COMPUTATIONAL METHODS FOR PREDICTINGPROTEIN-NUCLEIC ACIDS INTERACTION

A Dissertation
Submitted to the Graduate Faculty
of the
North Dakota State University
of Agriculture and Applied Science

By

Wen Cheng

In Partial Fulfillment of the Requirements
for the Degree of
DOCTOR OF PHILOSOPHY

Major Department:
Computer Science

April 2015

Fargo, North Dakota

North Dakota State University
Graduate School

Title

COMPUTATIONAL METHODS FOR PREDICTING PROTEIN-NUCLEIC ACIDS INTERACTION

By

Wen Cheng

The Supervisory Committee certifies that this ***disquisition*** complies with North Dakota State University's regulations and meets the accepted standards for the degree of

DOCTOR OF PHILOSOPHY

SUPERVISORY COMMITTEE:

Changhui Yan
Chair

Kendall Nygard

Kenneth Magel

Jun Kong, and Jing Shi

Approved:

July 9, 2015
Date

Brian M. Slator
Department Chair

ABSTRACT

Since the inception of various proteomic projects, protein structures with unknown functions have been discovered at a fast speed. The proteins regulate many important biological processes by interacting with nucleic acids that include DNA and RNA. Traditional wet-lab methods for protein function discovery are too slow to handle this rapid increase of data. Therefore, there is a need for computational methods that can predict the interaction between proteins and nucleic acids. There are two related problems when predicting protein-nucleic interactions. One problem is to identify nucleic acid-binding sites on the protein structures, and the other problem is to predict the 3-D structure of the complex that protein and nucleic acids form during interaction. The second problem can be further divided into two steps. The first step is to generate potential structures for the protein-nucleic acids complex. The second step is to assign scores to the poses generated in the first step.

This dissertation presents two computational methods that we developed to predict the protein-nucleic acids interaction. The first method is a scoring function that can discriminate native structures of protein-DNA complexes from non-native poses, which are also known as docking decoys. We analyze the distribution of protein atoms around each structural component of the DNA and develop spatial-specific scoring matrices (SSSMs) based on the observed distribution. We show that the SSSMs could be used as a knowledge-based energy function to discriminate native protein-DNA structures and various decoys.

Our second method discovers the graphs that are enriched on the protein-nucleic acids interfaces and then uses the sub-graphs to predict RNA-binding sites on protein structures and to assign scores to protein-RNA poses. First, the interface area of each RNA-binding protein is represented as a graph, where each node represents an interface residue. Then, common sub-

graphs being abundant in these graphs are identified. The method is able to identify RNA-binding sites on the protein surface with high accuracy. We also demonstrate that the common sub-graphs can be used as a scoring function to rank the protein-RNA poses. Our method is simple in computation, while its results are easier to interpret in biological contexts.

ACKNOWLEDGEMENTS

I am so honored and grateful to take this opportunity to thank Dr. Changhui Yan, Dr. Kendall Nygard, Dr. Kenneth Magel, Dr. Jun Kong, and Dr. Jing Shi for their willingness to serve on my Advisory Committee. Particularly, I want to show my deep gratitude to Dr. Changhui Yan, my Ph.D. adviser, for his dedication; patience; and endless time and effort while guiding and enlightening, supporting, and encouraging me during my Ph.D. program. I learned numerous things from his huge passion for doing research; his advice that is full of wisdom, leading me through tough time; his rigorous attitude in academic work; and his generosity and kindness. I could not be where I am right now without his help. I feel so lucky to have such a great adviser.

I want to thank Dr. Kendall Nygard for serving on my Advisory Committee. I always enjoy discussing many aspects of my research projects with him and listening to his helpful advice. I look forward to further discussions.

I especially want to thank Dr. Kenneth Magel, Dr. Jun Kong, and Dr. Jing Shi for serving on my Ph.D. Advisory Committee. I appreciate the comments and help that they provided as well as their willingness to serve on my committee shortly after my notification.

Finally, I want to thank my family members for their support, understanding, and encouragement during the hard time while pursuing my Ph.D.

DEDICATION

To my mum and grandma

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1. INTRODUCTION

Nucleic acids and protein are two essential biological molecules for every organism, where nucleic acids carry and transmit the gene information, and the protein is responsible for most cellular activities, such as catalyzing metabolic reactions, replicating DNA, and moving molecules. Protein is a sequence of units that are called amino acids. There are 20 types of amino acids, which are also referred to as residues when they exist within the polymeric chain of a protein. The protein sequence usually folds in a unique 3-D structure to stay stable. There are four aspects of a protein structure: primary structure, secondary structure, tertiary structure, and quaternary structure. The primary structure is linear sequence of amino acids. The secondary structure is the local spatial arrangement of amino acids and has periodic structural patterns: α-helix, β-sheet, and loop. The tertiary structure is a 3-D fold of one or multiple secondary structure elements from a single protein chain. The quaternary structure is the 3-D structure of several protein chains and describes how those chains join together to form a complex.

Since the end of 1980, thanks to various genome projects, the amount of genetic sequence data increased dramatically. There were more than 67 million sequences stored in Genbank by February 2008, and the sequences continued to accumulate exponentially. The genetic sequences resulted in a tremendous number of protein sequences as well because protein can be directly translated from DNA. A great deal of valuable structural and functional information about the protein was hidden behind the enormous quantity of data.

Being able to determine the proteins' structures can help to discover these proteins' cellular and evolutionary roles and also to develop drugs to bind to these proteins. The protein structures are mainly solved by X-ray crystallography and Nuclear Magnetic Resonance (NMR). However, these techniques are extremely time-consuming and labor-expensive, which made the

increasing number of protein structures fall far behind the exponential growth of protein sequential data. For instance, only 45,000 protein structures were solved by February 2008, and the number slowly increased to around 100,000 in the protein data bank by 2015. Therefore, it has been imperative to develop computational methods to predict protein structures automatically, efficiently, and accurately as an alternative.

In every species, protein is responsible for almost all the tasks in a cell, for instance, catalyzing glycolysis, fostering Kreb's cycle, defensing against germs, synthesizing ATP, transporting small molecules within cells, etc. Protein interacts with molecules, such as other protein, DNA, RibonucleicAcid (RNA), viruses, small molecules, and ligands, to fulfil its functions. Knowing about the interaction of protein and those molecules can significantly help to understand molecular mechanisms and to recognize potential drug targets. Besides the experimental methods used to solve protein structures, mutagenesis is also applied to determine the interactions. A person who wants to understand a protein's function not only needs its sequential information, but also its structural information. The combined factors make it even more urgent and challenging to develop powerful computational methods to predict the protein functions.

Among the interactions between protein and molecules, protein-nucleic acid interaction is critical to a variety of biological processes, such as regulating transcription, translation, DNA replication, repair and recombination, RNA processing and translocation, enzymatic events, and operating nucleic acids as substrates. Protein-nucleic acid complexes have a significant impact on the structure and function of the associated nucleic acid. There are two types of nucleic acids, DNA and RNA. Thus, there are two types of protein-nucleic acids interactions: protein-DNA and protein-RNA. Accurate information about the interactions between protein and DNA can give

insight about the regulation of gene expression. Protein-RNA interactions are involved with biological processes, ranging from messenger RNA (mRNA) processing and protein synthesis, to RNA transport and RNA splicing, to viral replication and cellular defense against pathogens. It is further observed that the RNA-protein interaction significantly impacts cellular defense and developmental regulation (Hall, 2002; Tian et al., 2004). Therefore, knowing how protein and nucleic acids recognize and interact with each other is crucial for understanding those biological processes. With respect to DNA-binding proteins, we can roughly classify them into two groups: specific binding and non-specific binding. With specific binding, a protein seeks to bind a specific sequence of DNA while, for non-specific binding, a protein can bind to a set of DNA sequences. This dissertation focuses on non-specific binding.

In addition, compared with protein-RNA interaction, protein-DNA interaction has received longer and more effort and investigation by researchers, mainly due to the diversity of RNA structures as well as the lack of structural information about RNA sequences. Different from DNA, which is known for its double-helix structure, RNA can be hairpins/stem-loops, bulges, and loops, making RNA's interactions with protein more complicated and unpredictable.

1.1. Motivation and Problem Statement

Given the structures of proteins and nucleic acids, the problem of predicting how protein and nucleic acids interact includes two tasks, binding-site identification and complex structure prediction. The latter task has two steps: generating potential poses of protein and nucleic acid, and then scoring each of the poses. A huge volume of DNA and proteins was discovered with their structural and coordinate information thanks to a series of genomic and proteomic projects. This valuable resource is important for studying the interactions between protein and nucleic acids. However, the traditional experimental methods to implement this task are far behind the

needs for methods to process the increasing amount of biological data. Thus, it is vital to develop computational algorithms in order to automatically and efficiently analyze the known protein-nucleic acid complexes and to predict the interactions between proteins, which have known sequential and structural information along with unknown functions and functional sites, and nucleic acids.

