction) around this point should be similar to the arrangements of atoms around the center of the nucleotide. We use triangles, of which two vertexes are the interacting atoms of the protein and the third vertex is the grid point (the third vertex is the center of the nucleotide in the template) to represent this spatial arrangements. For a given grid point, the grid-associated surface patch is defined as a collection of protein atoms that are within $R_1=R+2$Å to the grid point. By this definition, the radius R_1 of the grid-associated surface patch is slightly larger but still similar to the radius R of the binding site. We compare the triangles formed by the atoms of the template binding site and the triangles formed by the atoms of the grid-associated surface patches. The triangles are formed by two atoms of the grid-associated surface patch (or template binding site) and the grid point (or the center of the nucleotide). Among the vertexes, the grid point or the center of the nucleotide is invariant while the other two vertexes are chosen from a large number of combinations of the corresponding atoms. Therefore, we generate two sets of triangles for the binding site and the grid-associated surface patch. We say that a given triangle of the grid-associated surface patch matches a given triangle of the template binding site if the corresponding vertexes have the same atom type, residue type, and the

difference between the side length of the corresponding edges is less than or equal 2Å. A grid point and the associated surface patch are retained if at least 25% of the triangles in the template site match with the triangles of the surface patch, and the surface patch matches at least 50 triangles of the template binding site. In step 3C, we go back to step 3A and choose another grid point until all grid points are used. Finally, in step 3D, the retained grid points are passed to step 4.

Step 4. *Cluster the retained grid points*. Two retained grid points are assigned to the same cluster if they are neighboring grid points, i.e., the distance between the two points is 2Å. The clusters are sorted by the number of grid points and the top three clusters are selected; if the total number of clusters is smaller than 3 then all clusters are selected. We use two points to represent each cluster: the geometrical center of the grid points and the point with the maximal number of triangles that match the template binding site. We refer to these representative grid points as seeds and the associated surface patches as seed-associated surface patches.

Step 5. *Search for the maximal common sub-structure between the binding site and the seed-associated surface patches*. The seed-associated surface patch is defined as the collection of protein atoms that are within R_1=R+2Å to the seed (grid point), see step 5A. We search for the maximal common sub-structure between the template binding site and the seed-associated surface patches. We denote the atoms at the template binding site as a_1, a_2, …, a_n and the atoms at the seed-associated surface patch as b_1, b_2, …, b_m. An atom from the template site matches an atom from the surface patch if the two atoms have the same atom type and residue type. For every pair of the matched atoms, we create a corresponding vertex $g(a_i, b_j)$ on a new graph G, where a_i is the atom from the binding site and b_j is the atom from the selected surface patch and a_i matches b_j. Two vertices $g_k(a_i, b_j)$ and $g_l(a_s, b_t)$ in graph G are connected if two conditions are satisfied. First, $|D(a_i, a_s)-D(b_j, b_t)| \leq 2Å$,

where D(a_i, a_s) is the distance between a_i and a_s and D(b_j, b_t) is distance between b_j and b_t. Second, $a_i \neq a_s$ and $b_j \neq b_t$. Searching for the maximal common sub-structure between the template binding site and a given surface patch is equivalent to searching for the complete sub-graph in G. We used the backtracking algorithm to search for the complete sub-graph. The identified common sub-structure (atoms connected with green solid lines) between the template binding site and the seed-associated surface is shown in step 5B.

Step 6. *Superimpose the template binding site into the seed-associated surface patches.* In step 5 we identified a common sub-structure between the template binding site and a given seed-associated surface patch. By using the coordinates of the two sub-structures, we calculated the RMSD value between the two sub-structures and the translation vector (V) and the rotation matrix (M) to achieve this RMSD value. Based on V and M, we superimpose the nucleotide structure at the template binding site into the corresponding surface patch.

Step 7. *Select the next available seed and repeat steps 5 and 6 until all seeds are used.*

Step 8. *Select the next available binding site in the template library and repeat steps 2 to step 6 until all templates in the library are used*, see step 8A. Once all templates are used, step 8B passes the predictions (the locations of the binding nucleotides) to step 9.

Step 9. *Filter the predictions by using a docking energy function*. Our algorithm superimposes a number of nucleotide structures on the surface of the predicted protein. That is, both the coordinates of the protein and the coordinates of the superimposed nucleotides are known. Therefore, we can assess the predictions based on docking energy functions. We use the AMBER force field for energy calculation. Since protein and the

nucleotides are not covalently linked, we only considered van der Waals and electrostatic energies. The predicted nucleotide structures with an energy that suggests weak interaction between this nucleotide and the protein are discarded.

Step 10. *Generate the predicted binding sites and binding residues*. The NSiteMatch method predicts both binding sites and binding residues. For each of the superimposed nucleotide structure, we calculate its geometrical center and the residues that interact with this structure. The binding sites and binding residues are generated separately. For the generation of the binding sites, the geometrical centers are clustered based on the distances between them. Two geometrical centers are assigned to the same cluster if the distance between them is less than 4Å. The clusters are ranked by the number of centers of each cluster. We use the geometrical center of all centers of one cluster to represent this cluster. The geometrical centers of the top n clusters are outputted as the predicted binding sites; by default n=5. For each residue in the predicted protein, we count the number of nucleotide structures that the residue interacts with. The residues are sorted and scored by these counts in the descending order. The scores are used to annotate a given residue as binding or non-binding based on a cutoff threshold. We selected 2 thresholds that result in predictions that match the highest precision or recall, respectively, achieved by the other methods, including Findsite, MetaPocket, and Q-SiteFinder.

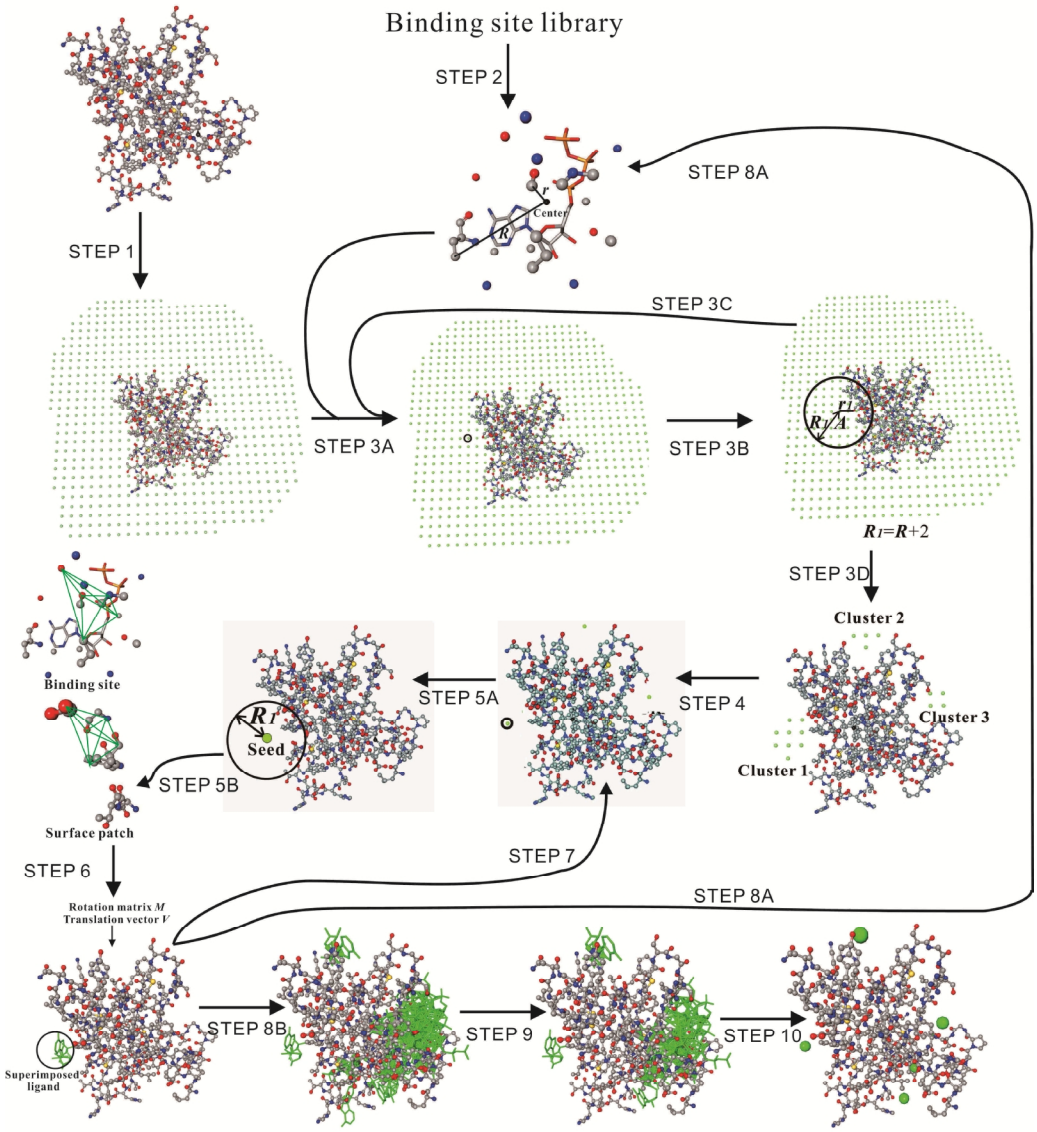


Figure 7-1: The overall flow of the NSiteMatch algorithm, which includes 10 steps. The details of the algorithm are given in section 7.5.2.

7.6 Evaluation protocol

7.6.1 Evaluation measures

The NSiteMatch, Findsite, MetaPocket, and Q-SiteFinder generate both the coordinates of the binding sites and the list of the binding residues. Therefore, their predictions are evaluated at two levels:

Evaluation of the predicted coordinates of the binding sites by using D_{CC}, which is the minimal *distance* from the *center* of the predicted binding site to the *center* of the ligand. The D_{CC} index was used in the evaluation of binding site predictors in a few recent studies (Chen et al., 2011; Skolnick and Brylinski, 2008) and was discussed in Chapter 5. For a given predicted protein with n native nucleotide-binding sites we take the top n predictions for each of the four considered methods. A given binding site is assumed to be correctly predicted if the minimal D_{CC} between this site and any of the n predictions from a given method is below a threshold D. We calculate a success rate over the entire dataset for a given value of D, which is defined as the number of correctly predicted binding sites divided by the total number of sites.

Evaluation on the predicted binding residues. A given residue is defined as the binding residue if a non-hydrogen atom of that residue is within 3.9Å to a non-hydrogen atom of the nucleotide. The same 3.9Å threshold was used in the investigation of protein-DNA and protein-small ligand interactions (Chen and Kurgan, 2009; Luscombe *et al*., 2001). For a given predicted protein we extract the binding residues for the top n predictions generated by each of the four methods and we compare them with the native binding residues using the following three measures

Precision (PREC) = $TP/(TP+FP)$
Recall (REC) = $TP/(TP+FN)$
MCC = $(TP*TN–FP*FN) / \text{sqrt}[(TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)]$

where *TP* (true positives) and *TN* (true negatives) are the counts of correctly predicted binding and non-binding residues, respectively, *FP* (false positives) are the non-binding residues that were predicted as the binding residues, and *FN* (false negatives) are the binding residues that were predicted as the non-binding residues.

7.6.2 Evaluation setup

The NSiteMatch method is compared with the Findsite, MetaPocket and Q-SiteFinder on three benchmark datasets that concern ADP-protein, ATP-protein and AMP-protein interactions, respectively. The NSiteMatch and Findsite are template-based methods and therefore the predictive quality of these two methods depends on the similarity between the predicted protein and the template library. We use four filters to assess the ability of these two methods to predict binding sites on proteins that are dissimilar to the template library. For a predicted protein, we use only the template structures that share at most 40% sequence similarity and that are in a different protein family, superfamily, and fold to the predicted protein, respectively; the latter three filters are based on the SCOP annotations (Andreeva et al., 2008). Proteins that lack the SCOP labels were not used to perform the evaluation for the homology-based filters, but they were used to asses with the 40% identity filter. We note that the first, sequence similarity-based filter may use templates from the same family. A study that analyzed SCOP annotations demonstrated that pairs of proteins that share > 25% sequence similarity are assigned to same protein family in 99% of the cases (Levitt, 2007). Similarly, the CATH database (Cuff et al., 2009), which is another protein homology classification system, automatically assigns two proteins that share > 35% sequence similarity to the same family. The evaluation of the NSiteMatch and Findsite are based on the jackknife test, where each protein in the dataset is selected once as the test/predicted protein and the remaining chains are used as the template library.

The predictions for the template-free MetaPocket and Q-SiteFinder were performed using the corresponding web servers. This means that it is possible that some of the predicted proteins were used to build these predictive models. This should not lead to a significant advantage since both of these methods use prediction models that do not utilize templates and which were computed using a large dataset of diverse proteins-ligand complexes.

7.6.3 Statistical analysis

The statistical analysis follows the procedures for the comparative analysis of existing binding site predictors, which was performed in Chapter 5. For a protein with n binding sites we take the top n predictions for every considered prediction method. For each of the n binding site, the minimum distance is calculated between this site and the top n predictions. Consequently, for a dataset with m binding sites, a set of minimal distances $\{d_i;\ i=1,2,\ldots,m\}$ are generated for each method. We assume that the predictions from different methods that are farther than 10Å away from the native site are equally wrong, i.e., they are too far away to be meaningful, and thus we round them down to 10Å. The significance of the differences between a given pair of predictors is measured by evaluating the corresponding, for the same m, minimal distance values. Since the distances for the considered predictors are not normally distributed, per the Shapiro-Wilk test of normality at $p = 0.05$, we use the non-parametric Wilcoxon signed-rank test. We assume that the differences are significant if $p < 0.05$.

7.7 Results

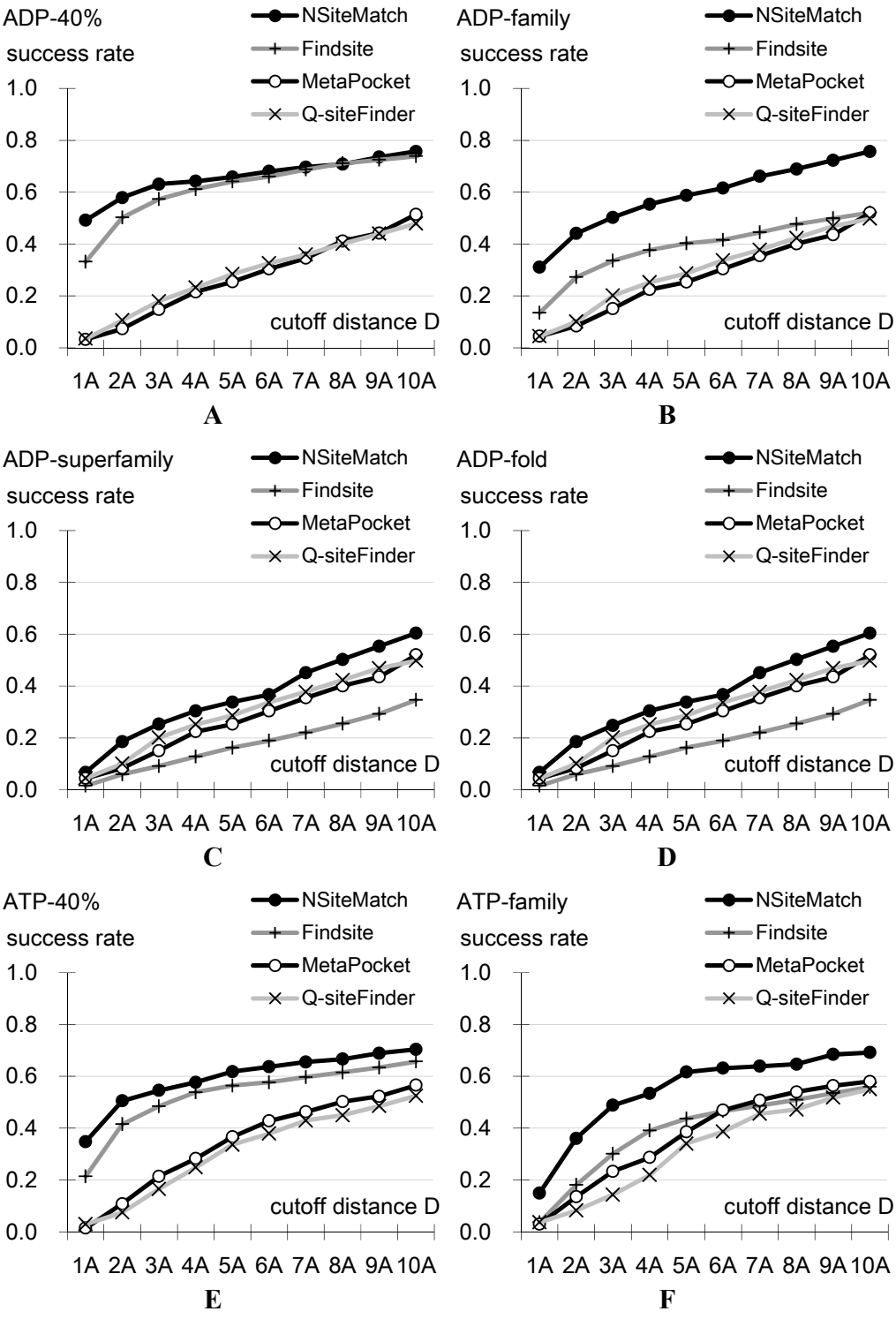
7.7.1 Evaluation of the predicted binding sites

The success rates of NSiteMatch, Findsite, MetaPocket and QsiteFinder quantified using D_{CC}, which measures the distance from the center of the predicted site to the center of the ligand in its native location are shown in Figure 7-2. For each nucleotide type, the success rates are calculated using the four filters: the 40% sequence similarity, and the family-, superfamily- and fold-based homology.

For the 40% sequence similarity filter, the NSiteMatch and Findsite achieve higher success rates than the Q-SiteFinder and MetaPocket for the three types of the nucleotides; see panels A, E and I in Figure 7-2. For the cutoff $D = 4$Å, which was suggested by Skolnick and colleagues (Skolnick and Brylinski, 2008), the success rates of NSiteMatch, Findsite, MetaPocket and Q-SiteFinder are 64%,

61%, 22%, and 23% for the ADP; 58%, 54%, 28% and 25% for the ATP; and 43%, 38%, 23% and 31% for the AMP, respectively. When considering the cutoff distances *D* between 1Å and 5Å, NSiteMatch achieves 7-8%, 16-43%, and 21-45% higher success rates than the Findsite, Q-SiteFinder and MetaPocket, respectively.

At the family level, the NSiteMatch again outperforms the remaining methods; see panels B, F and J in Figure 7-2. For the cutoff D = 4Å, the success rates of NSiteMatch are 55%, 53%, and 41% for the ADP, ATP, and AMP, respectively. To compare, the corresponding success rates for the Findsite, Q-SiteFinder and MetaPocket are 38%, 25%, and 23% for the ADP; 39%, 22%, and 29% for the ATP; and 19%, 32%, and 25% for the AMP. Although Findsite obtains higher success rates than the Q-SiteFinder and MetaPocket for the ADP and ATP, its success rates for the AMP are lower than the rates of the other two methods. This is likely because the template library for the AMP is smaller than the libraries for the ADP and ATP. When we exclude the proteins for which the SCOP label is not assigned and thus which cannot be used for the prediction when the homology filter is applied, the template libraries contain 158, 114, and 66 structures for the ADP, ATP, and AMP, respectively. The number of the available templates is even smaller once we also exclude the structures that belong to the same family as the predicted protein. Consequently, the lower success rates of Findsite and NSiteMatch for the AMP, when compared with the ATP and ADP, are due to the fact that fewer templates can be used. Moreover, the rates of the Q-SiteFinder and MetaPocket are relatively similar across the three nucleotides since these methods do not utilize templates. The Q-SiteFinder and MetaPocket methods also should not be sensitive to the filter.


ADP-40%
success rate
NSiteMatch
Findsite
MetaPocket
Q-siteFinder
1.0
0.8
0.6
0.4
0.2
0.0
cutoff distance D
1A 2A 3A 4A 5A 6A 7A 8A 9A 10A
A
ADP-family
success rate
NSiteMatch
Findsite
MetaPocket
Q-siteFinder
1.0
0.8
0.6
0.4
0.2
0.0
cutoff distance D
1A 2A 3A 4A 5A 6A 7A 8A 9A 10A
B
ADP-superfamily
success rate
NSiteMatch
Findsite
MetaPocket
Q-siteFinder
1.0
0.8
0.6
0.4
0.2
0.0
cutoff distance D
1A 2A 3A 4A 5A 6A 7A 8A 9A 10A
C
ADP-fold
success rate
NSiteMatch
Findsite
MetaPocket
Q-siteFinder
1.0
0.8
0.6
0.4
0.2
0.0
cutoff distance D
1A 2A 3A 4A 5A 6A 7A 8A 9A 10A
D
ATP-40%
success rate
NSiteMatch
Findsite
MetaPocket
Q-siteFinder
1.0
0.8
0.6
0.4
0.2
0.0
cutoff distance D
1A 2A 3A 4A 5A 6A 7A 8A 9A 10A
E
ATP-family
success rate
NSiteMatch
Findsite
MetaPocket
Q-siteFinder
1.0
0.8
0.6
0.4
0.2
0.0
cutoff distance D
1A 2A 3A 4A 5A 6A 7A 8A 9A 10A
F

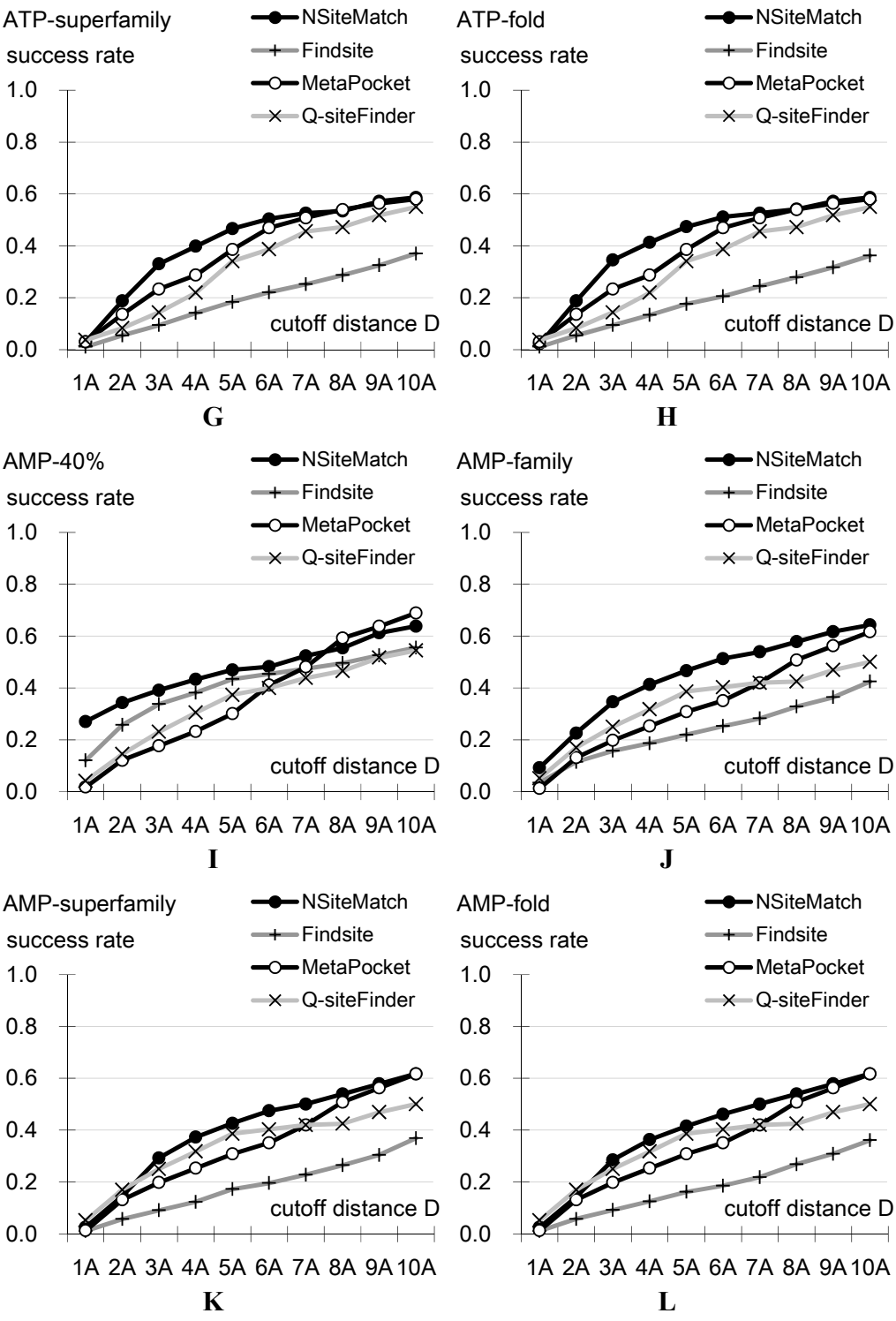
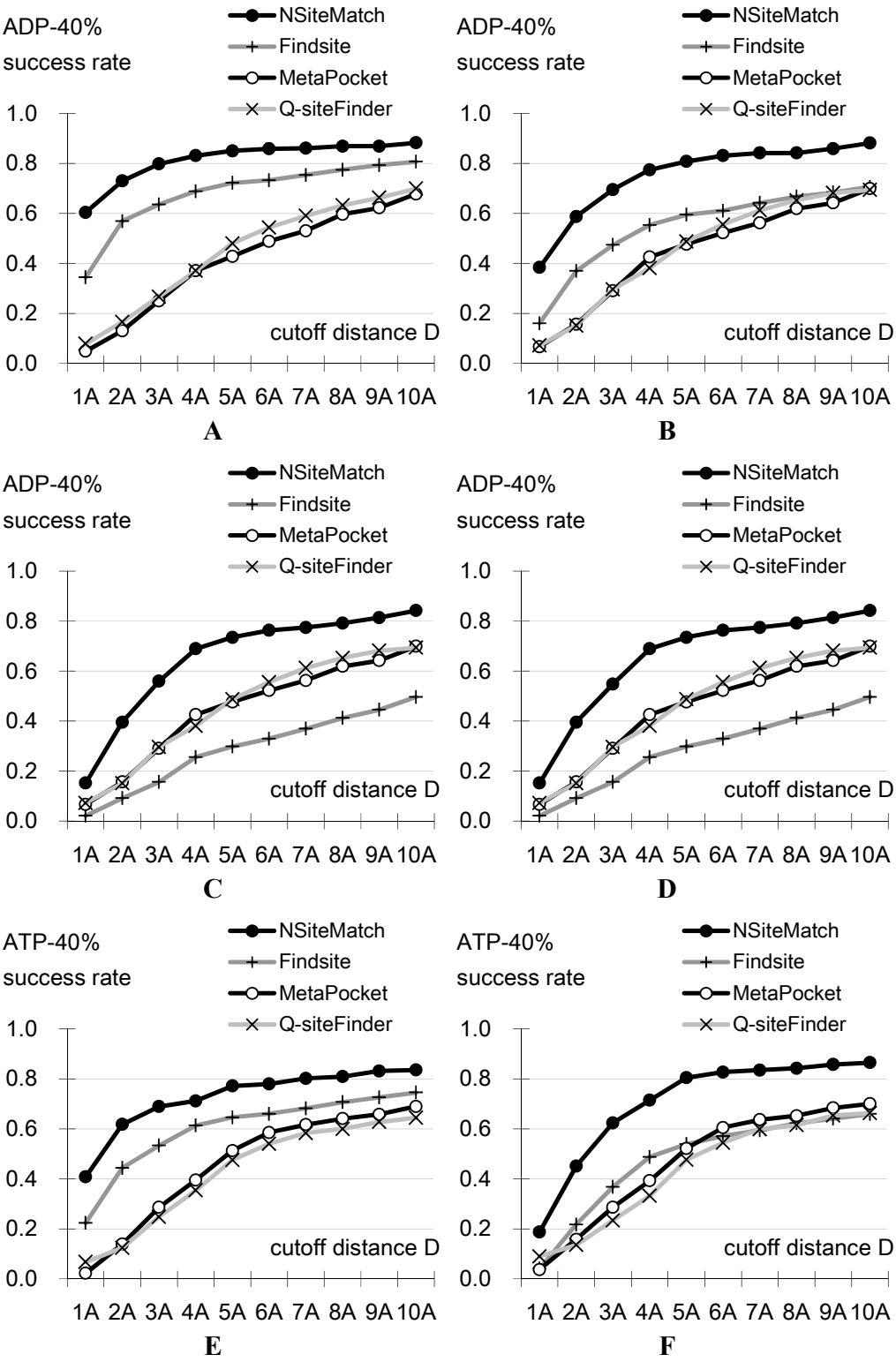