There are two problems involved with predicting protein-nucleic acid interaction. One problem is to predict the protein interface or binding site, and the other problem is to predict the protein's interaction with nucleic acids or the 3-D structure of the protein-nucleic acids complex. A protein-nucleic acids complex assembly is the aggregation, arrangement, and bonding together of protein and nucleic acids. With respect to the first problem, the input data are a protein sequence with known sequential and structural information but unknown binding-site information. The output is supposed to be an annotation of the protein residues that interact with the nucleic acids. Regarding the second problem, the input is a protein sequence and nucleic acids pair with sequential and structural information available for both protein and nucleic acids. We know that they will bind together but do not know how they will interact with each other, i.e., their relative geometric position to each other. Thus, the problem is to specify the 3-D structure of a protein-nucleic acids complex on an atom-atom level.

The existing computational algorithms usually address only one of the aforementioned problems, and most methods that address the second problem are complex and computationally expensive. The rest of the chapter describes the existing computational methods for coping with the two problems, followed by a description of this work's contribution. Section 1.6 demonstrates the structure for dissertation.

1.2. Binding-Site Prediction

We can roughly classify the methods used to identify the protein's binding sites into two categories based on the ways that they handle the prediction: feature- and template-based. Template-based methods are centered on the hypothesis that similar protein structures lead to comparable functions. In this approach, the protein structure for which one wants to predict binding sites is referred to as the query or target protein. In order to predict binding sites on the query, this approach needs another protein structure called a template. The requirements for the template are that it is very similar to the query and that the template's binding sites are known. The query is structurally aligned with the template. Then, the region on the query that aligns with the template's binding sites is predicted to be the binding site on the query protein. Some representative methods include DNA-binding Domain Hunter (DBD-Hunter) (Gao and Skolnick, 2008), a structure-based method that incorporates the statistical energy function (Zhao et al., 2011), and RBRDetector (Yang et al., 2014), which combines the feature-and template-based strategies. The benefit of using the template-based methods is that they can accurately identify the binding sites if the algorithm can find a template in the database that is structurally similar to the target protein. However, if the correct template is absent, the approach can hardly identify reliable binding sites on the query protein.

Compared to template-based methods, feature-based methods, which commonly take advantage of machine-learning techniques to do the classification, have received longer and more attention. One important question for understanding the protein-nucleic acid interaction is "why the binding event occurs here but not there on the protein?" Many studies have found that nucleic acids bound the protein by recognizing specific sequential or structural patterns of the binding sites on the protein. For instance, it was observed that positively charged amino acids,

e.g. arginine, prefer to appear at nucleic-acid binding sites (Gallet et al., 2000). It was also observed that protein residues that are in contact with the DNA are better conserved than the remaining residues on the protein surface (Luscombe and Thornton, 2002). Jones et al (2003) found that DNA-binding sites on the protein surface are much more prone to having positive electrostatic potentials compared to non-binding sites. Therefore, it was natural to utilize those binding-site properties as features to predict the protein interface.

Feature-based approaches extract a variety of features from the annotated protein's nucleic-acids binding sites to describe and characterize those patterns. The features are fed to machine-learning classifiers, such as neural networks, Naïve Bayes, Support Vector Machine (SVM), and random forest, allowing the classifiers to learn the difference between binding sites and non-binding sites. Usually a cross-validation procedure is followed to test the classifiers' performance. During the prediction process, when given a protein with an unknown binding site, the features are encoded first, and then, the features are used to predict the binding site.

The most-used machine-learning classifiers are SVM, including BindN (Wang and Brown, 2006), BindN+ (Wang et al., 2010), PPRInt (Kumar et al., 2008), PiRaNhA (Murakami et al., 2010), PRINTR (Wang et al., 2008), RBRDetector (Yang et al., 2014), and some others (Spriggs et al., 2009; Cai and Lin, 2003; Han et al., 2004; Shao et al., 2009; Cheng et al., 2008). In addition, other kinds of machine-learning approaches are also applied, e.g., Naïve Bayes (Trribilini et al., 2006, 2007), decision tree (Carson, 2010), random forest (Liu et al., 2010; Ma et al., 2011), and neural networks (Jeong et al., 2004).

The attributes used by the feature-based methods can be roughly divided into sequential and structural ones. Sequential features mainly contain the identities for 20 types of amino acids and the residues' physicochemical information. In terms of amino-acid identity, it was reported

that arginine frequently occurs at DNA-binding sites and that both arginine and lysine are enriched at RNA-binding sites (Terribilini et al., 2007). Because the protein residues need to interact with the nucleic acids' negatively charged phosphate backbone, positively charged amino acids are inclined to show in nucleic-acids binding sites. As a matter of fact, the combination of residue identity and residue charge is also used to predict the proteins' binding sites (Carson, 2010). Physicochemical features include properties such as mixture of side-chain pK_a value, the hydrophobicity index, and the molecular mass of an amino acid in BindN+ (Wang and Brown, 2006); and a mixture of residue interface propensity, predicted residue accessibility, and residue hydrophobicity in PiRaNhA (Murakami et al., 2010). After the researchers explored the sequential-features space, the methods' performances to predict interface residues achieved a bottleneck. It was found that method using combination of only hydrophobicity and evolutionary information (Chen and Li, 2010) could achieve similar results as method integrating a large number of features (Chen and Jeong, 2009). Then it was the time to look for more effective features or information to improve the capacity of approaches to predict binding-site residues.

As an increasing number of 3-D protein structures and protein-nucleic acid complex structures becomes available, structure-based methods are gaining more thorough investigation. Structural features are often extracted from the solvent-accessibility surface areas and the 3-D coordinates of the protein structure. Generally speaking, structural features are more difficult to obtain than sequential ones, so methods based on sequential features have been studied earlier than structure-based ones. However, structural features usually achieve better performance for predicting binding-site residues for three main reasons. One reason is that the 3-D protein structure contains more information that does not exist in the primary sequence, such as the spatial contacting status between each pair of amino acids as well as the interaction between the

amino acid and nucleotide pair. The second reason is that the protein structure is more conserved than the protein sequence (Whisstock and Lesk, 2003). The third reason is that the features are extracted from the protein surface which removes the noises caused by non-surface residues. In addition, studies show a high correlation between a protein's secondary structure and protein-nucleic acids interactions (Allers and Shamoo, 2001), and the preferences between protein secondary structures and bases (Zhang et al., 2010).

Examples of structural features include protein surface patches that are built on electrostatics and geometric information (Shazman and Mandel-Gutfreund, 2008; Chen and Lim, 2008), accessible surfaces, a betweenness and retention coefficient (Maetschke and Yuan, 2009), structural neighboring information (Li et al., 2010), the fusion of a secondary structure, solvent accessibility, a side-chain environment and interaction propensity (Liu et al., 2010), hybrid features based on amino-acid identity, surface roughness, interface propensity, and protrusion score (Towfic et al., 2010). In addition, some structure-based methods use scoring functions to predict the interface areas. Among them, one method (Kim, 2006) exploits 3-D neighbored residue pairs to score each surface residue for its potential to interact with RNA nucleotides. As another example, the Optimal Protein-RNA Area (OPRA) (Perez-Cano and Fernandez-Recio, 2010) calculates the energy score for each residue using interface propensity weighted by accessible surface area. Besides template- and feature-based methods, geometry-based approaches, which mainly focus on the attributes of protein surfaces, such as size, shape, depth of clefts, and height of protruded areas (Iwakiri et al., 2012), are also used for predicting binding-site residues.

Evolutionary information is also used to improve binding site prediction methods. The information can provide the mutation history for a certain kind of protein family and can indicate

the conserved regions on the protein sequence. A popular form of evolutionary information is the position-specific scoring matrix (PSSM) which is created for each protein of interest. Given a protein, evolutionary information is commonly built by aligning the protein against the National Center for Biotechnology Information's non-redundant (NCBI-NR90) database (Ahmad and Sarai, 2005) using Position-Specific Iterative Basic-Local-Alignment Search Tool (PSI-BLAST) (Altschul et al., 1997). In the resultant matrix, a residue at the protein sequence's *i*th position is presented by a vector consisting of the loglikelihood for 20 types of amino acids.

1.3. Docking Methods

Docking methods are usually used to predict protein-nucleic acid interactions by generating decoys and scoring them, where decoys are potential ways to bind the protein and nucleic acids with different positions and conformations. This approach not only identifies the protein's binding sites, but also predicts the structure of the resulting protein-nucleic acids complex. This approach can be divided into two steps: the first step is to generate decoys using the docking approaches; the second step is to evaluate each decoy complex by using a scoring function. During the docking process, the generated complex structures not only reveals the nucleic-acid-binding sites on the protein structures, but also show the detailed atomic interaction between the protein and the nucleic acids. A docking method discretizes, or continuously searches, the conformation space of the complex with various rotations, translations, or both. During a docking procedure, one of the molecules usually the heavier one-stays static, and the other molecule approaches until they bind. The factors considered during the docking the procedure include electrostatic and geometric complementarity, shape complementarity, desolvation, and biochemical or biophysical information. There are several docking methods developed in the past two decades. The well-known docking methods include FTDock is

designed for two molecules based on geometric and electrostatic complementarities using Fourier transforms to speed up the searching procedure (Gabb et al., 1997). The Global Range Molecular Matching (GRAMM) methodology is an empirical approach that enhances the protein-protein docking quality by taking hydrophobic groups at contact sites into account (Vakser and Aflalo, 1994). High Ambiguity Driven Docking (HADDOCK) is for protein-protein docking that makes use of biochemical and/or biophysical interaction data (Dominguez et al., 2002). RPDock is based on FTDock and incorporates features specific to RNA-protein interfaces (including looser atom packing at the interface, the preference of positively charged residues at RNA-protein interfaces, and stacking interactions between the bases of nucleotides and the aromatic rings of the charged amino acids) (Huang et al., 2013). Because the methods need to exhaustively try the entire space of complex poses, bottlenecks for the docking methods exist in the tradeoff between tremendous computation and docking quality.

1.4. Scoring Functions

Docking methods do not evaluate the each protein-nucleic acid pose's fitness, so we need to build scoring functions to measure how similar a pose is to the native complex. Many scoring functions have been developed to predict protein-nucleic acid interactions (Liu et al., 2005; Zhang et al., 2005; Zhou and Zhou, 2002; Xu et al., 2009; Zhao et al., 2010; Samudrala and Moult, 1998; Robertson and Varani, 2007; Huang and Zou, 2014; Li et al., 2012; Zhao et al., 2011; Perez-Cano and Fernandez-Recio, 2010; Perez-Cano et al., 2010; Chen et al., 2004; Zheng et al., 2007; Tuszynska and Bujnicki, 2011; Huang et al., 2013). These methods can be divided into two groups. The first group simulates various physical and chemical forces between atoms (Liu et al., 2005; Zhang et al., 2005; Xu et al.; 2009, Zhao et al., 2010). The second group uses knowledge-based statistical energy functions that are derived from the observed contacting pairs

across the interface (Samudrala and Moult, 1998; Zhou and Zhou, 2002; Robertson and Varani, 2007).

Knowledge-based methods utilize statistics for the interactions between proteins and nucleic acids that are collected from a database, e.g. Protein Data Bank (Berman et al., 2003), with known protein and nucleic-acids structures. Berman et al. assume that protein-nucleic acids interactions can be described by energy functions with multiple parameters when they developed the knowledge-based methods. According to the granularity of contacting pairs, knowledge-based scoring functions can be further divided into three categories: amino acid and nucleotide pairs, fragment pairs, and atom pairs. An amino acid and nucleotide pair contains one residue from the protein's binding site and one nucleotide from the nucleic acids. For example, Iwakiri et al. (2012) dissect the paring preferences of amino acids and nucleotides at the interface area. Due to the increased number of high-resolution protein-nucleic acids complex structures, more knowledge-based scoring functions derive binding affinity based on interactions between protein atoms and nucleic-acid atoms involved with binding (Xu et al., 2009; Robertson and Varani, 2007). However, the weakness of atom-level methods is their complexity which demands greater computation compared to other methods. In addition, the methods do not consider the binding motifs that are known to occur among unrelated proteins (Denessiouk and Johnson, 2000; Denessiouk et al., 2001; Kinoshita et al., 1999; Kobayashi and Go, 1997). Motivated by the methods for predicting ligand-binding "hot spots" (Brenke et al., 2009; Kasahara et al., 2010), the fragment-pairs based methods gain more attention for solving the protein-nucleic acid problem (Guo and Wang, 2011) due to its tradeoff between accuracy and computation amount. In addition, the fragment-based interaction methods can disclose the 3-D distributions of protein atoms around nucleotide fragments and vice versa. The key point for using fragment-based

methods lies in how to segment the amino acid and nucleotide. For instance, a nucleotide can be divided into three fragments: sugar ring, base, and phosphate group.

Finally, the aforementioned methods are assessed by their capacities to discriminate between near-native decoys and coarse decoys, or between native complex structures and near-native decoys. A native complex is the structure of protein and nucleic acids that are naturally bound to each other. The decoy complex is a non-native complex structure that is generated from a native complex through a docking procedure. Based on the C_α root mean square deviation (RMSD) of the decoy to the native complex (after superimposing the native and decoy nucleic-acid structures), the decoy complex can be classified as a near-native decoy or a coarse decoy. The near-native decoy structure is more similar to the native complex and has a lower RMSD compared to a coarse decoy. In the testing stage, for each native complex, we generate a set of decoys using the docking procedure. Then, we utilize the potential function to calculate the score for each decoy and rank the decoys according to the scores. Ideally, the native complex should be ranked in the first position by its score. Thus, the better the scoring function is, the higher the native complex should be ranked. There are various ways to measure the scoring function's discriminatory ability, among which z-score is a popularly used one. Given a list of scores, an individual's z-score is the result of dividing the difference between the individual's score and the average score by the standard deviation. The z-score indicates how far away that individual is from the mean in terms of the standard deviation in the normal distribution. If an individual's z-score is higher, there are more standard deviations between the mean and this individual. In the context of our study, a higher z-score for a native complex means that the scoring function can recognize the native complex as a native complex more effectively.

1.5. Contributions

For the protein-DNA interaction, we developed a knowledge-based and fragment-atom pair energy function to study the protein-DNA interaction. We divided the nucleotide into three fragments: phosphate group, base, and sugar ring. We analyzed the distribution of protein atoms around each structural component of the DNA and developed spatial-specific scoring matrices (SSSMs) based on the observed distribution. We showed that the SSSMs could be used as a knowledge-based energy function to discriminate between the native protein-DNA structures and various decoys.

For the protein-RNA interaction, we developed a graph-mining method that could identify the RNA-binding site on proteins and predict the protein-RNA complexes' structures. Our method extracted common sub-graphs from the interface areas of proteins with known binding sites, assuming that the binding sites share similar graphlets which can be used to characterize and reveal the repetitive patterns that exist at the binding sites. Another advantage of our method is its simplicity and efficiency. Once the common sub-graphs are collected during the training process, they can be used as features either to help the classifier recognize binding sites, or to discriminate between near-native decoys and docking decoys of proteins with unknown functions by simply calculating the number of occurrences for the selected common sub-graphs in the decoys' interface areas.

1.6. Dissertation Overview

This dissertation introduces two computational methods: one is a fragment-based method that serves as a scoring function to evaluate the fitness of each protein-DNA complex pose; the other one is a graph-mining method that utilizes common sub-graphs to recognize the RNA-

binding site on the protein surface and to score the generated poses. The remaining chapters of this dissertation illustrate my Ph.D. research work. They are organized as follows.

In Chapter 2, we will present a knowledge-based computational method for discriminating native protein-DNA complexes from decoys. A paper derived from this chapter was published in the proceedings of the 7th International Conference on System Biology (ISB).

Chapter 3 presents a graph-mining method that uses common sub-graphs to identify RNA-binding sites on proteins and to predict 3-D protein-RNA complexes. Publications derived from this chapter are under preparation.

In Chapter 4, several plans are discussed. These plans will be used to improve the performance of the common sub-graph method for predicting RNA-binding sites and discriminating near-native protein-RNA complexes from the docking decoys.

2. A METHOD FOR DISCRIMINATING NATIVE PROTEIN-DNA COMPLEXES FROM DECOYS USING SPATIAL-SPECIFIC SCORING MATRICES

Decoding protein-DNA interactions is important for understanding gene regulation and has been investigated by worldwide scientists for a long time. However, many aspects of the interactions still need to be uncovered. The crystal structures of protein-DNA complexes reveal detailed atomic interactions between the proteins and DNA and are an excellent resource for investigating the interactions. This study profiles the spatial distribution of protein atoms around six structural components of the DNA; the four bases, the deoxyribose sugar, and the phosphate group. The resultant profiles not only revealed the preferred atomic interactions across the protein-DNA interface, but also captured the interaction's spatial orientation. The profiles are a useful tool for investigating protein-DNA interactions. We tested the profiles' strength with two experiments: discrimination of native protein-DNA complexes from decoys with mutant DNA and discrimination of native protein-DNA complexes from near-native docking decoys. The profiles achieved an average Z-score of 7.41 and 3.22, respectively, on benchmark datasets for the tests; both experimental results are better than other knowledge-based energy functions that model protein-DNA interactions based on atom pairs.