Figure 7-2: The success rates (y-axis) of the NSiteMatch and the three competing methods (Findsite, MetaPocket, and Q-SiteFinder) measured using D_{CC} (the minimal distance from the center of the predicted site to the center of the ligand)

on the benchmark datasets. A given binding site is regarded as correctly predicted if the minimal distance between this site and the top *n* predictions is below the cutoff distance *D* (*x*-axis), where *n* is the number of binding sites of the protein that includes the evaluated binding site. All methods are evaluated at 4 filter levels, the 40% sequence similarity level (panels A, E and I), family level (panels B, F and J), superfamily level (panels C, G and K) and fold level (panel D, H and L). Panels A, B, C and D show results for the ADP. Panels E, F, G and H show results for the ATP and panels I, J, K and L show results for the AMP. The 40% sequence similarity level indicates that all chains in the template library that were used for the prediction share less than 40% sequence similarity to the test protein. The family, superfamily and fold levels indicate that all chains in the template library that were used for the prediction are classified as belonging to a different family, superfamily and fold (annotated using the SCOP database), respectively, when compared with the annotation of the test protein.

The results at the superfamily and fold levels are similar; see panels C, D, G, H, K and L in Figure 7-2. Although the NSiteMatch still achieves higher success rates than the remaining methods, the corresponding improvements are smaller. When considering the cutoff distances *D* between 1Å and 5Å, the NSiteMatch achieves 14-20%, 1-12%, and 7-8% better success rates than the Findsite, Q-SiteFinder, and MetaPocket, respectively. We observe that at the superfamily and fold filter levels, the success rates of the Q-SiteFinder and MetaPocket are higher than the success rates of Findsite.

We also calculate the success rates of the NSiteMatch, Findsite, MetaPocket, and Q-SiteFinder quantified using D_{CC} by taking the top 5 predictions for every predicted protein, see Figure 7-3. Although the success rates of the four methods are improved due to the inclusion of additional predictions, the relative ranking does not change when compared with the evaluations based on the *n* predictions. For instance, at the 40% sequence similarity level, the NSiteMatch achieves success rates that are higher than the rates of the other three methods, and Findsite is the runner-up. At the superfamily and fold filter levels, the NSiteMatch outperforms the MetaPocket and Q-SiteFinder, which in turn improve over the Findsite.


ADP-40%
success rate
NSiteMatch
Findsite
MetaPocket
Q-siteFinder
1.0
0.8
0.6
0.4
0.2
0.0
cutoff distance D
1A 2A 3A 4A 5A 6A 7A 8A 9A 10A
A
B
C
D
ATP-40%
E
F

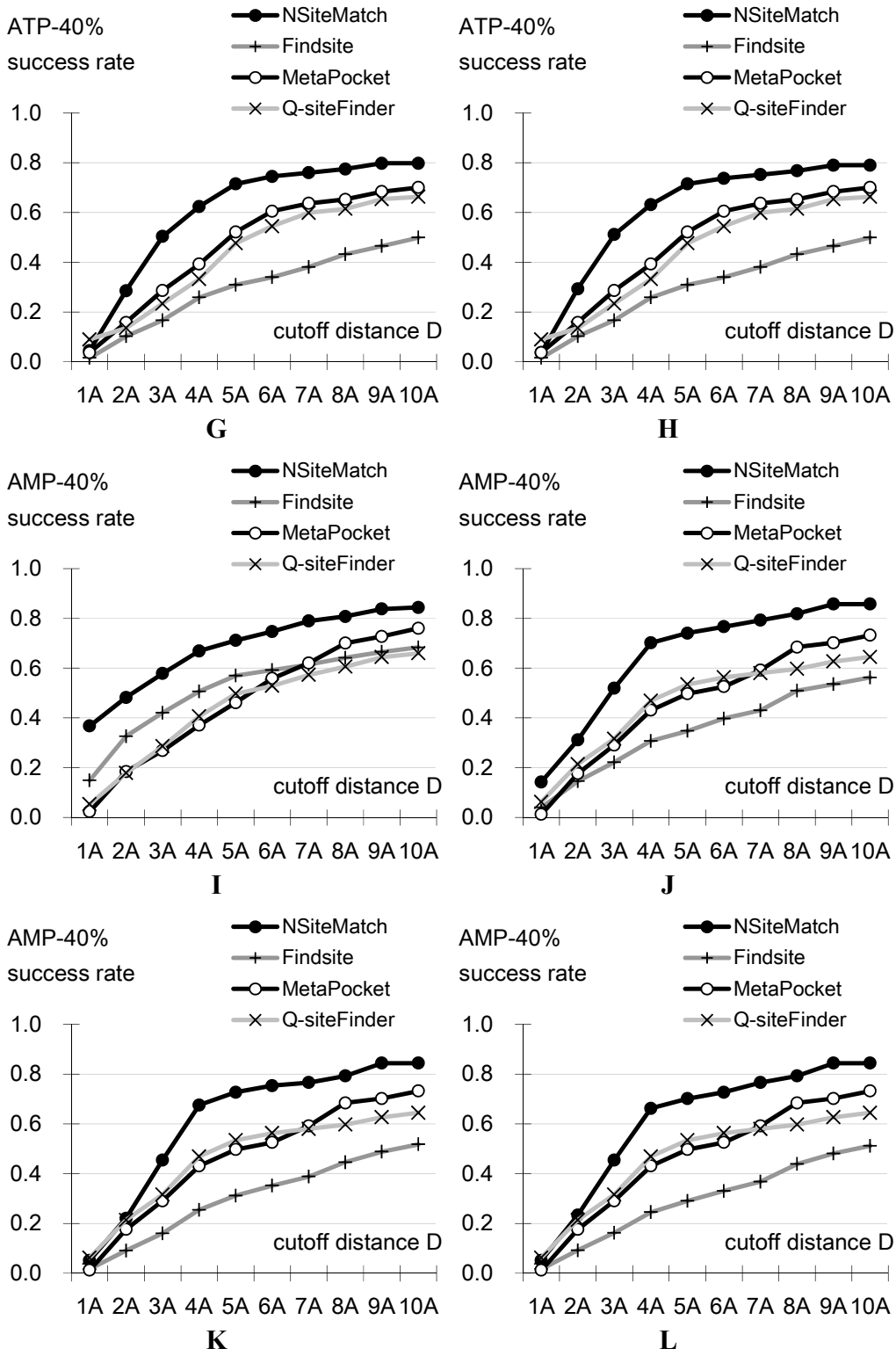


Figure 7-3: The success rates (*y*-axis) of the NSiteMatch and the three competing methods (Findsite, MetaPocket, and Q-SiteFinder) measured using D_{CC} (the

minimal distance from the center of the predicted site to the center of the ligand) on the benchmark datasets. A given binding site is regarded as correctly predicted if the minimal distance between this site and the top 5 predictions is below the cutoff distance *D* (*x*-axis). All methods are evaluated at 4 filter levels, the 40% sequence similarity level (panels A, E and I), family level (panels B, F and J), superfamily level (panels C, G and K) and fold level (panels D, H and L). Panels A, B, C and D show results for the ADP. Panels E, F, G and H show results for the ATP and panels I, J, K and L show results for the AMP. The 40% sequence similarity level indicates that all chains in the template library that were used for the prediction share less than 40% sequence similarity to the test protein. The family, superfamily and fold levels indicate that all chains in the template library that were used for the prediction are classified as belonging to a different family, superfamily and fold (based on the SCOP database), respectively, when compared with the annotation of the test protein.

Among the four predictors, the NSiteMatch and Findsite are template-based and therefore they depend on the availability of suitable templates. As expected, Figure 7-4 reveals that the predictive quality of these two methods declines with the decrease of the structure similarity to the templates, i.e., when more distant homologs are used. For instance, when considering the cutoff $D = 4$Å, the success rates of the NSiteMatch are 64%, 55%, 31%, and 31% for the ADP at the 40% sequence similarity, family, superfamily, and fold filter levels, respectively (panel A in Figure 7-4). Similarly, for the ATP and AMP, the success rates of the NSiteMatch are 58% and 43%, 53% and 41%, 40% and 37%, and 41% and 36% for the four filters, respectively (panels B and C in Figure 7-4). Similar declining trends are observed for the Findsite; see panels D, E, and F in Figure 7-4. Although the results indicate that the availability of similar templates has a relatively strong impact on the predictive quality of these two predictors, we note that the NSiteMatch maintains higher success rates when predicting with the help of more distant homolog. This advantage is due to the use of local similarity and consequently, as shown in Figure 7-2, our method also outperforms the two template-free methods, MetaPocket and Q-SiteFinder, at the superfamily and fold levels even for the hard (characterized by the small template library) AMP ligand. To compare, these two approaches improve over the Findsite when the templates are filtered at the superfamily and fold levels for each of the three ligands.

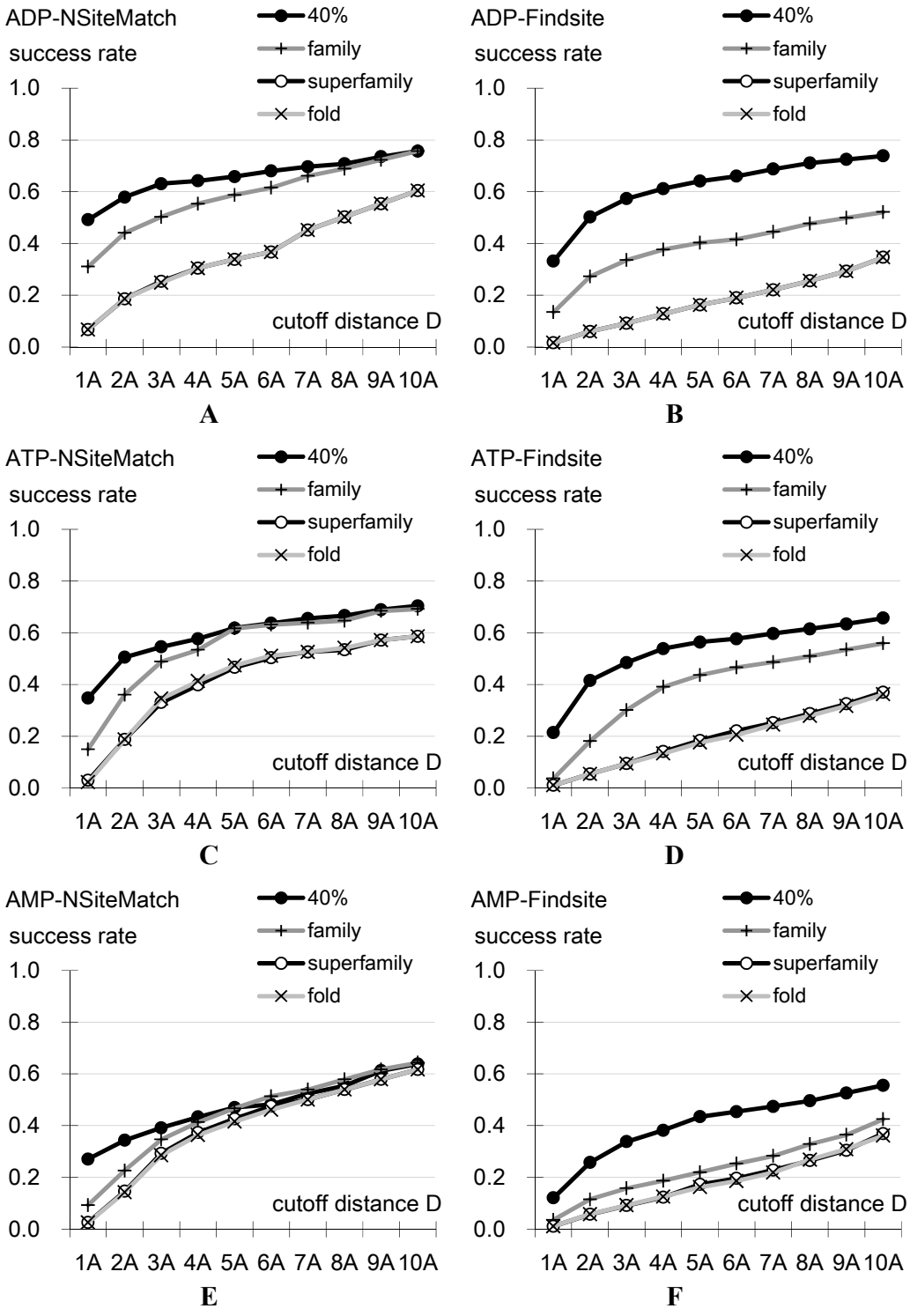


Figure 7-4: The relation between the predictive quality of the NSiteMatch and Findsite and the similarity between the predicted protein and template library. The success rates (y-axis) are measured using D_{CC} (the minimal distance from the