2.1. Related Work

Many computational methods have been developed for predicting DNA-binding sites on protein structures. Some methods focus on the geometrical and physiochemical properties of the DNA-binding sites and use a data-mining or statistical approach to predict potential DNA-binding sites for new protein structures (Chikhi et al., 2010; Guo and Wang, 2012; Ito et al., 2012; Sael and Kihara 2012; Blanchi et al., 2012; Konc and Janezic 2010; Zhou and Yan, 2010). These methods usually represent patches on the protein surface using vectors or graphs, and then

compare the patcheswith known DNA-binding sites. These methods usually suffer relatively low accuracy, and some of them are very computationally demanding. Other methods rely on structural alignment (Wass et al., 2010; Kinoshita and Nakamura 2009). These methods maintain a database of protein structures for which DNA-binding sites are known. To predict DNA-binding sites for a new protein (also known as query protein), the new protein is used to query the database to find structures (also known as templates) that share a high similarity with it. The query protein structure is then aligned with the templates. The region on the query protein that superimposes with the known DNA-binding sites on the templates is predicted to be a DNA-binding site. These methods' success strongly depends on the availability of templates and the level of similarity between the query and templates.

Other researchers use docking approaches to predict the structure of the protein-DNA complex. The resultant complex structure not only reveals the DNA-binding sites on the protein structure, but also shows the detailed atomic interaction between the protein and the DNA. A docking method searches the conformation space of the complex and uses an energy function to score the conformations. Different docking methods vary in the energy function used. Some docking methods use functions that model various physical and chemical forces between atoms (Liu et al., 2005; Zhang et al., 2005; Zhou and Zhou, 2002; Xu et al., 2009; Zhao et al., 2010). Others use knowledge-based statistical energy functions derived from the observed interacting atom pairs across the interface (Samudrala and Moult, 1998; Robertson and Varani, 2007).

2.2. Methods and Materials

2.2.1. Datasets

The testing dataset for the first test, the DNA mutation decoy test, was composed of 51 non-redundant complexes from Kono and Sarai (1999). For the second test, the near-native

docking decoys were generated using FTDock (Katchalski-Katzir et al., 1992) from 45 protein-DNA complexes that were collected by Robertson and Varani (2007). The training datasets for both tests were derived from the 212 protein-DNA complexes used in Xu et al. (2009); these complexes were extracted from the Protein Data Bank (PDB) database and culled by the PISCES server (Wang and Dunbrack, Jr., 2003), which is a protein sequence culling server, such that pairwise similarity was less than 35%. For both tests, we removed complexes that had more than 35% similarity with any protein in the test sets. As a result, the training set for the DNA mutation decoy test contained 166 protein-DNA complexes, and the training set for the near-native docking decoy discrimination test contained 167 protein-DNA complexes.

2.2.2. Spatial-Specific Scoring Matrices (SSSMs)

We first divided the DNA into six repeating structural components: the four bases, the deoxyribose sugar, and the phosphate group. We collected the protein atoms that were in contact with these components and investigated how they were distributed around the components in the space. For each component, we defined a new coordinate system that centered on it. Using the new coordinate system as a grid, we divided the space into X*X*X cubes with X bins on each axis. The grid size was customized so that all the protein atoms that were in contact with the component fell into the cubes. We tried several values for X, such as 6, 8, 10, 16, and 20. When dividing the space into too few cubes, e.g., 6*6*6, the cube was too large and could not effectively represent the 3-D distributions of the protein atoms around the nucleotide component. One the other hand, if X was too large, only a few atoms existed in a cube, and most cubes were vacant. The large X caused low estimations too, and, furthermore, an increased computation amount. Finally, we decided to use 16*16*16 cubes, and the experimental results demonstrated that this choice yielded the best performance.

Inspired by previous research work (Murphy et al., 2000; Weathers et al., 2004; Peterson et al., 2009; Bacardit et al., 2009), we classified the protein atoms into 14 types based on the environment around them, as described in Petsalaki et al. (2009), and then counted the number of different types of atoms that fell into each cube. Previous research work (Murphy et al., 2000; Weathers et al., 2004; Peterson et al., 2009; Bacardit et al., 2009) showed that, by using a reduced alphabet for amino acids, the capacity of recognizing protein structures from protein sequences could be improved while the computation amount computation was reduced. The reason was because the simplified amino-acid alphabet kept sufficient information representing protein structures, and most importantly, it reduced noise. Inspired by Petsalaki et al. (2009), we classified the protein-atom into 14 types as described in their work based on the environment around the atoms. Then, we counted the number of different types of atoms that were in each cube. By using the compact atom types, we hoped to purge the unnecessary information and, meanwhile, simplify our 3-D model. The 14 atom types were: C3 (aliphatic carbons; sp3), C= (carbonyl carbon; sp2), O= (carbonyl oxygen; sp2), N2H (nitrogen of amides; sp2; also sp2 neutral nitrogen of side chains), Car (aromatic carbon; sp2; general), O2- (negatively charged oxygens (-1/2) in carboxylates; sp2), SH (sulphur in thiols; sp3), OH (hydroxyl group; sp3), NarH (aromatic nitrogen with a hydrogen; sp2), NarH+ (aromatic nitrogen with a hydrogen and a postive charge; sp2), Set (sulphur in thioethers; sp3), C+ (carbon of carbocations; sp2), N3H+ (sp3 nitrogen with a hydrogen and a positive charge), and N2H+.

Therefore, the distribution of protein atoms around a component was described using a 16*16*16*14 matrix. The matrix counts were normalized by the total count. Therefore, each cell in the matrix corresponded to one atom type and one cube in the space, and the cell's value showed how likely the atom would be to contact the DNA component from a location

corresponding to the cube. These matrices were populated using protein-DNA complexes in the training set. The resultant six matrices (which are referred as SSSMs) were used as scoring matrices to discriminate native protein-DNA complexes from various decoys. For a given structure (native or decoy) of protein-DNA complex, a score was assigned using the following method:

$$S = \sum_{i=1}^{6} \sum_{j=1}^{16*16*16} \sum_{k=1}^{14} O_{ijk} P_{ijk}$$

, where O_{ijk} is the number of atoms of type *k* that contact component *i* from the location corresponding to cube *j*, and P_{ijk} is the value in the cell of the scoring matrix for component *i* that corresponds to atom type *k* and cube *j*. Higher scores mean that the complex was more likely to be the native structure.

2.3. Experiments and Results

2.3.1. Test 1: To Discriminate Native Structures from DNA Mutation Decoys

For this test, 166 protein-DNA complexes were used as a training set to derive the 6 scoring matrices, and a disjoint test set consisting of 51 protein-DNA complexes was used to generate decoys. For each of native complex, we generated 50,000 decoys by replacing a nucleotide base with a different type of base that had equal opportunity. The new base was placed in the same plane as the native one. Then, we calculated the scores for the native complex and the decoys. Because the native complex only differed from the decoys in the bases, only the four SSSMs corresponding to bases were utilized in this test to calculate the scores. We used Z-score to evaluate the performance of discriminating the native complex and decoys. Here, Z-score = (S_{avg}- S_{native}) / *SD*, where S_{avg} and *SD* were the average and standard deviation for the scores of 50,000 decoys, and S_{native} was the score for the native structure. Because the native

structure was expected to have a higher score than the decoys, a lower negative Z-score meant that the scoring system was able to distinguish the native structure from decoys with a better performance. Our method achieved an average Z-score -7.41 with the test set. The Z-scores for each complex are shown in Table 1.

Many researchers have tried to develop knowledge-based energy functions for protein-DNA interactions based on the observed atomic contacts across the interface. Zhou and Zhou (2002) first applied a distance-scaled, finite ideal-gas (DFIRE) energy function for protein-DNA interaction. Gromiha et al. (2004) also developed energy functions based on intermolecular and intramolecular contacts. Xu et al. (2009) developed five variants for the DFIRE energy functions, among which the variant (named vcFIRE) with the low-count correction and volume correction achieved the best result. Xu et al. (2009) evaluated and compared these methods using the same training and test datasets that were utilized for the present study. We used the results from their study and compared our method with others. Table 1 shows that our method achieved better Z-scores than all other methods in all but two complexes. The only exceptions were 1cjg and 1xbr (shaded in gray in Table 1). For 1xbr, our Z-score was very close to the best. The paired t-test showed that our method outperformed all others with $p<0.0001$. The average Z-score for our method with the dataset was -7.41, which was much better than that of any other methods.

Table 1. Z-scores for different methods in the test of discriminating native structures from mutation decoys

PDB ID	Gromiha et al. (2004)	Zhou and Zhou (2002)	Xu et al. (2009)	Our Method
1a02	-1.8	-2.27	-3.29	-18.27
1a74	0.7	1.50	-4.17	-5.50
1b3t	-2.1	-1.15	-2.38	-2.44
1bhm	-1.3	-0.05	-3.26	-6.20
1bl0	-2.5	-2.23	-3.25	-8.56
1cdw	-0.6	1.64	-0.02	-5.45
1cjg	-1.4	-2.58	-0.81	-0.10

Table 1. Z-scores for different methods in the test of discriminating native structures from mutation decoys (continued)

PDB ID	Gromiha et al. (2004)	Zhou and Zhou (2002)	Xu et al. (2009)	Our Method
1cma	-1.6	1.02	-1.59	-2.69
1e66	-1.7	-3.22	-3.12	-4.01
1dp7	-0.7	0.76	-3	-3.02
1ecr	-1.1	0.53	-1.58	-5.01
1fjl	-1	2.59	-2.63	-11.53
1gat	-1.7	1.73	-1.27	-2.12
1gdt	-1.7	-0.04	-3.75	-10.70
1glu	-1.1	1.72	-1.95	-12.03
1hcq	-2.5	-0.85	-4.09	-10.11
1hcr	0.4	-0.25	-2.43	-3.70
1hdd	-1.8	0.95	-1.57	-6.48
1hlo	-1.6	0.29	-3.95	-5.83
1hry	-0.9	0.23	-1.33	-3.76
1if1	-1.7	-1.62	-1.96	-8.64
1ign	-2.2	-0.23	-5.32	-8.32
1ihf	-2.3	1.79	-1.81	-2.35
1j59	-0.8	-2.33	-3.79	-12.29
1lmb	-4.3	-1.48	-4.25	-7.04
1mdy	-2.5	2.81	-2.83	-14.06
1mey	-2.2	-1.52	-4.92	-9.84
1mhd	-1.9	0.56	-2.74	-7.12
1mnm	-3	0.20	-4.04	-8.24
1mse	-2	-0.69	-2.13	-4.46
1oct	-2.1	-0.37	-2.85	-8.96
1par	-1.7	-0.96	-2.42	-5.34
1pdn	-2.5	-1.06	-1.92	-8.41
1per	-1.1	0.20	-1.92	-8.53
1pue	-2.7	-1.27	-2.21	-11.13
1rep	-3.2	-2.2	-3.01	-12.57
1rv5	-0.3	0.11	-1.67	-3.99
1srs	-2.4	0.67	-3.62	-8.44
1svc	-2.2	-1.68	-4.27	-9.04
1tc3	-2.5	-0.24	-2.29	-6.46
1tf3	-2.3	-1.19	-3.56	-5.45
1tro	-3.1	-0.19	-4.05	-7.41
1tsr	-1.2	-2.38	-2.68	-8.74
1ubd	-2.1	-0.12	-4	-7.26
1xbr	-2.4	-2.76	-2.4	-2.21
1yrn	-2.9	-0.05	-3.78	-9.10

Table 1. Z-scores for different methods in the test of discriminating native structures from mutation decoys (continued)

PDB ID	Gromiha et al. (2004)	Zhou and Zhou (2002)	Xu et al. (2009)	Our Method
1ysa	-2.1	0.14	-4.01	-8.88
2bop	-1.7	-2.16	-3.12	-4.04
2drp	-2.3	1.40	-4.75	-21.02
3cro	0.3	-1.52	-0.57	-9.61
6cro	-2.3	-3.86	-3.79	-5.38
Mean	-1.8	-0.43	-2.86	-7.41

2.3.2. Test 2: To Discriminate Native Structures from Near-Native Docking Decoys

This experiment was designed to test the SSSMs' ability to discriminate native complexes from near-native docking decoys. We created 10,000 docking decoys for each of the 45 native complexes using FTDock. The 2,000 lowest-RMSD decoys (i.e., the 2,000 decoys that were most similar to the native complex) were selected; we refer to them as near-native decoys. Six SSSMs constructed from the 16*16*16 3-D model were derived using the 167 protein-DNA complexes from the training set. Then, these SSSMs were used to compute scores for the native complex and the near-native decoys.

For this test, we compared our method with the DFIRE-based methods developed by Zhou and Zhou (2002), Xu et al. (2009), and an all-atom distance-based method developed by Robertson and Varani (2007). These methods were all evaluated using the same training and test datasets that were utilized for this study. Our method achieved an average Z-score of -3.22, which was the best among all methods (Table 2). Our method achieved the best Z-score for 29 of the 45 protein-DNA complexes. The paired t-test confirmed that our method outperformed the others with $p<0.0001$.

Table 2. Comparisons of methods in terms of Z-scores for the test of discriminating native structures from near-native docking decoys

PDB ID	Zhou and Zhou (2002)	Xu et al. (2009)	Robertson and Varani, (2007)	Our method
1qna	-1.21	-1.79	-1.57	-2.36
1d02	-1.47	-2.63	-1.95	-4.91
1eon	-1.66	-3.09	-1.98	-3.52
1ckq	-1.02	-1.94	-1.14	-2.77
1dmu	-1.55	-4.16	-2.06	-3.06
1qpz	-2.2	-3.48	-2.55	-3.04
1au7	-1.52	-2.55	-1.96	-3.86
1je8	-1.85	-2.91	-2.04	-2.43
2cgp	-0.97	-1.99	-1.42	-2.07
1b3t	-1.38	-2.99	-1.94	-2.27
1tc3	-1.56	-2.67	-1.56	-3.02
1g9z	-2.63	-5.45	-3.29	-3.89
1zme	-2.13	-2.38	-2.26	-4.01
1a73	-1.85	-3.41	-2.3	-5.90
1jko	-1.77	-3.12	-2.16	-3.21
1bdt	-1.77	-3.19	-1.88	-3.13
2bop	-1.68	-2.97	-2.13	-2.55
1a1i	-1.44	-2.49	-1.98	-5.09
1bc8	-1.5	-2.67	-2.1	-3.22
1pdn	-1.45	-2.47	-2.17	-3.13
1skn	-1.23	-2.6	-2.06	-4.98
1mjo	-2.09	-2.55	-2.16	-3.12
1bl0	-0.96	-1.92	-1.4	-1.70
2dgc	-1.46	-2.36	-2.06	-1.30
3pvi	-1.65	-2.34	-1.86	-2.19
2hdd	-2.37	-3.13	-2.7	-4.82
1ign	-1.74	-3.44	-2.3	-3.06
1qpi	-2.12	-3.67	-3.07	-3.26
1a3q	-1.46	-2.49	-1.91	-3.02
1dfm	-1.23	-2.6	-1.51	-1.97
1lq1	-1.94	-3.26	-2.38	-2.73
1tro	-1.43	-2.78	-2.05	-2.86
1fjl	-1.36	-2.12	-1.58	-3.45
1h8a_a	-1.29	-2.35	-2	-1.52
1h8a_b	-1.02	-2.18	-1.59	-4.71
1f4k	-1.16	-2.58	-2.1	-2.74
6pax	-1.21	-2.74	-1.96	-1.28
1hlv	-1.77	-3.17	-2.23	-2.48

Table 2. Comparisons of methods in terms of Z-scores for the test of discriminating native structures from near-native docking decoys (continued)

PDB ID	Zhou and Zhou (2002)	Xu et al. (2009)	Robertson and Varani, (2007)	Our method
1mnn	-1.59	-3.4	-2.49	-5.68
1dsz	-1.12	-2.38	-1.82	-2.79
1hwt	-1.77	-1.96	-2.4	-2.65
1per	-1.44	-2.7	-2.08	-3.62
1l3l	-1.76	-3.1	-2.42	-4.54
3hts	-0.95	-3.03	-2.05	-3.32
3bam	-1.66	-2.86	-1.99	-3.70
Mean	**-1.56**	**-2.8**	**-2.06**	**-3.22**

While the above results were obtained with the SSSM based on a 16*16*16 grid, we also tested the SSSM's performance using the 6*6*6, 8*8*8, 10*10*10, and 20*20*20 models. We measured their performances in terms of the total number of decoys with potential scores that were higher than the native complexes, average Z-score, standard deviation of Z-score, and t-test. We can see from the comparisons (shown in Table 3) that the 6*6*6 and 8*8*8 models led to too many decoys with higher scores than the native complexes while their average Z-scores were much lower than the other models. Therefore, we did not consider these two models. The 20*20*20 model had the smallest number of decoys with higher scores, but was only a little better than the 16*16*16 model. The 10*10*10 model yielded the highest average Z-score, 3.45, slightly better than the ones for 16*16*16 and 20*20*20. However, its standard deviation of Z-score is much higher than the other models. Thus, we excluded the 10*10*10 model, too. To choose from the 16*16*16 and 20*20*20 models, the former one outperformed with a higher average Z-score, smaller standard deviation for the Z-score, and smaller t-test score. Thus, we finally decided to use the 16*16*16 model to build the SSSM as a scoring function. The 5 models' performances also agrees with our speculation of choosing appropriate 3-D model.