center of the predicted site to the center of the ligand) on the benchmark datasets. A given binding site is regarded as correctly predicted if the minimal distance between this site and the top n predictions is below the cutoff distance D (x-axis), where n is the number of binding sites of the protein that includes the evaluated binding site. Panels A, C, and E evaluate results of the NSiteMatch for the ADP, ATP, and AMP, respectively; Panels B, D, and F summarize the corresponding results for the Findsite.

We investigate significance of differences in the prediction quality measured with D_{CC} between NSiteMatch and the other predictors, see Table 7-1. We compare the D_{CC} values that are calculated by taking top n predictions for each protein where n is the number of the nucleotide-binding sites for a given protein. At 40% sequence similarity and family levels, the NSiteMatch is significantly better than the other three methods. Similarly, our method significantly outperforms the competing solutions by using the superfamily and fold filters for the ADP and ATP ligands, and the improvements are not significant only when compared with the MetaPocket and Q-SiteFinder for the AMP.

Table 7-1: Statistical significance of the differences in distances measured using D_{CC} between the predicted and the actual location of the binding site measured using Wilcoxon signed-rank test. The "+" indicates that NSiteMatch is significantly better than a method in a given column with $p < 0.05$ and "=" denotes that NSiteMatch and a method in a given column is not significantly different.

Ligand type	SCOP level	MetaPocket	Findsite	Q-SiteFinder
ADP	40%	+	+	+
	Family	+	+	+
	Superfamily	+	+	+
	Fold	+	+	+
ATP	40%	+	+	+
	Family	+	+	+
	Superfamily	+	+	+
	Fold	+	+	+
AMP	40%	+	+	+
	Family	+	+	+
	Superfamily	=	+	=
	Fold	=	+	=

7.7.2 Evaluation of the predicted binding residues

Besides the coordinates of the predicted binding site, the NSiteMatch, Findsite, MetaPocket, and Q-SiteFinder also predict the binding residues. For the NSiteMatch, each residue in the predicted protein structure is assigned with a numerical score which indicates the number of ligands that this residue interacts with (details concerning the annotation of the binding residues for the NSiteMatch are given in 7.5.2). A given residue is regarded as a binding residue if its score is above a certain threshold. The selection of this threshold controls the trade-off between precision (fraction of the correctly predicted binding residues among all predicted binding residues) and recall (fraction of the correctly predicted binding residues among all native binding residues). Since the precision and recall values achieved by the Findsite, MetaPocket, and Q-SiteFinder vary substantially, we selected two thresholds that allow for a direct comparison. Similarly as in (Zhang et al., 2008), we set the threshold such that the precision/recall of the NSiteMatch is equal to the highest precision/recall achieved by the other methods for a given ligand and a given filter. The predictions are evaluated based on the recall (also called sensitivity), precision, and MCC; see Table 7-2. The MCC quantifies correlation between predictions and the native annotations and thus higher MCC values correspond to more accurate predictions.

Table 7-2: Comparison of the predictive qualities of the NSiteMatch, MetaPocket, Findsite and Q-SiteFinder for the prediction of binding residues for ADP, ATP, and AMP. The PRE, REC, and MCC stand for precision, recall, and Matthews Correlation Coefficient, respectively. The NSiteMatch generates a real value for each residue (propensity to bind), which is thresholded to make binary (binding vs. non-binding residue) predictions. The rows annotated as the "NSiteMatch ^{p}", are based on the thresholds that generate precision values which match the highest precision obtained by the MetaPocket, Findsite, and Q-SiteFinder for a given ligand type; similarly, the "NSiteMatch ^{r}" rows correspond to thresholds for which the highest value of recall is matched. The matching recall and precision values are shown in italics and the highest MCC values are given in bold font.

Ligand type	Method	40%			Family			Superfamily			Fold		
		PRE	REC	MCC	PRE	REC	MCC	PRE	REC	MCC	PRE	REC	MCC
ADP	NSiteMatch ^{p}	*0.48*	0.79	0.6	*0.43*	0.68	0.52	*0.43*	0.49	**0.44**	*0.43*	0.48	**0.43**
	NSiteMatch ^{r}	0.76	*0.53*	**0.62**	0.53	*0.57*	**0.53**	0.37	*0.57*	0.43	0.37	*0.57*	**0.43**
	MetaPocket	0.41	0.13	0.21	*0.43*	0.13	0.22	*0.43*	0.13	0.22	*0.43*	0.13	0.22
	Findsite	*0.48*	0.67	0.55	0.41	0.45	0.42	0.31	0.32	0.3	0.31	0.3	0.3
	Q-SiteFinder	0.29	*0.53*	0.36	0.31	*0.57*	0.39	0.31	*0.57*	0.39	0.31	*0.57*	0.39
ATP	NSiteMatch ^{p}	*0.49*	0.68	**0.56**	*0.46*	0.54	**0.47**	*0.46*	0.41	0.41	*0.46*	0.41	0.41
	NSiteMatch ^{r}	0.61	*0.52*	0.54	0.47	*0.51*	**0.47**	0.4	*0.51*	**0.42**	0.4	*0.51*	**0.42**
	MetaPocket	0.47	0.16	0.26	*0.46*	0.14	0.23	*0.46*	0.14	0.23	*0.46*	0.14	0.23
	Findsite	*0.49*	0.52	0.5	0.38	0.43	0.39	0.29	0.34	0.31	0.29	0.34	0.31
	Q-SiteFinder	0.31	*0.52*	0.36	0.29	*0.51*	0.35	0.29	*0.51*	0.35	0.29	*0.51*	0.35
AMP	NSiteMatch ^{p}	*0.47*	0.62	**0.51**	*0.47*	0.36	0.39	*0.47*	0.3	0.35	*0.47*	0.28	0.34
	NSiteMatch ^{r}	0.47	*0.62*	**0.51**	0.33	*0.6*	**0.41**	0.29	*0.6*	**0.38**	0.29	*0.6*	**0.38**
	MetaPocket	*0.47*	0.16	0.26	*0.47*	0.15	0.25	*0.47*	0.15	0.25	*0.47*	0.15	0.25
	Findsite	0.44	0.49	0.44	0.31	0.34	0.31	0.29	0.33	0.3	0.29	0.32	0.29
	Q-SiteFinder	0.29	*0.62*	0.39	0.29	*0.6*	0.38	0.29	*0.6*	**0.38**	0.29	*0.6*	**0.38**

For the 40% sequence similarity filter, the NSiteMatch achieves higher precision, recall and MCC values than the Findsite, Q-SiteFinder, and MetaPocket for all three types of the nucleotides. The NSiteMatch generates predictions with a substantially higher precision when its recall is the same as the highest recall produced by the other predictors. Similarly, our method has higher recall when its precision matches the highest precision produced by the other methods. The Findsite obtains the second best MCC values for the three types of the nucleotides. We observe that predictions of MetaPocket are characterized by the precision that is higher than the recall, while the Q-SiteFinder has the recall values higher than the precision. This indicates that the MetaPocket and Q-SiteFinder under- and over-predict the binding residues, respectively.

At the family level, the NSiteMatch also provides the highest precision, recall, and MCC values when compared with the other methods for the three nucleotides. However, as expected, the predictive quality of the NSiteMatch and Findsite declines when compared to the 40% sequence similarity filter. Based on the MCC value, Findsite outperforms the Q-SiteFinder and MetaPocket for the ADP and ATP but is inferior to the Q-SiteFinder for the AMP. The results at the superfamily and fold level filters are similar to each other. The NSiteMatch maintains the highest precision, recall, and MCC values for the ADP and ATP. However, for the AMP, the predictions of the NSiteMatch have quality that is comparable to the Q-SiteFinder and higher than the MetaPocket and Findsite.

The results concerning prediction of binding residues are consistent with our statistical analysis based on the D_{CC} values. The lower predictive quality of the NSiteMatch (and Findsite) for the AMP, when compared with the ADP and ATP, is due to the relatively small size of the template library for that ligand.

Similarly as for the prediction of the binding sites, we assessed the impact of the similarity between the predicted protein and the corresponding template library on the predictive qualities of the NSiteMatch and Findsite for the prediction of binding residues. The MCC values achieved by NSiteMatch for the ADP, ATP, and AMP are 0.6, 0.56, and 0.51, respectively, at the 40% sequence similarity level; 0.52, 0.47, and 0.39, respectively, at the family level; 0.44, 0.41, and 0.35, respectively, at the superfamily level; and 0.43, 0.41, and 0.34, respectively, at the fold level. As expected, the results indicate that NSiteMatch generates better predictions when the predicted protein has a higher structural similarity to the template library. Similar relation is observed for the Findsite; see Table 7-2. However, based on the MCC values, the Findsite is outperformed by the template-free Q-SiteFinder for the prediction of binding residues of the three nucleotides for the superfamily and fold filters. Importantly, we observe that our method outperforms Findsite for each filter and each ligand type, and it also improves over the Q-SiteFinder and MetaPocket, except for the AMP with the superfamily and fold level filters where it provides predictive quality that is comparable to the

Q-SiteFinder. This demonstrates that our local similarity-based approach provides one of the best solutions for the structure-based prediction of nucleotide binding residues, even when predicting for structures from novel/uncharacterized folds and superfamilies.

7.7.3 Case studies

We present two case studies. The first compares the utility of the NSiteMatch and the existing binding site predictors, and the second demonstrates the ability of the NSiteMatch to identify similar binding sites across protein folds.

We use the chain A of the MJ1225 protein (PDB code: 3KH5) (Gómez-García et al., 2010) for the first case study. This structure was released after our benchmark dataset was created and this sequence shares less than 25% similarity to any sequence in our benchmark dataset. We used the web servers of the MetaPocket and Q-SiteFinder and the standalone implementation of the Findsite and our NSiteMatch to generate the predictions. The template library of Findsite and NSiteMatch includes all structures from the benchmark dataset. Since the MJ1225 protein includes 3 ADP-binding sites and 1 AMP-binding site, the top 4 predictions generated by each predictor were assessed. For the cutoff distance D = 4Å, the NSiteMatch and Findsite correctly predict 4 and 3 of the binding sites, respectively, while the Q-SiteFinder and MetaPocket find 2 and 1 of the binding sites, respectively, see panel A in Figure 7-5. The lower quality of the Q-SiteFinder and MetaPocket predictions can be explained by the fact that these methods predict sites for a generic class of small ligands, while Findsite and NSiteMatch use a library that is specific to the three nucleotides. In spite of using the same template library, we show that NSiteMatch is more accurate that Findsite, which is due to the use of the local similarity.