Table 3. Comparisons of SSSM method using various grid dimensions for the test of discriminating native structures from near-native docking decoys

SSSM models	# of decoys with higher scores than native complex	Average z-score	Standard Deviation of z-score	t-test
6*6*6	4284	1.76	0.69	4.78E-12
8*8*8	2909	2.05	0.74	1.07E-09
10*10*10	642	3.45	3.45	0.080738
16*16*16	278	3.22	1.45	0.013865
20*20*20	256	3.13	1.59	0.062958

2.4. Summary

We developed a knowledge-based scoring function to assess protein-DNA interactions. We divided the DNA into six repeating structural components and used spatial-specific scoring matrices (SSSMs) to capture the distribution of protein atoms around these components in the 3-D space. The proposed method was able to discriminate native protein-DNA complexes from various decoys with better performance than other knowledge-based energy functions. Compared with other energy functions derived from the observed atom contacts, the proposed SSSMs not only reflected the preferences for atomic interactions across the protein-DNA interface, but also captured the interactions' spatial orientations. The SSSMs will be a useful tool to investigate protein-DNA interactions.

3. A NOVEL GRAPH-MINIG METHOD FOR THE PREDICTION OF RNA BINDING SITES ON PROTEINS AND FOR THE PREDICTION OF THE PROTEIN-RNA COMPLEXES' THREE-DIMENSIONAL STRUCTURE

It is well known that protein-RNA interactions play important roles in various biological process, e.g., mRNA processing, gene expression, protein synthesis, DNA replication and repair, and cellular defense against pathogens. Therefore, understanding the underlying mechanism of that interaction is an imperative task. Due to the increasing number of discovered protein-RNA complexes, it is possible for researchers to analyze and characterize the interacting areas in order to acquire insight about this biological issue. Furthermore, thanks to proteomic projects, tons of proteins with 3-D structure information become available while their functions and functional sites remain unknown. All those factors make it natural and compulsory to develop computational algorithms to identify the RNA-binding sites on proteins and to recognize 3-D protein-RNA complexes automatically and efficiently. Compared with RNA, the interactions between protein and DNA are investigated with more mature techniques, mainly due to the wider diversity for RNA structures. While DNA usually exists in the formation of a double helix, RNA structures can be hairpins/stem-loops, bulges, and loops, making RNA's interactions with protein more complicated and unpredictable.

We introduce a graph-mining method that could identify the RNA-binding site on protein and predict the protein-RNA complexes' structure. Our method extracts common sub-graphs from the interface areas of proteins with known binding sites, assuming that the binding sites share similar graphlets which can be used to characterize and reveal the repetitive patterns that exist at the binding sites. Another advantage of our method is its simplicity and efficiency. Once the common sub-graphs (CSs) are collected during the training process, they can be used as

features either to be fed to classifier to recognize the binding sites, or to discriminate between near-native decoys and docking decoys for a protein with unknown functions by simply calculating the occurrences of the selected CSs in the decoys' interface areas.

3.1. Related Work

Regarding the computational methods developed for tackling the protein-RNA interaction issues, we can generally divide the methods into two categories in terms of the tasks they attempted to address: one category predicts the RNA-binding sites on proteins; the other category focuses on predicting the protein-RNA complex's structure, i.e., how a protein binds with RNA. It is worth noting that most approaches were designed for either task, but not both. Methods that predict binding sites on proteins (Chen et al., 2013; Spriggs et al., 2009; Murakami et al., 2010; Yang et al., 2014) can be labeled as either feature-based (also machine learning) or template-based methods. The feature-based methods employ information about the protein's sequence, structure, or both for prediction. In most cases, the approaches which use structure information can achieve better performances because the protein structure contains more information than the sequence. Some methods also incorporate sequential or structural profiles to improve the prediction accuracy. The classifiers used by feature-based methods include Support Vector Machine (SVM) (Kumar et al., 2007; Wang and Brown, 2006; Cai and Lin, 2003), Naïve Bayes (Terribilini et al., 2006, 2007), random forest (Liu et al., 2010), and neural networks (Jeong et al., 2004). Template-based approaches (Zhao et al., 2011) align the query protein's sequences or structures with a non-redundant protein dataset that has known binding sites for prediction. There are also a few other algorithms to predict RNA-binding sites on proteins by analyzing the relationship between protein shapes and preferred RNA bases (Iwakiri et al., 2012) or by using pairs of amino acids to characterize binding sites on the protein (Kim et al., 2007).

When predicting the structure of protein-RNA complexes, most published methods use knowledge-based scoring functions (Huang and Zou, 2014; Li et al., 2012; Zhao et al., 2011; Perez-Cano and Fernandez-Recio, 2010; Perez-Cano et al., 2010) to evaluate the fitness of a binding mode between protein and RNA. The methods are usually assessed by their ability to discriminate between near-native complex decoys and docking decoys, which are various binding modes generated by docking procedure. The advantage with these methods is their high accuracy. However, they are too computationally complicated and time-consuming due to their nature.

3.2. Methods and Materials

3.2.1. Datasets

Our study used three datasets. The first one was referred to as the training set. It included 3-D structures for the protein-RNA complexes which were experimentally determined using wet-lab methods such as x-ray crystallography and NMR. Each complex structure showed a native binding mode between the protein and RNA. In this study, the training set was used to discover common sub-graphs that wereabundant on the RNA-binding sites. The training dataset was obtained from the RCSB Protein Data Bank (PDB) database. We used an advanced search to retrieve all protein-RNA complexes from PDB, and the research returned 1,570 results. Then, the dataset was culled using PISCES (Wang and Dunbrack, Jr., 2003) with the mutual sequence similarity no more than 25%. After this filtering step, 144 protein-RNA complexes remained and included 153 chains. The second and third datasets included not only structures of the protein-RNA complexes, but also unbound structures of the involved proteins and RNA. They were used to test our method. Therefore, they were referred to as Testing Sets I and II. The sets consisted of 64 and 72 protein-RNA complexes, respectively. Testing Set I was from part of the extended

protein-RNA docking benchmark collected by Perez-Cano et al (2012). All 64 complexes were X-ray or NMR experimental structures with an available, unbound protein structure, where five had an unbound RNA structure, four had a pseudo-unbound (i.e., bound to a protein that had less than 35% sequence similarity with respect to that in the reference complex structure) RNA structure due to the lack of fully free structures, and the other 55 cases had bound RNA information. Testing Set II was a non-redundant set taken from another benchmark for protein-RNA docking in Huang and Zou's (2013) study. First, 87 bound protein-RNA complexes were obtained from the Protein Data Bank, and they were X-ray crystal structures with a resolution better than 4.0 Å, with less than 30% pairwise similarity for protein structures, and with less than 70% similarity for RNA. Then, BLAST was applied to the bound structures to acquire the corresponding unbound protein and RNA structures. Finally, 72 complexes were kept because the other 15 did not have available unbound protein or RNA structures. Both testing sets were also used in Huang et al.'s (2013) study to show the performance of their protein-RNA complex prediction method.

3.2.2. Extracting Interface Residues

The protein's RNA-binding sites are created with interface residues that are defined in Jones et al.'s (2003) study. We used the NACCESS software, an atomic solvent accessible area calculation program, to calculate the accessible surface area (ASA) of each amino acid for both the bounded and unbounded situations. If an amino acid's ASA in unbounded form was at least $1Å^2$ more than that in the bounded format, we considered this amino acid an interface residue. To obtain a general idea about how large a binding site is within a protein, we calculated the size of RNA-binding site based on the training dataset. The statistics showed that the binding-site size ranged from 4 to 82 residues. The average size of the binding sites was 33 residues, and the

average length of the protein sequences is 314, almost 10 times larger than binding site, while the standard deviation of the binding-site size was 19.3.

3.2.3. Building a Binding-Site Graph

Each RNA-binding site was represented using a graph, where each node illustrated an interface residue and where an edge was added between two nodes if their residues were in contact. Each node was labeled with its residue type. There were 20 residue types: alanine, arginine, asparagine, aspartic acid, cysteine, glutamic acid, glutamine, glycine, histidine, isoleucine, leucine, lysine, methionine, phenylalanine, proline, serine, threonine, tryptophan, tyrosine, and valine (A, R, N, D, C, E, Q, G, H, I, L, K, M, F, P, S, T, W, Y, and V). Two residues were considered as being in contact if the nearest distance between their heavy atoms was less than 0.5Å. Each edge was also associated with a label. If the two nodes at the end of an edge were sequence neighbors on the protein chain, then the edge was labeled as type one; otherwise, the edge was labeled as type two.

3.2.4. Finding Common Sub-Graphs that Are Abundant at the RNA-Binding Sites

There are 144 RNA-binding sites in the training set, and each one is represented as a graph. We refer to these graphs as binding-site graphs. We aim to find common sub-graphs that occur frequently at the RNA-binding sites. We implement the VF2 algorithm (Cordella et al., 2001, 2004) to fulfill this task. Generally, the algorithm is designed to find whether one graph is isomorphic to the sub-graph of the other by trying to match each pair of nodes from both graphs and ends when an isomorphism is found or when all pairs are searched without finding an isomorphism. Details of the VF2 algorithm are given in the following paragraphs.