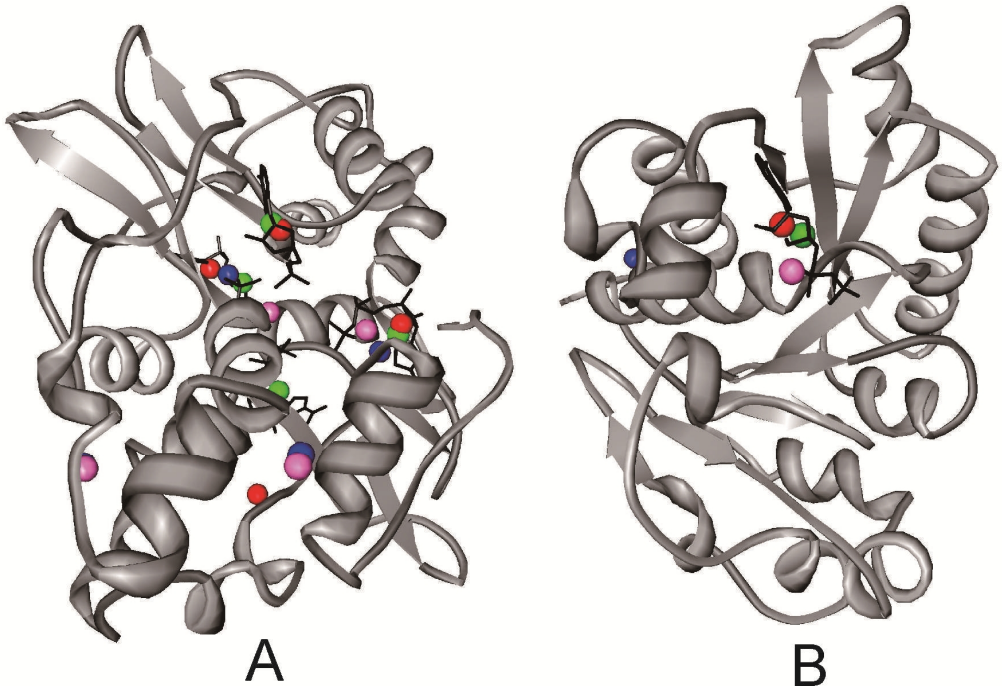


Figure 7-5: Binding sites predicted by the NSiteMatch, Findsite, MetaPocket, and Q-SiteFinder for chain A of the MJ1225 protein (panel A) and chain A of the cell division inhibitor mind protein (panel B). The predictions by NSiteMatch, Findsite, MetaPocket, and Q-SiteFinder are denoted with green, red, purple, and blue spheres, respectively. The ligands are in the stick format and are colored in black. The MJ1225 contains 3 ADP-binding sites and 1 AMP-binding site and the top 4 predictions from each method are shown. The cell division inhibitor mind protein has 1 ADP-binding site and the top prediction for each method is shown.

We use chain A of the probable cell division inhibitor mind protein (PDB code: 1ION) (Sakai et al., 2001) to demonstrate that NSiteMatch is capable of identifying similar binding sites across protein folds. This would imply that the function of a given protein could be inferred from other proteins that have different topologies. This structure includes 1 ADP-binding site and thus we assess the top prediction from each method. The distances between the predicted and the native center of the ligand are 0.6Å, 1.3Å, 3.0Å, and 22.8Å for the NSiteMatch, Findsite, Meta-Pocket, and Q-SiteFinder, respectively, see panel B in Figure 7-5. The ADP-binding site is implemented by the "GTGKTT" sequence segment and this protein is assigned to the "P-loop containing nucleoside triphosphate hydrolases" superfamily based on the SCOP annotation. The

NSiteMatch uses potentially multiple templates to find a single binding site. We analyze the templates that the NSiteMatch finds as similar to the 1ION protein in the predicted binding region and which were used to predict this site. Three of these templates belong to superfamilies that are different to the superfamily of the predicted protein; see Table 7-3. The first template is chain A of phosphoenolpyruvate carboxykinase (PDB code: 1K3C) (Sudom et al., 2001), which is assigned to the "PEP carboxykinase-like" superfamily in SCOP. The other two templates are chain A of UDP-N-Acetylmuramoylalanine-D-Glutamate ligase (PDB code: 2JFG) (Kotnik et al., 2007) and chain A of Thermosome alpha subunit (PDB code: 1Q3S) (Shomura et al., 2004), which belong to the "MurD-like peptide ligases" and "catalytic domain and GroEL equatorial domain-like" superfamilies, respectively. We superimpose these three templates into the predicted, 1ION protein using Fr-TM-align (Pandit and Skolnick, 2008); see panels A, B and C in Figure 7-6. The Figures reveal that the templates are dissimilar in their overall topology when compared with the 1ION protein. The alignment of the binding segments for the three templates and the predicted protein, which is given in Table 7-3, reveals that they share key binding residues, i.e., the Gly, Lys, and Thr residues. The NSiteMatch works by finding local similarity in the binding region between the predicted and the template proteins, and we superimposed these regions; see panels D, E, and F in Figure 7-6 where the residues are displayed in ball and stick format and the ADP is shown in the stick format. The binding site of the phosphoenolpyruvate carboxykinase is very similar to the binding site of the predicted protein (panel D in Figure 7-6); we found 30 atoms which overlap between these two superimposed sites. The overlap between the binding site of the UDP-N-Acetylmuramoylalanine-D-Glutamate ligase and the predicted protein includes 16 atoms (panel E in Figure 7-6) which mainly involve the Gly114, Lys115, and Thr117 residues on the template and the Gly15, Lys16, and Thr18 residues on the predicted chain. The binding site of the thermosome alpha subunit is less similar to the predicted protein when compared with the other two templates (panel F in Figure 7-6); 11 atoms overlap and they correspond to Gly96, Thr98, and Thr99 residues on the template and Gly15,

Thr17, and Thr18 residues on the predicted sequence. We observe that the ADP binds to the predicted protein and the first two templates mainly through the β-phosphate group, while it interacts with the third template mainly through the α-phosphate group, which explains the lower similarity. However, even when the interaction group changes, the NSiteMatch was still able to capture a similar spatial arrangement of residues at the binding site. This example demonstrates that our method can perform annotation of binding sites based on templates with distant homology. In contrast to the NSiteMatch, the templates used by Findsite to predict the 1ION protein belong to the same "P-loop containing nucleoside triphosphate hydrolases" superfamily, i.e., Findsite was not able to capture the distant functional relationship between proteins from different superfamilies.

Table 7-3: The templates identified by NSiteMatch for the probable cell division inhibitor mind protein. Three templates, including phosphoenolpyruvate carboxykinase, UDP-N-Acetylmuramoylalanine-D-Glutamate Ligase and thermosome alpha subunit have different topologies but similar binding segments and binding sites to the predicted protein.

Polymer name	PDBcode:(chain)	Binding segment	SCOP label
Cell division inhibitor mind protein	1ION:A	G T - G K T T	c.37.1.10
Phosphoenolpyruvate carboxykinase	1K3C:A	G T - G K T T	c.91.1.1
UDP-N-Acetylmuramoylalanine-D-Glutamate ligase	2JFG:A	G S N G K S T	c.72.2.1
Thermosome alpha subunit	1Q3S:A	G D - G T T T	a.129.1.2

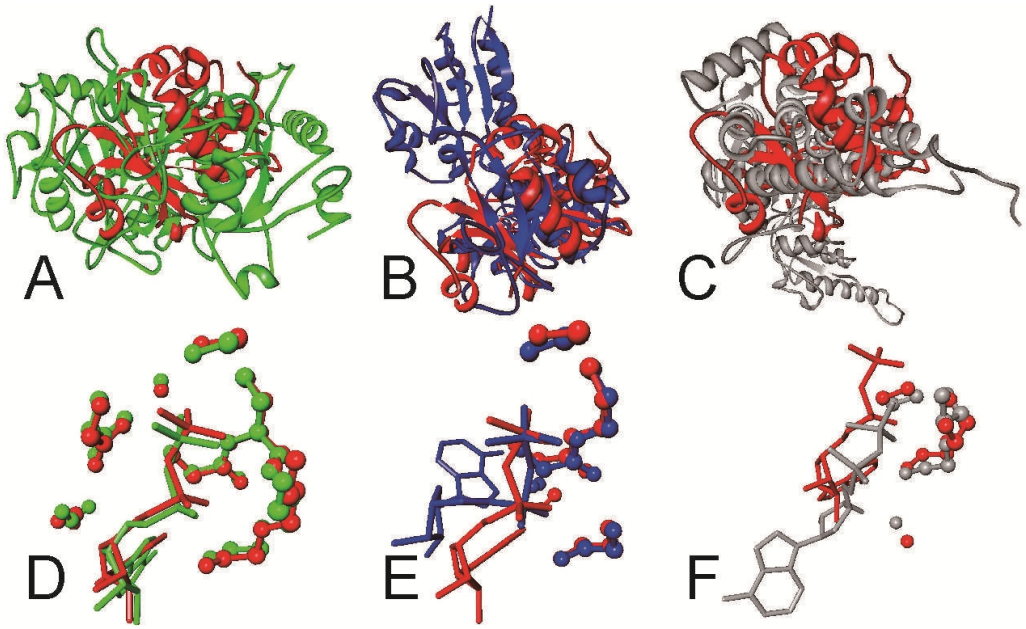


Figure 7-6: Comparison of the structures of templates identified by the NSiteMatch as similar to the chain A of the probable cell division inhibitor mind protein (PDB code: 1ION), which are classified as belonging to different superfamily as the 1ION protein. The 1ION structure is shown in red, while the three templates, phosphoenolpyruvate carboxykinase, UDP-N-Acetylmuramoylalanine-D-Glutamate ligase, and thermosome alpha subunit are in green, blue, and grey, respectively. Panels A, B, and C superimpose each of the templates to the 1ION structure by using Fr-TM-align. Panels D, E, and F are the common sub-structures between the 1ION structure and a given template, which were identified by the NSiteMatch. The residues are displayed in the ball and stick format and the ADP is shown in the stick format.

7.8 Conclusions

Motivated by the importance and the substantial interest in protein-nucleotide interactions and the lack of accurate computational predictors, we designed a novel and accurate structure-based nucleotide-binding site predictor for the three most commonly occurring nucleotides. Empirical test shows that the proposed NSiteMatch method significantly outperforms generic, template-free binding site predictors, except for the AMP nucleotide, for which NSiteMatch generates results that are comparable to the Q-SiteFinder and MetaPocket at the superfamily and fold filter levels. We also show that the template-based NSiteMatch and Findsite generate better predictions for proteins that share higher similarity with

their template library. However, NSiteMatch significantly outperforms Findsite when the predicted protein shares low structural similarity to the template library. Contrary to the Findsite which relies on identification of templates that have similar topology to the topology of the predicted protein, our method recognizes templates that share local similarity in the binding area and which are not necessarily similar in their overall topology. This allows us to identify similar binding sites across potentially very different protein structures. Our method can accurately, when compared to the current state-of-the-art, find distant functional relationships between proteins from different families, superfamilies, and folds. Although the NSiteMatch targets predictions for a few specific nucleotides, our methodology constitutes a generic platform that could be extended to predict interactions with other small ligands.

CHAPTER 8 Summary and conclusions

8.1 Summary

This dissertation addresses computational characterization and prediction of protein-small ligand interactions, with emphasis on the interactions between protein and nucleotides. The four main parts include the investigation of atomic level interactions between proteins and small ligands, a comparative survey of existing structure-based binding site predictors for small organic compounds, a method that predicts the nucleotide-binding residues for protein sequences, and an algorithm that identifies the nucleotide-binding sites for protein structures.

In Chapter 4, we investigate the atomic level patterns that describe the protein-small ligand interactions. This study opens the dissertation because such analysis is crucial not only to understand and summarize these interactions, but also to investigate whether computational prediction of these interactions would be possible. Our study demonstrates that the protein- small organic compound, protein-metal ion, protein-inorganic anion, and protein-inorganic cluster interactions are governed by different interaction forces. Therefore, different interaction patterns were found for each type of interactions. The 10 proposed patterns describe 56% of the protein-small ligand complexes in PDB.

Among the four abovementioned ligand groups, we focus on the small organic compounds driven by the fact that they constitute significant majority of the drugs approved by the U.S. Food and Drug Administration. In Chapter 5, we assessed the predictive performance of the current structure-based predictors of the protein-small organic compound interactions. This study not only evaluates the current efforts, but most importantly investigates whether new and improved predictors are needed and points out potential directions for future research. We demonstrated that the predictive quality of these methods was significantly improved during the past decade. However, we found that there is a large room for further improvements, which suggests that new solutions are needed. We also

discovered several limitations of these predictors. For instance, the best-performing Findsite is largely dependent on the completeness of its template library while the runner-up methods, ConCavity, MetaPocket and Q-SiteFinder generate less accurate predictions when using the apo structures instead of using the holo structures. These limitations motivate further research on the prediction of the protein-organic compound interactions. We observe that the predictive performance of the top-performing methods varies for different organic compound types, which implies that separate models should be designed for specific types of organic compounds. Our comparative survey shows that improvements can be obtained when using a consensus-based approach and that threading is a promising approach, but given that its performance for lower similarity templates would be improved. We use these insights to design a new method, which is described in Chapter 7.

Among the different organic compound types, we focus on nucleotides because these ligands are abundant, ubiquitous, and have important functions. We proposed two novel methods, one that predicts the nucleotide-binding residues (see Chapter 6) from protein sequences and the other that identifies the nucleotide-binding sites from protein structures (see Chapter 7). These methods are designed by taking advantage of the information derived from the preceding chapters, i.e., the sequence-based method is based on a consensus of algorithms that target predictions for specific nucleotides and the structure-based method utilizes an improved (local) threading and hybridizes the geometry-, energy-, and threading-based approaches. We demonstrate that both of these methods provide statistically significant improvements over the state-of-the-art existing solutions.

The major contributions of this dissertation include:

- Creation of a dataset for the analysis of protein-small ligand interactions. The proteins in the dataset have a high quality, i.e., the structure resolution is below 2.0Å and the R-factor value is below 0.25, and they adequately sample the sequence space, i.e., the sequence similarity is reduced to 25%.

- First-of-its-kind comprehensive characterization of the molecular-level interactions between proteins and small ligands that covers both covalent and non-covalent interactions.

- Discovery of ten molecular-level interaction patterns that cover significant majority of the protein-small ligand complexes in PDB.

- Creation of a benchmark dataset for the assessment of the existing structure-based binding site predictors. The structures in the dataset are annotated with multiple binding sites (using structural alignment and clustering), the dataset samples the homology space (proteins belong to different families), and the biological-irrelevant ligands are excluded.

- First-of-its-kind comprehensive comparative analysis of the predictive performance of ten representative structure-based binding site predictors.

- Development of a quality index O_{PL}, which quantifies *overlap* between the predicted *binding site* and the *ligand* and provides additional insights into the performance of the structure-based binding site predictors.

- Assessment of the impact of the structural similarity between the predicted protein and the template library on the predictive quality of the top-performing threading-based binding site predictor, Findsite.

- Assessment of the differences between the predictions generated by top-performing structure-based predictors when using apo and holo structures, respectively.

- Assessment of the impact of the ligand size and ligand type on the predictive quality of the top-performing structure-based binding site predictors.

- Investigation of the complementarity between the predictions generated by the top-performing structure-based binding site predictors.

- Creation of three benchmark datasets for the sequence-based prediction of the nucleotide-binding residues.

- Development of five accurate sequence-based predictors (the NsitePred method) that identify binding residues for the five common nucleotides, including ATP, ADP, AMP, GTP, and GDP.

- Empirical evaluation of the NsitePred method and comparative analysis against the state-of-the-art existing sequence-based solutions.

- Investigation of the relation between the spread index and the predictive quality of the proposed NsitePred method.

- Development of an algorithm, called NSiteMatch, which predicts the nucleotide-binding sites and residues from protein structures.

- Empirical evaluation of the NSiteMatch method and comparative analysis against the state-of-the-art existing structure-based solutions.

- Development of an evaluation protocol to assess the performance of template-based binding site predictors by controlling the structural similarity between the predicted protein and the template library.

The list below outlines the most important findings of the aforementioned studies:

Investigation of the atomic level patterns in protein-small ligand interactions

- The interactions between proteins and different ligand types are governed by different interaction forces. The protein-organic compound complexes are governed by the hydrogen bonds, van der Waals contacts and covalent bond. The protein-metal ion complexes are based on the electrostatic force and coordination bonds while the protein-anion complexes are governed by the electrostatic force, hydrogen bonds and van der Waals contacts. Finally, the

protein-inorganic cluster complexes are established mostly due to the coordination bonds.

- 73% of the covalent bonds are formed between proteins and organic compounds are covered by three interaction patterns, 1) thioether bond formed between the thiol of Cys residue and the carbon atom of an organic compound; 2) disulfide bond formed between thiol of Cys residue and the sulfur atom of an organic compound; 3) the covalent contacts formed between nitrogen atom of Lys residue and the carbon atom of an organic compound.

- 65.8% of all hydrogen bonds are formed between proteins and organic compounds are established between NH- group (as the donor) of an AA and the oxygen atom of an organic compound.

- 96.4% of all coordination bonds are formed between proteins and metal ions are covered by three interaction patterns, 1) coordination bond established between metal and the nitrogen atom in the side chain of His residue; 2) coordination bond formed between metal and sulfur atom of Cys residue; 3) coordination bond established between metal and the oxygen atoms in the side chain of Asp/Glu residues;

- 87.1% of all hydrogen bonds are formed between proteins and inorganic anions are established between NH- group (as the donor) of an AA and the oxygen atom of an anion.

- 98.5% of all coordination bonds are formed between an inorganic cluster and a protein are covered by two interaction patterns, 1) coordination bond formed between the iron atom of the cluster and the sulfur atom of Cys residue; 2) coordination bond formed between the iron atom of the cluster and the nitrogen atom in the side chain of His residue.

Assessment of the existing structure-based binding site predictors

- Among the 10 considered binding site predictors, Findsite is significantly more accurate than the other considered predictors. The ConCavity, Q-SiteFinder, MetaPocket and PocketPicker methods are second-best and not significantly different between each other (except for the ConCavity which significantly improves over the Q-SiteFinder), and this group is significantly better than LIGSITE^{csc}, SURFNET, PASS, PocketFinder and Fpocket.

- The predictive performance of the threading-based Findsite is largely dependent on the completeness of its template library.

- The geometry-based ConCavity, energy-based Q-SiteFinder, and consensus-based MetaPocket predictors benefit from the usage of the holo structure and generate less accurate predictions when using the apo structures.

- Predictive quality of Findsite, ConCavity, Q-SiteFinder, and MetaPocket are strongly positively correlated with size of binding sites.

- The predictive performance of same of the top-performing methods substantially differs for different organic compound types. Considering the four major organic compound groups, Findsite and ConCavity achieve the highest success rates for the cofactors, followed by the mononucleotides and acids, and lower accuracies for the carbohydrates. In contrast, the differences between the success rates for different organic compound groups for the Q-SiteFinder and MetaPocket are relatively minor.

- The predictions by different binding site predictors are complementary and a simple consensus-based approach improves the success rates of the best-performing Findsite.

Prediction of the nucleotide-binding residues from protein sequence

- The predictions by machine-learning based approach and alignment-based method are complementary. A consensus of these two approaches improves the predictive quality.

- The proposed NsitePred method is significantly more accurate in identification of nucleotide-binding residues when compared with the existing ATPint and GTPbinder methods, as well as solutions based on the sequence alignment and residue conservation scoring.

- The NsitePred method predicts fewer nucleotide-binding residues for the non-binding chains than for the nucleotide-binding chains, and outperforms the ATPint and GTPbinder methods on this aspect.

- Predictive quality of NsitePred is strongly negatively correlated with the spread value of the binding sites. NsitePred performs particularly well for the binding sites in which the binding residues are clustered together in the sequence.

- The nucleotide-binding residues are associated with certain sequence-derived characteristics, including specific arrangements of secondary structures, dihedral angles, and certain amino acid pairs in the specific neighboring positions in the sequence.

Prediction of the nucleotide-binding sites for protein structure

- The proposed NSiteMatch predictor is superior to the template-base Findsite for all ligand types and at all filter levels. This supports the utility of our novel predictive approach that is based on local similarity in the binding region.

- The proposed NSiteMatch is significantly more accurate than the template-free binding site predictors Q-SiteFinder and MetaPocket, except for the

AMP nucleotide, for which NSiteMatch generates comparable results to the Q-SiteFinder and MetaPocket at the superfamily and fold filter levels.

- The template-based NSiteMatch and Findsite generate better predictions for proteins that share higher structural similarity to their template library.

- The NSiteMatch method is capable of identifying templates that have dissimilar overall topology and similar binding area when compared with the predicted protein, i.e., our method is capable of finding distant functional relationships.

8.2 Limitations and future directions

The field of protein-small ligand interactions covers a large number of research subjects and the studies performed in this dissertation concern only a few selected topics. We acknowledge that several other topics deserve research attention, for instance, prediction of the protein-anion and the protein-inorganic cluster interactions, and building of a comprehensive structure-based consensus method that utilizes the predictions from the existing structure-based predictors. Below we discuss the limitations and future directions that are specifically related to the work presented in this dissertation.

Investigation of the atomic level patterns in protein-small ligand interactions

The pattern for protein-organic compound interactions, i.e., hydrogen bonds established between NH- group (as the donor) of an AA and the oxygen atom of an organic compound, is relatively generic. Since PDB contains more than 10,000 organic compounds, it is unlikely to discovery a pattern that covers all protein-organic compound interactions. An alternative solution is to divide the organic compounds into a number of subtypes and summarize the interaction patterns for each of these types. For instance, the structural and sequence motifs were previously proposed for the protein-ATP interactions (Walker et al., 1982).

Prediction of the nucleotide-binding residues from protein sequence

The limitation of the proposed NsitePred is that this method is designed specifically for prediction of the nucleotide-binding residues and cannot be used for prediction of binding residues of other organic compounds. Our method needs to be rebuilt to predict the other compounds. Moreover, NsitePred provides less accurate predictions for the binding sites with larger spread values. These two weaknesses will be addressed in our future work.

Prediction of the nucleotide-binding sites for protein structure

Although the proposed NSiteMatch method provides more accurate predictions than the template-free methods, i.e., MetaPocket and Q-SiteFinder, the NSiteMatch method is more computationally expensive than the template-free methods. We are planning to modify the grid scanning steps to reduce the amount of computations. Since NSiteMatch is a template-based method, its predictive quality is largely dependent on the structure similarity between the predicted protein and the template library. In the case that the template library contains a small number of (dissimilar) binding sites, the NSiteMatch would generate less accurate predictions. Our future work will address this shortcoming by iteratively enlarging the library, as more native complexes become available.

Bibliography

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*., 25: 3389-402.

Andreeva, A., Howorth, D., Chandonia, J.M., Brenner, S.E., Hubbard, T.P., et al. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res*. 36: D419-D425.

Angkawidjaja, C., You, D.J., Matsumura, H., Kuwahara, K., Koga, Y., et al. (2007) Crystal structure of a family I.3 lipase from Pseudomonas sp. MIS38 in a closed conformation. *FEBS Lett.* 581:5060–4.

Ashburner, M., Ball, CA., Blake, J.A., Botstein, D., and Butler, H. (2000) Gene ontology: tool for the unification of biology. *Nat Genet*. 25:25-9.

Ashkenazy, H., Erez, E., Martz, E., Pupko, T., and Ben-Tal, N. (2010) ConSurf 2010: calculating evolutionary conservation in sequence and structure of proteins and nucleic acids. *Nucleic Acids Res*., 38:W529-33.

Bárány, M., Barron, J.T., Gu, L., and Bárány, K. (2001) Exchange of the actin-bound nucleotide in intact arterial smooth muscle. *J. Biol. Chem*. 276:48398–403.

Barrell, D., Dimmer, E., Huntley, R.P., Binns, D., O'Donovan, C., et al. (2009) The GOA database in 2009--an integrated Gene Ontology Annotation resource. *Nucleic Acids Res*. 37:D396-D403.

Benning, M.M., Taylor, K.L., Liu, R.Q., Yang, G., Xiang, H., et al. (1996) Structure of 4-chlorobenzoyl coenzyme A dehalogenase determined to 1.8 A resolution: an enzyme catalyst generated via adaptive mutation. *Biochemistry*, 35:8103–9.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., et al. (2000) The Protein Data Bank. *Nucleic Acids Res*. 28:235-242.

Brady, G., and Stouten, P. (2000) Fast prediction and visualization of protein binding pockets with PASS. *J Comput Aided Mol Des*. 14:383-401.

Brooijmans, N., and Kuntz, I.D. (2003) Molecular recognition and docking algorithms. *Annu Rev Biophys Biomol Struct*. 32:335-73.

Capra, J.A., and Singh, M. (2007) Predicting functionally important residues from sequence conservation. *Bioinformatics*, 23:1875-82.

Capra, J.A., Laskowski, R.A., Thornton, J.M., Singh, M., and Funkhouser, T.A. (2009) Predicting protein ligand binding sites by combining evolutionary sequence conservation and 3D structure. *PLos Comput Biol*. 5, 12.

Chauhan, J.S., Mishra, N.K. and Raghava, G.P. (2009) Identification of ATP binding residues of a protein from its primary sequence. *BMC Bioinformatics*, 10, 434.

Chauhan, J.S., Mishra, N.K. and Raghava, G.P. (2010) Prediction of GTP interacting residues, dipeptides and tripeptides in a protein from its evolutionary information. *BMC Bioinformatics*, 11, 301.

Chen, K., and Kurgan, L. (2009) Investigation of atomic level patterns in protein-small ligand interactions. *PLoS ONE*, 4, e4473.

Chen, K., Mizianty, M., Gao, J., and Kurgan, L.A. (2011) A critical comparative assessment of predictions of protein binding sites for biologically relevant organic compounds. *Structure*, 19:613-621.

Chou, W.Y., Chou, W.I., Pai, T.W., Lin, S.C., Jiang, T.Y., et al. (2010) Feature-incorporated alignment based ligand-binding residue prediction for carbohydrate-binding modules. *Bioinformatics*, 26:1022-8.

Cornell, W.D., Cieplak, P., Bayly, C.I., Gould, I.R., Merz, K.M.Jr., et al. (1995) A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules. *J. Am. Chem. Soc*. 117:5179–5197.

Cortes, C., and Vapnik, V. (1995) Support-Vector Networks. *Machine Learning*, 20:273-97.

Cuff, A.L., Sillitoe, I., Lewis, T., Redfern, O.C., Garratt, R., et al. (2009) The CATH classification revisited--architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Res*. 37:D310-4.

Davis, A.M., Teague, S.J., and Kleywegt, G.J. (2003) Application and limitations of X-ray crystallographic data in structure-based ligand and drug design. *Angew Chem Int Ed Engl*. 42: 2718–36.

Declercq, J.P., Evrard, C., Lamzin, V., and Parello, J. (1999) Crystal structure of the EF-hand parvalbumin at atomic resolution (0.91 A) and at low temperature (100 K). Evidence for conformational multistates within the hydrophobic core. *Protein Sci*. 8:2194–2204.

Denessiouk, K.A. and Johnson MS. (2000) When fold is not important: a common structural framework for adenine and AMP binding in 12 unrelated protein families. *Proteins*. 38:310-26.

Dessailly, B.H., Lensink, M.F., Orengo, C.A., and Wodak, S.J. (2008) LigASite--a database of biologically relevant binding sites in proteins with known apo-structures. *Nucleic Acids Res*. 36: D667-73.