Assume that there are two graphs, G and G', with *m* vertices and *n* vertices, respectively. The entire procedure is like a Depth-First Search of a tree (Figure 1), where the root represents

the start and where each tree node means matching between a pair of G and G' vertices. The maximum length of a tree's branch is MIN$\{m, n\}$, and the maximum number of branches is MAX$\{m, n\}$. The entire search stops either when it reaches a tree node at the deepest level of the tree (See the red node in Figure 1.) and finds a solution, which exists in the path from the tree root to that node; or after it searches all the pairs of G and G' vertices without finding a solution. During the search, the algorithm goes back one level up from the current node whenever it finds that the path from the tree root to the current node does not contain a match between the two graphs. Figure 2 gives an example of finding whether graph G' is an isomorphism of G using VF2.

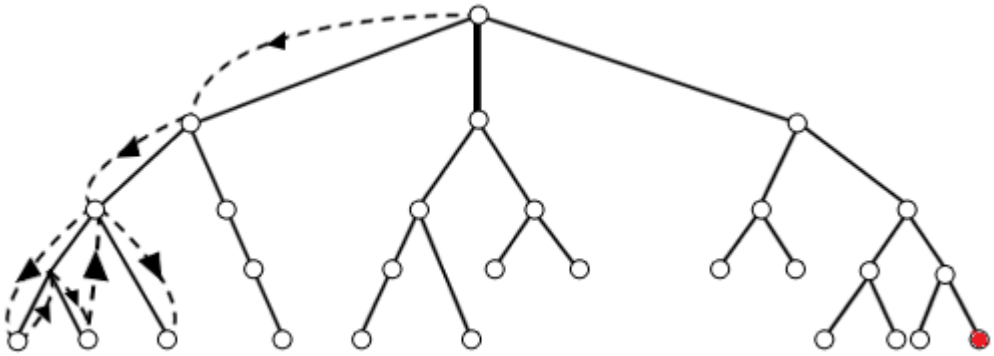

Figure 1. Illustration of a depth-first search of a tree

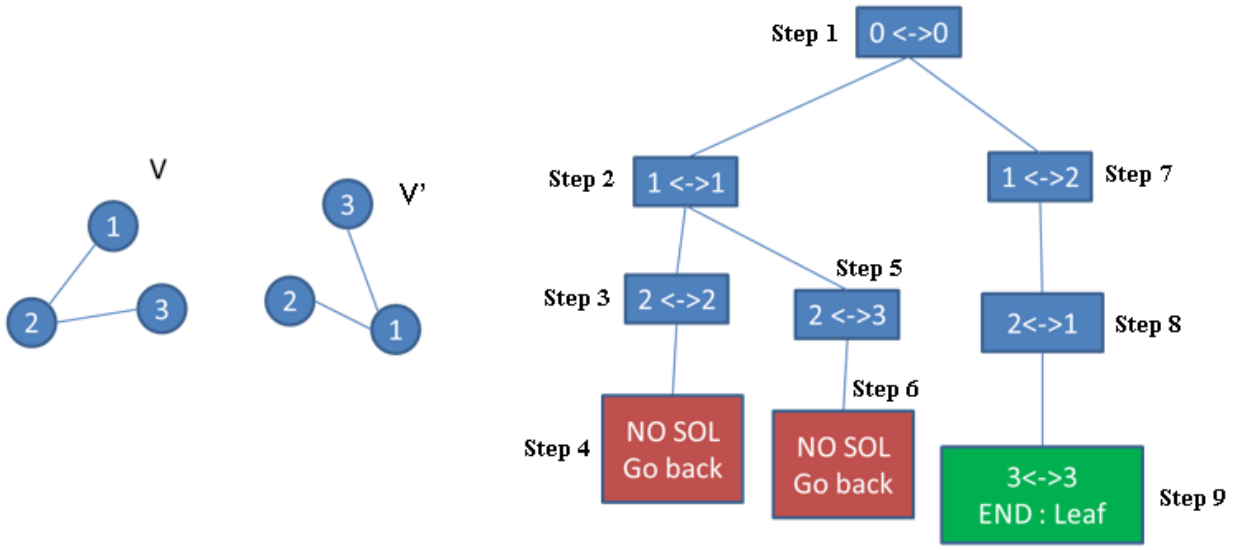

Figure 2. Example of the VF2 algorithm to find whether graphs V and V' are isomorphic

In Figure 2, graph V has three nodes {1, 2, 3} and V' has three nodes {1', 2', 3'}. The following steps illustrate how VF2 works:

Step 1: Match empty V with empty V', and it always works.

Step 2: Try to match node 1 with node 1', and it works.

Step 3: Try to match node 2 with node 2', and it works because {1, 2} and {1', 2'} are isomorphic.

Step 4: Try to match node 3 with node 3', and it does not work because there is an edge between nodes 2 and 3 with no edge between nodes 2' and 3'. In addition, there is no edge between 1 and 3, but nodes 1' and 3' are connected. Because all of the nodes in V' are used to match node 3 in V, we go back to step 2.

Step 5: Try to match nodes 2 with 3', and it works.

Step 6: Try to match nodes 3 with 2'. It does not work because there is an edge between nodes 2 and 3, but no edge exists between nodes 3' and 2'. Furthermore, nodes 1 and 3 are not connected, but nodes 1' and 2' are connected. We want to go back to step 2, but all

the nodes in V' (except 1' because it is now being matched with 1) have already been tried, so we can only to go back to step 1.

Step 7: Try to match node 1 with 2', and it works.

Step 8: Try to match node 2 with 1', and it works.

Step 9: Try to match node 3 with 3', and it works too.

Finally, we use VF2 to find that V and V' are isomorphic, and the match between them is (1, 2'), (2, 1'), and (3, 3').

For each pair of binding-site graphs, we found their CSs of sizes 3 and 4. Then, we obtained a set of CSs, among which some may be isomorphic to each other. After removing the duplicated CSs, we obtained 3,363 unique 3-node CSs and 7,482 unique 4-node CSs. For each CS, we used a vector with 144 values to indicate the CS's presence or absence in the RNA-binding sites of the 144 proteins, with 1 being presence and 0 absence. This vector was referred as a positive vector. To find the CSs that occurred with higher frequency at the RNA-binding sites than on the rest of the protein surface, we also collected non-binding sites on the protein surface. A protein's non-binding site was a surface patch that had the same number of residues as the binding site with the requirement that the non-binding site did not overlap the binding site. First, one non-binding site was randomly collected from each RNA-binding protein. Then, for each CS, a vector, named the negative vector, with 144 values was built to represent its presence or absence at the non-binding sites of the 144 proteins. We repeated these two steps 5 times in order to generate 5 negative vectors. Then, an average vector was computed such that each of its values was the average of the 5 corresponding values from the negative vectors.

We discarded the CSs that had fewer times of presence in the positive vector than in the average negative vector. Then, for the remaining CSs, we performed a t-test to compare the

positive vector and the average negative vector. A lower t-test score indicated that the CS was more favored by RNA-binding sites. We ranked the CSs in order of increasing t-test scores. Thus, the CSs at the top of the list were more likely to occur at the RNA-binding sites.

3.3. Experiments and Results

3.3.1. Test 1: Binding Site Evaluation

When protein and RNA interact, the interface adopts a certain conformation to achieve the required binding affinity. We hypothesize that conserved conformations utilized by the protein-RNA interactions could be characterized using small graphs enriched at the protein-RNA binding sites. Thus, discovering such small graphs helps to identify RNA-binding sites on the protein surface and to elucidate the interaction's mechanism.

In the previous section, we identified common sub-graphs (CSs) that were enriched at the RNA-binding sites. To verify that these CSs were crucial for binding the protein to RNA, we tested their ability to discriminate RNA-binding sites from non-RNA-binding sites. For this test, a set of surface patches from the protein was used. The surface patches included 144 RNA-binding sites and 144 non-RNA binding sites. We picked *n* high-ranking CSs from the top of the list. Then, each surface patch was encoded using a vector of *n* elements, such that each element indicated the presence or absence of a CS on the surface patch. Then, libSVM (Chang and Lin, 2011) was utilized to classify these surface patches as RNA-binding sites and non-binding sites using five-fold cross-validation and an RBF kernel.

The CSs used for this evaluation include the top-ranked 100, 200, 300, 400, and 500 CSs of 3-node size as well as the top-ranked 100, 200, 300, and 400 CSs of 4-node size. Table 4 shows the performance of libSVM using five-fold cross-validation and an RBF kernel, where

$$\text{precision} = \text{true positive} / (\text{true positive} + \text{false positive})$$

Table 4. Performance of classifying protein-RNA binding sites and non-binding sites using libSVM with 5-fold cross-validation and RBF kernel

Size of CS	# of CSs	TP	FP	FN	TN	Precision	Cross Valid Acc.	Acc.
3-nodes	Top 100	108	48	36	96	69.2308%	73.958%	70.833%
	Top 200	94	28	50	116	77.0492%	77.431%	72.917%
	Top 300	101	30	43	114	77.0992%	76.736%	74.653%
	Top 400	93	13	51	131	87.7358%	77.778%	77.778%
	Top 500	95	16	49	128	85.5856%	76.389%	77.431%
4-nodes	Top 100	88	15	56	129	85.4369%	76.042%	75.347%
	Top 200	112	14	32	130	88.8889%	84.028%	84.028%
	Top 300	118	29	26	115	80.2721%	80.903%	80.903%
	Top 400	109	21	35	123	83.8462%	80.903%	80.556%

From Table 4, we can see that, among all the selected CSs shown, the top 200 CSs of 4-node size yield the best precision, 88.89%; cross-validation accuracy, 84.03%; and accuracy, 84.03%.