Dudev, T., and Lim, C. (2008a) Metal binding affinity and selectivity in metalloproteins: insights from computational studies. *Annu Rev Biophys.* 37:97–116.

Dudev, T., and Lim, C. (2008b) Principles governing Mg, Ca, and Zn binding and selectivity in proteins. *Chem Rev*. 103:773–88.

Ellis, J.J., Broom, M., and Jones, S. (2007). Protein-RNA interactions: structural analysis and functional classes. *Proteins*. 66:903–11.

Fan, R.E., Chang, K.W., Hsieh, C.J., Wang X.R, and Lin, C.J. (2008) LIBLINEAR: A library for large linear classification. *J Mach Learn Res*. 9:1871-4.

Fan, R.E., Chen, P.H., and Lin, C.J. (2005) Working set selection using second order information for training SVM. *J Mach Learn Res*. 6:1889-1918.

Faraggi, E., Xue, B., and Zhou, Y. (2009) Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network. *Proteins*. 74:847-56.

Fields, R.D., and Burnstock, G. (2006) Purinergic signalling in neuron-glia interactions. *Nat Rev Neurosci*. 7:423–36.

Fiorucci, S., and Zacharias, M. (2010) Prediction of protein-protein interaction sites using electrostatic desolvation profiles. *Biophys J*. 98:1921-30.

Fredholm, B.B., Abbracchio, M.P., Burnstock, G., Daly, J.W., Harden, T.K., et al. (1994) Nomenclature and classification of purinoceptors. *Pharmacol Rev.* 46:143–156.

Gao, M., and Skolnick, J. (2008) DBD-Hunter: a knowledge-based method for the prediction of DNA-protein interactions. *Nucleic Acids Res*. 36:3978-92.

Gifford, J.L., Walsh, M.P., and Vogel, H.J. (2007) Structures and metal-ion-binding properties of the Ca2+-binding heli.-loop-helix EF-hand motifs. *Biochem J*. 405:199–221.

Gilman, A.G. (1987) G Proteins: Transducers of Receptor-Generated Signals. *Annu Rev Biochem*. 56:615–649

Goldenberg, O., Erez, E., Nimrod, G., and Ben-Tal, N. (2009) The ConSurf-DB: pre-calculated evolutionary conservation profiles of protein structures. *Nucleic Acids Res*. 37:D323-7.

Gómez-García, I., Oyenarte, I., and Martínez-Cruz, L.A. (2010) The crystal structure of protein MJ1225 from Methanocaldococcus jannaschii shows strong conservation of key structural features seen in the eukaryal gamma-AMPK. *J Mol Biol*. 399:53-70.

Goto, S., Okuno, Y., Hattori, M., Nishioka, T., and Kanehisa, M. (2002) LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res*. 30:402-4.

Grishkovskaya, I., Avvakumov, G.V., Sklenar, G., Dales, D., Hammond, G.L., et al. (2000) Crystal structure of human sex hormone-binding globulin: steroid transport by a laminin G-like domain. *EMBO J*. 19:504–12.

Gutteridge, A., and Thornton, J.M. (2005) Understanding nature's catalytic toolkit. *Trends in Biochem Sci*. 30:622–29.

Hendlich, M., Rippmann, F., and Barnickel G. (1997) LIGSITE: automatic and efficient detection of potential small molecule-binding sites in proteins. *J Mol Graph Model*. 15:359-63.

Horst, J.A., and Samudrala, R. (2010) A protein sequence meta-functional signature for calcium binding residue prediction. *Pattern Recognit Lett*. 31:2103-2112.

Howard, J., and Hyman, A.A. (2007) Microtubule polymerases and depolymerases. *Curr Opin Cell Biol*. 19:31-5.

Huang, B. (2009) MetaPocket: a meta approach to improve protein ligand binding site prediction. *OMICS.* 13:325-30.

Huang, B., and Schroeder, M. (2006) LIGSITEcsc: predicting ligand binding sites using the Connolly surface and degree of conservation. *BMC Struct Biol*. 6, 19.

Jones, S., and Thornton, J.M. (1996) Principles of protein-protein interactions. *Proc Natl Acad Sci U S A.* 93:13–20.

Kabsch, W., and Sander, C. (1983) Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers*, 22:2577-637.

Kotnik, M., Humljan, J., Contreras-Martel, C., Oblak, M., Kristan, K., et al. (2007) Structural and functional characterization of enantiomeric glutamic acid derivatives as potential transition state analogue inhibitors of MurD ligase. *J Mol Biol*. 370:107-15.

Kurgan, L., Homaeian, L. (2006) Prediction of Structural Classes for Protein Sequences and Domains - Impact of Prediction Algorithms, Sequence Representation and Homology, and Test Procedures on Accuracy. *Pattern Recognition*, 39:2323-2343.

Kyte, J., and Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein. *J Mol Biol.* 157:105–132.

Lan, F., Collins, R.E., De Cegli, R., Alpatov, R., Horton, J.R., et al. (2007) Recognition of unmethylated histone H3 lysine 4 links BHC80 to LSD1-mediated gene repression. *Nature*, 448:718–22.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., et al. (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*, 23:2947-2948.

Laskowski, R. (1995) SURFNET: a program for visualizing molecular surfaces, cavities and intermolecular interactions. *J Mol Graph*. 13:323-330.

Laurie, A.T., and Jackson, R.M. (2005) Q-SiteFinder: an energy-based method for the prediction of protein-ligand binding sites. *Bioinformatics*, 21:1908-16.

Le, Guilloux V., Schmidtke, P., and Tuffery, P. (2009) Fpocket: an open source platform for ligand pocket detection. *BMC Bioinformatics*, 10,168.

Levitt, M. (2007) Growth of novel protein structural data. *Proc Natl Acad Sci U S A*. 104:3183-8.

Li, W., and Godzik, A. (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22:1658-1659.

Linderstrøm-Lang, K.U. (1952) *Proteins and Enzymes*. Stanford University Press.

Liu, D., Thomas, P.W., Momb, J., Hoang, Q.Q., Petsko, G.A., et al. (2007) Structure and specificity of a quorum-quenching lactonase (AiiB) from Agrobacterium tumefaciens. *Biochemistry*, 46:11789–99.

Luscombe, N.M., Laskowski, R.A., and Thornton, J.M. (2001) Amino acid-base interactions: a three-dimensional analysis of protein-DNA interactions at an atomic level. *Nucleic Acids Res*. 29:2860-74.

Ma, B., Elkayam, T., Wolfson, H., and Nussinov, R. (2003) Protein-protein interactions: structurally conserved residues distinguish between binding sites and exposed protein surfaces. *Proc Natl Acad Sci U S A.* 100:5772–7.

Ma, Q., Zhao, X., Nasser, E.A., Geerlof, A., Li, X., et al. (2006) The Mycobacterium tuberculosis LipB enzyme functions as a cysteine/lysine dyad acyltransferase. *Proc Natl Acad Sci U S A*. 103:8662–7.

Maret, W. (2005) Zinc coordination environments in proteins determine zinc functions. *J Trace Elem Med Biol.* 19:7–12.

Maton, A., Jean, H., Charles, W., Susan, J., Maryanna, Q.W., et al. (1993) *Human Biology and Health*. Englewood Cliffs, New Jersey, USA: Prentice Hall.

McDonald, I.K., and Thornton, J.M. (1994) Satisfying hydrogen bonding potential in proteins. *J Mol Biol.* 238:777–93.

McGuffin, L.J., Bryson, K., and Jones, D.T. (2000) The PSIPRED protein structure prediction server. *Bioinformatics*, 16:404-5.

Meng, X.Y., Zhang, H.X., Mezei, M., Cui, M. (2011) Molecular docking: a powerful approach for structure-based drug discovery. *Curr Comput Aided Drug Des*. 7:146-57.

Mitchison, T., and Kirschner, M. (1984) Dynamic instability of microtubule growth. *Nature*, 312:237–42.

Moodie, S.L., Mitchell, J.B. and Thornton, J.M. (1996) Protein recognition of adenylate: an example of a fuzzy recognition template. *J Mol Biol*. 263:486-500.

Morris, G. M., Goodsell, D. S., Halliday, R.S., Huey, R., Hart, W. E., et al. (1998) Automated Docking Using a Lamarckian Genetic Algorithm and and Empirical Binding Free Energy Function. *J.Comput.Chem*. 19:1639-1662.

Mukherjee, S., Acharya, B.R., Bhattacharyya, B., and Chakrabarti, G. (2010) Genistein arrests cell cycle progression of A549 cells at the G(2)/M phase and depolymerizes interphase microtubules through binding to a unique site of tubulin. *Biochemistry*, 49:1702-12.

Murakami, Y., Spriggs, R.V., Nakamura, H., and Jones, S. (2010) PiRaNhA: a server for the computational prediction of RNA-binding residues in protein sequences. *Nucleic Acids Res*. 38:W412-6.

Pandit, S.B., and Skolnick, J. (2008) Fr-TM-align: a new protein structural alignment method based on fragment alignments and the TM-score. *BMC Bioinformatics*, 9, 531.

Pei, J., Grishin, N.V. (2001) AL2CO: calculation of positional conservation in a protein sequence alignment. *Bioinformatics*, 17:700-12.

Petrey, D., Fischer, M., and Honig, B. (2009) Structural relationships among proteins with different global topologies and their implications for function annotation strategies. *Proc Natl Acad Sci U S A*. 106:17377-82.

Popova, Y., Thayumanavan, P., Lonati, E., Agrochão, M., and Thevelein, J.M. (2010) Transport and signaling through the phosphate-binding site of the yeast Pho84 phosphate transceptor. *Proc Natl Acad Sci U S A*. 107:2890-5.

Pruitt, K.D., Tatusova, T., Klimke, W., Maglott, D.R. (2009) NCBI Reference Sequences: current status, policy, and new initiatives. *Nucleic Acids Res*. 37:D32-6.

Pupko, T., Bell, R.E., Mayrose, I., Glaser, F., and Ben-Tal, N. (2002) Rate4Site: an algorithmic tool for the identification of functional regions in proteins by surface mapping of evolutionary determinants within their homologues. *Bioinformatics*, Suppl 1:S71-7.

Que, E.L., Domaille, D.W., and Chang, C.J. (2008) Metals in neurobiology: probing their chemistry and biology with molecular imaging. *Chem Rev*. 108:1517-49.

Rajamani, D., Thiel, S., Vajda, S., and Camacho, C.J. (2004) Anchor residues in protein-protein interactions. Proc Natl Acad Sci U S A 101: 11287–92.

Rich, P.R. (2003) The molecular machinery of Keilin's respiratory chain. *Biochem. Soc. Trans*. 31:1095–105.

Rosenberg, M.R., and Casarotto, M.G. (2010) Coexistence of two adamantane binding sites in the influenza A M2 ion channel. *Proc Natl Acad Sci U S A*. 107:13866-71.

Sakai, N., Yao, M., Itou, H., Watanabe, N., Yumoto, F., et al. (2001) The three-dimensional structure of septum site-determining protein MinD from Pyrococcus horikoshii OT3 in complex with Mg-ADP. *Structure*, 9:817-26.

Saraste, M., Sibbald, P.R., and Wittinghofer, A. (1990) The P-loop–A common motif in ATP-binding and GTP-binding proteins. *Trends Biochem Sci*. 15:430-434.

Schnell, J.R., and Chou, J.J. (2008) Structure and mechanism of the M2 proton channel of influenza A virus. *Nature*, 451:591-5.

Senes, A., Gerstein, M., and Engelman, D.M. (2000) Statistical analysis of amino acid patterns in transmembrane helices: the GxxxG motif occurs frequently and in association with beta-branched residues at neighboring positions. *J Mol Biol*. 296:921-36.

Shapiro, S.S., Wilk, M.B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, 52:591–611.

Shin, J.M., and Cho, D.H. (2005) PDB-Ligand: a ligand database based on PDB for the automated and customized classification of ligand-binding structures. *Nucleic Acids Res*. 33:D238-41.

Shomura, Y., Yoshida, T., Iizuka, R., Maruyama, T., Yohda, M., et al. (2004) Crystal structures of the group II chaperonin from Thermococcus strain KS-1: steric hindrance by the substituted amino acid, and inter-subunit rearrangement between two crystal forms. *J Mol Biol*. 335:1265-78.

Silva, J. J., and Williams, R.J. (1991) *The Biological Chemistry of the Elements*, Clarendon Press: Oxford.

Skolnick, J., and Brylinski, M. (2008) A threading-based method (FINDSITE) for ligand-binding site prediction and functional annotation. *Proc Natl Acad Sci U S A*. 105:129-34.

Smith, B.J., Colman, P.M., Von, I.M., Danylec, B., and Varghese, J.N. (2001) Analysis of inhibitor binding in influenza virus neuraminidase. *Protein Sci*. 10:689–96.

Stouffer, A.L., Acharya, R., Salom, D., Levine, A.S., Di, Costanzo L., et al. (2008) Structural basis for the function and inhibition of an influenza virus proton channel. *Nature*, 451:596-9.

Sudom, A.M., Prasad, L., Goldie, H., and Delbaere, L.T. (2001) The phosphoryl-transfer mechanism of Escherichia coli phosphoenolpyruvate carboxykinase from the use of AlF(3). *J Mol Biol*. 314:83-92.

Tate, R.F. (1954). Correlation between a discrete and a continuous variable. Point-biserial correlation. *Ann. Math. Statist*, 25:603-7.

Tegge, A.N., Wang, Z., Eickholt, J., and Cheng, J. (2009) NNcon: improved protein contact map prediction using 2D-recursive neural networks. *Nucleic Acids Res*. 37:W515-8.

UniProt Consortium. (2010) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res*. 38:D142-8.

Walker, J.E., Saraste, M., Runswick, M.J., and Gay, N.J. (1982) Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J*. 1:945-951.

Wang, G., and Dunbrack, R.L.Jr. (2003) PISCES: a protein sequence culling server. *Bioinformatics*, 19:1589–91.

Wang, K., and Samudrala, R. (2006) Incorporating background frequency improves entropy-based residue conservation measures. BMC Bioinformatics, 7, 385.

Weisel, M., Proschak, E., and Schneider, G. (2007) PocketPicker: analysis of ligand binding-sites with shape descriptors. *Chem Cent J*. 1, 7.

Whittard, J.D., Sakurai, T., Cassella, M.R., Gazdoiu, M., and Felsenfeld, D.P. (2006) MAP kinase pathway-dependent phosphorylation of the L1-CAM ankyrin binding site regulates neuronal growth. *Mol Biol Cell*. 17:2696-706.

Wikimedia, Foundation. (2006) Wikipedia: The Free Encyclopedia. http://en.wikipedia.org/.

Wilcoxon, F. (1945) Individual comparisons by ranking methods. *Biometrics Bulletin*, 1:80–83.

Wishart, D.S., Knox, C., Guo, A.C., Cheng, D., Shrivastava, S., et al. (2008) DrugBank: a knowledgebase for drugs, drug actions and drug targets. *Nucleic Acids Res*. 36:D901-6.

Word, J.M., Lovell, S.C., Richardson, J.S., and Richardson, D.C. (1999) Asparagine and glutamine: using hydrogen atom contacts in the choice of side-chain amide orientation. *J Mol Biol*. 285:1735–47.

Wu, X., Kumar, V., Quinlan, J.R., Ghosh, J., Yang, Q., et al. (2007) Top 10 algorithms in data mining. *Knowledge and Information Systems*. 14:1-37.

Zhang, J., Adrián, F.J., Jahnke, W., Cowan-Jacob, S.W., Li, A.G., et al. (2010) Targeting Bcr-Abl by combining allosteric with ATP-binding-site inhibitors. *Nature*, 463:501-6.

Zhang, T., Zhang, H., Chen, K., Shen, S., Ruan, J., et al. (2008) Accurate Sequence-based Prediction of Catalytic Residues. *Bioinformatics*, 24:2329-2338.

Zhang, Y., and Skolnick, J. (2005) TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res*. 33:2302-9.

Zhu, H., Sommer, I., Lengauer, T., and Domingues, F.S. (2008) Alignment of Non-Covalent Interactions at Protein-Protein Interfaces. *PLoS ONE*, 3:e1926.

Zoltowski, B.D., Schwerdtfeger, C., Widom, J., Loros, J.J., Bilwes, A.M., et al. (2007) Conformational switching in the fungal light sensor Vivid. *Science*, 316:1054–7.

APPENDIX A

Appendix A presents the list of protein chains used to design and evaluate the NsiteMatch predictor in Section 7.4. Each entry contains 5 characters, where the first 4 characters indicate the PDB id of the protein and the last letter indicates the chain id.

List of chains interact with ADP.

1A9XA	1HI5A	1O0HA	1UKYA	2AKOA	2FV7A	2NCDA	2V1XA	3B6VA	3FE2A
1AO0A	1HQYE	1O51A	1UM8A	2AWNA	2G2IA	2NO0A	2V2ZA	3BF1A	3FH0A
1AONA	1HTWA	1O6BA	1V1AA	2AXNA	2G5IB	2NR8A	2VASA	3BFNA	3FI8A
1B62A	1HUXA	1O92B	1VA6A	2B9FA	2GK6A	2NUNA	2VEDA	3BK7A	3FMPB
1B6SA	1HW8B	1OH9A	1VHLA	2BEJA	2GKSA	2O0JA	2VF7A	3BRBA	3FWRA
1BS1A	1I58B	1OL5A	1VTKA	2BFRA	2GL6A	2O1VA	2VOSA	3BXZA	3FWYA
1CNFA	1IAHA	1OSNA	1W44A	2BTDA	2GR0A	2OJWA	2W41A	3C4NA	3G15A
1CQIB	1II6A	1OXUA	1W5SA	2BUFB	2GRJA	2OLCA	2W4KA	3C4ZA	3G2FA
1CZAN	1IN4A	1P72A	1W78A	2BVCA	2GRYA	2OLJA	2W58A	3C7NA	3GLFB
1DJNA	1IONA	1PFKA	1W7IA	2C2AA	2GXAC	2ONMA	2W5AA	3C7NB	3GR4A
1E19A	1IOVA	1PKGA	1WBPA	2C31A	2H1FA	2OTGA	2W6ED	3C9UA	3GVIA
1E3MA	1IQPA	1Q3SA	1WNLA	2C9OA	2HENA	2OWMA	2W8RA	3CR3A	3H4SA
1E4EA	1J1CA	1Q8YA	1X3MA	2CDNA	2HGSA	2OXCA	2WQNA	3CWQA	3HB9A
1E79B	1J7LA	1R0YA	1X6VB	2CDUA	2HMFA	2P05A	2WW4A	3D36A	3HY6A
1E8HA	1JBPE	1R7RA	1XJKA	2CGJA	2HMVA	2PL3A	2X6TA	3D54A	3HYOA
1E9FA	1JEDA	1RFUA	1XMVA	2CH6A	2HV7B	2PO0A	2YWVA	3D5WA	3HZ6A
1EA6A	1JJ7A	1RK2A	1XRJA	2CN5A	2HYDA	2PYWA	2YX6B	3D8BA	3I0OA
1EHIA	1K3CA	1RZUA	1XTJA	2CNQA	2I5BA	2Q14A	2Z4RA	3DC4A	3I61A
1EQMA	1KSFX	1S4EA	1XW4X	2CVXA	2I8CA	2Q2RA	2ZAOA	3DINA	3I73A
1F48A	1L0OA	1SVLA	1XX6A	2D0OA	2IF8A	2QB5A	2ZBDA	3DKPA	3ICEA
1FNNA	1L8QA	1SXJE	1Y63A	2D2FA	2IO8A	2QBYB	2ZDGA	3DLSA	3ICSB
1FP6A	1LTQA	1T3TA	1Y8OA	2DCNA	2IOPA	2QENA	2ZGVA	3DSRA	3IG8A
1FWKA	1LVGA	1T5CA	1YP4A	2DHRA	2IUUA	2QQ0B	2ZJ5A	3DZVA	3IN1A
1G6HA	1M15A	1T6XA	1Z2NX	2DPYB	2IW3A	2QR1G	2ZPAA	3ECCA	3K5IA
1G6OA	1MWMA	1TF2A	1Z59A	2DR3A	2IYQA	2QSYA	2ZS8A	3EGIA	3KALA
1G8XA	1N06A	1TY8A	1Z6TA	2DWCA	2J0WA	2QV7A	2ZTSA	3EHHA	3KB1A
1G99A	1NKSF	1TZDA	1ZARA	2E2PA	2J9DB	2R6FA	3A0TA	3EPQA	3KJGA
1GC5A	1NKTA	1U0JA	1ZS6A	2ECKA	2JA3A	2R7NA	3A1DA	3EQGA	3KO3D
1GKIA	1NQTA	1U2VA	1ZTHA	2EWVA	2JCBA	2REPA	3A37A	3EX7C	3KQLA

1GKZA	1NVAA	1U2VB	1ZXNA	2F1JA	2JFGA	2RIOA	3A4MA	3EZ2A	3KX2B
1GLBG	1NY3A	1U3FA	2A2CA	2FNAA	2JGVB	2SHKB	3A7JA	3F61A	3L8KA
1GSAA	1NY5A	1UC9A	2AD5A	2FSNA	2JLSA	2V1UA	3B5ZA	3FD6A	3LV8A
1GZFC									

List of chains interact with ATP.

1A0IA	1G3IA	1M83A	1SU2A	1YFRA	2FAQA	2NVUB	2YWWA	3CRCA	3H1QA
1A49A	1G5TA	1MB9A	1SVMA	1YIDB	2FGHA	2NYJA	2Z02A	3D2EA	3H39A
1A82A	1GN8A	1MIWA	1TF7A	1YP3A	2FGJA	2O0HA	2Z08A	3DKCA	3H5NA
1ATPE	1GOLA	1MJHA	1TILA	1YUNA	2FSGA	2OGXA	2Z1UA	3DNTA	3H8VA
1AYLA	1H4QB	1MO8A	1TQPA	1Z0SA	2HIXA	2OH5A	2ZANA	3DWLA	3HAVA
1B0UA	1H8HF	1MV5A	1TWAB	1Z7EA	2HMUA	2P09A	2ZDQA	3E1YA	3HGMA
1B76A	1HI1A	1N5IA	1TYQB	1ZFNA	2HS0A	2PBZA	2ZHZA	3E7EA	3HMNA
1B8AA	1HP1A	1NGEA	1U5RA	1ZP9A	2I4OA	2Q0DA	2ZSFA	3EA0A	3HRCA
1BCPE	1II0A	1NSFA	1UA2A	1ZYDA	2IA6A	2Q7GA	2ZT7A	3EFSA	3I7VA
1BCPF	1J09A	1NYRA	1UF9C	2A5YB	2IAJA	2QB8A	3A8TA	3EHGA	3IBQA
1CSNA	1J1ZA	1O93A	1V1BA	2AQXA	2IDXA	2QK4A	3B2QA	3ETHA	3IE7A
1D9ZA	1J7KA	1OBDA	1VJCA	2ARUA	2IJMA	2QKMB	3BG5A	3F5MA	3IKHA
1DY3A	1JI0A	1OJLE	1WKLB	2B6FA	2ILYA	2QRDG	3BJUA	3FDXA	3IN5A
1E2QA	1JJVA	1OL6A	1X01A	2BEKA	2IVPA	2QUIA	3BLQA	3FKQA	3INNA
1E4GT	1JKNA	1PHKA	1XDNA	2C01X	2IXEA	2R7LA	3BU5A	3FPBA	3IQ0A
1E8XA	1KJ8A	1PJ4A	1XDPA	2CBZA	2IYWA	2R9VA	3C16A	3FVQA	3K5HA
1EE1A	1KMNA	1PK8A	1XEXA	2CG9A	2J3MA	2RD5C	3C16B	3G59A	3KMWA
1ESQA	1KO5A	1Q12A	1XMIA	2CJAA	2J9LA	2V92E	3C4WA	3G6VA	3LGXA
1F2UA	1KP2A	1Q97A	1XNGA	2DDOA	2JJXA	2VHQA	3C5EA	3GBUA	3LKIB
1F2UB	1KP8A	1QHGA	1XSCA	2E5YA	2JK8A	2VT3A	3C9RA	3GNIB	3LSSA
1F9AA	1KVKA	1QHXA	1Y56A	2E89A	2KMXA	2W00A	3CISA	3GQNA	3R1RA
1FMWA	1L2TA	1R8BA	1Y8PA	2EWWA	2NPIA	2W02A	3CQDA	3H0RE	4AT1B
1G21E	1LHRA	1S9IA	1Y8QB	2F02A	2NT8A	2W5GA			

List of chains interact with AMP.

12ASA	1GPMA	1NH8A	1TUUB	1ZBUA	2F3DA	2OUNA	2W4YA	3DDJA	3GC0A
1AMUA	1H3DA	1NKSC	1U9ZA	2A7XA	2FJBA	2PZAA	2WEFA	3DHVA	3GLVB
1ANKA	1HTOA	1O94C	1UA4A	2AK3A	2G1UA	2QGAB	2YRXA	3DLZA	3H7EA
1CJAA	1HTTA	1OBDA	1UKZA	2ARTA	2GM3A	2QJTB	2YVOA	3DQXA	3HF7A
1CT9A	1J20A	1OBTA	1UUYA	2BWJC	2GMKA	2QRCG	2ZE5A	3DRWA	3HW5A
1DELB	1JWBB	1OREA	1UXNA	2C38A	2GSUA	2QRKA	2ZR2A	3ERRA	3I0QA

1DGSA	1K9YA	1P8LA	1V26A	2C38B	2HBLA	2R7MA	3A7AA	3FEGA	3IB8A
1DMAB	1KHTB	1QB8A	1V8SA	2C5SA	2HCRA	2RIFA	3B6JB	3FHJA	3IUYA
1DS5A	1KPFA	1RAOA	1VD1A	2CFMA	2I4IA	2V8QE	3BERA	3FHMA	3JWPA
1ECJA	1KTGA	1RY2A	1XQSC	2D1QA	2IVTA	2VARA	3BLWA	3FIUA	3KD6A
1EFVB	1LTKA	1S68A	1Y1PA	2DCLA	2J91A	2VFKA	3C0HA	3FNAA	3LFRA
1FA9A	1MC1A	1T6YA	1YXUA	2DSDA	2J9DE	2VIIA	3C85C	3FWZA	3LHHA
1G51B	1MD9A	1TB5A	1Z6SA	2EQAA	2JB7A	2VSOA	3CJ7A	3G1ZA	3LOQA
1GPH1	1MF0A	1TBWA	1Z84A	2F17A	2JGDA	2VZEA	3CW9A	3G89B	3LW7B