3.3.2. Test 2: Validation Using the Results from Biological Experiments

Various experimental methods were used to study the protein-RNA interactions. These experiments confirmed that some interfaces' amino acids were crucial to the interactions. These experimentally confirmed binding-site residues were collected in the UniProt knowledgebase (UniProtKB) (UniProt Consortium, 2010). To further verify the importance of the CSs that we discovered, we compared the CSs with the UniProt binding-residue information.

For each of the 144 RNA-binding sites, we first find which of the top 200 CSs of size 4 occurred in each binding site and which residues are the nodes of those CSs. Then, we compare these residues with UniProt to see which ones are annotated as RNA-binding residues. In UniProt, the RNA-binding residues are labeled as one of the three categories: MUTAGEN (mutagenesis), REGION, or SITE. Residues with MUTAGEN labels are residues with functions have been tested using the mutagenesis experiment. With such an experiment, the residue is mutated to another type of residue. If the mutation affects the protein's ability to bind to RNA, then the residue is crucial for the RNA-binding function; therefore, it is on the RNA-binding site.

Otherwise, the residue is not crucial for RNA binding. REGION stands for the extent of an RNA-binding region in the protein sequence while SITE refers to single RNA-binding amino acid site on the sequence.

Table 5. Fractions of interface residues appearing in sites annotated as MUTAGEN in UniProtKB

PDBID	UniProt ID	# of CS residues in interface area	# of residues in MUTAGEN area	# of overlapped residues	ratio
2F8K	Q08831	5	2	1	0.2
1FEU	P56930	9	8	4	0.444
2XS2	Q64368	8	4	3	0.375
1H2C	Q05128	0	2	0	0
2A1R	O95453	4	4	1	0.25
4HOR	Q13325	0	21	0	0
3ZD6	O95786	4	1	0	0
1J1U	Q57834	0	1	0	0
3RC8	Q8IYB8	0	8	0	0
3O7V	Q96J94	0	2	0	0
3MOJ	P42305	5	2	0	0
3L25	Q05127	7	2	0	0
2A8V	P0AG30	4	2	2	0.5
2XLK	Q02MM2	7	9	0	0
2DU3	O30126	7	3	0	0
2VNU	Q08162	5	1	0	0
3RW6	Q9UBU9	13	24	0	0
2Y8Y	Q53WG9	9	1	0	0
4KXT	Q9UL18	20	4	0	0
2BGG	O28951	10	5	3	0.3
1WPU	P10943	13	13	0	0
2NUG	O67082	12	3	0	0
2PO1	Q9V119	0	3	0	0
3OIN	Q06287	0	4	0	0
3EQT	Q96C10	0	2	0	0
3PEY	P20449	4	8	2	0.5
1N78	P27000	16	1	0	0
3MDI	O43809	14	5	2	0.143
2XZO	Q92900	10	5	0	0

The ratio of residues that not only exist at a binding site and the top 200 CSs, but are also annotated as a certain category of RNA-binding residues in UniProt, to the residues existing in that binding site and the top 200 CSs is calculated for each of the 144 RNA-binding sites. Tables 5, 6, and 7 list the aforementioned ratios corresponding to MUTAGEN, REGION, and SITE respectively. It is worth noting that, because the annotation and information for the above three site types are incomplete in the UniProt knowledgebase, the tables do not show accurate results for our approach.

Table 6. Fractions of interface residues appearing in sites annotated as REGION in UniProtKB

PDBID	**UniProt ID**	**# of CS residues in interface area**	**# of residues in REGION are**	**# of overlapped residues**	**ratio**
2AZ0	P68831	0	73	0	0
1YVP	P42700	14	165	12	0.857
4G0A	Q03243	0	37	0	0
4HOR	Q13325	0	7	0	0
1J1U	Q57834	0	10	0	0
1K8W	P60340	17	29	4	0.235
1KNZ	P03536	4	146	4	1
3O7V	Q96J94	0	141	0	0
3MOJ	P42305	5	76	5	1
4IG8	P00973	13	59	8	0.615
2ZKO	P03496	10	73	10	1
1H4S	Q5SM28	4	30	0	0
3DH3	P32684	22	8	4	0.182
2A8V	P0AG30	4	17	1	0.25
3FOZ	P16384	19	28	13	0.684
3RW6	Q9UBU9	13	117	0	0
4KXT	Q9UL18	20	95	14	0.7
1JID	P09132	4	9	0	0
2BH2	P55135	16	24	6	0.375
3EQT	Q96C10	0	84	0	0
1N78	P27000	16	16	0	0
3MDI	O43809	14	146	10	0.714

Table 7. Fractions of interface residues appearing in sites annotated as SITE in UniProtKB

PDBID	UniProt ID	# of CS residues in interface area	# of residues in SITE area	# of overlapped residues	ratio
4HOR	Q13325	0	7	0	0
1J1U	Q57834	0	1	0	0
2JLV	Q2YHF0	0	1	0	0
3O7V	Q96J94	0	1	0	0
3BT7	P23003	4	3	0	0
2Q66	P29468	8	8	1	0.125
4IG8	P00973	13	1	0	0
2A8V	P0AG30	4	1	0	0
3FOZ	P16384	19	3	1	0.053
2BH2	P55135	16	2	1	0.063
3OIN	Q06287	0	4	0	0
3FTF	O67680	4	4	2	0.5
1N78	P27000	16	2	0	0
3MDI	O43809	14	3	1	0.071

3.3.3. Test 3: To Discriminate Between Near-Native Decoys and Docking Decoys

The purposes of our experiments were to test the abilities of CSs extracted from protein-RNA binding sites in our training set in order to distinguish between native protein-RNA complexes and decoy complexes. The experiments were conducted on two testing sets, Testing Set I and Testing Set II, and we compared our results to Huang et al. (2013) who developed a novel computational protocol, 3dRPC, to predict RNA-protein complexes. Based on our knowledge, their method outperformed the other algorithms in terms of accuracy for the 3-D structure prediction of protein-RNA complexes with the two testing datasets used in this study. Program 3dRPC consists of two parts: a docking algorithm, RPDock, which was designed for protein-RNA docking and DECK-RP which is a distance- and environment-dependent and knowledge-based potential function for discriminating native protein-RNA structures from decoy

complexes. We applied Huang's RP-Dock approach to our two test sets in order to generate the 1,000 best protein-RNA decoys for each native structure.

To measure the performance of the proposed approach, we ranked each set of 1,000 decoys using CSs; then, we compared our success rates and hit counts with the ones for DECK-RP over a series of prediction numbers, N_p. The success rate was defined as the fraction of complexes where at least one of their top N_p decoys, ranked by CSs, was a near-native structure. Herein, the near-native structure had a root mean square deviation (RMSD) of RNA less than 10 Å after the protein's superposition. The hit count was the average number of near natives within the top N_p decoys per complex. To rank the decoys, we extracted interface residues and built the binding-site graph from each decoy, like we did with the training set, and then simply counted the occurrence number for selected CSs at the binding site on each decoy.

Performances for the proposed method on Testing Set I are shown in Figure 3, where the top two graphs illustrate the testing results from all 64 cases of Testing Set I and the bottom two graphs are results from 38 cases of this testing set. The reason we examined 38 testing cases was because 26 cases in our training set also existed in Testing Set I. Thus, to make the CS method more persuasive, we also tested the non-overlapped part of the testing sets. The figure also shows a performance comparison for the proposed approach with top 1,400-ranked 3-node CSs, and DECK-RP to demonstrate the effectiveness of our method for detecting the best decoys from a massive number of candidates. For all 64 cases in Testing Set I, the success rates at N_p = 1, 10, and 100 are 17.07%, 34.15%, and 70.73% for the CS method, respectively; and 19.51%, 48.78%, and 75.61% for DECK-RP, respectively. The hit counts at N_p = 1, 10, and 100 were 0.17, 1.22, and 5.61 for the CS method; and 0.20, 1.71, and 6.07 for DECK-RP, respectively. For the 38 cases from Testing Set I, the success rates at N_p = 1, 10, and 100 were 8.70%, 26.09%, and

60.87% for the CS method; and 21.74%, 52.17%, and 82.61% for DECK-RP, respectively. The hit count for these 3 prediction numbers were 0.09, 0.87, and 4.57 for the CS approach; and 0.22, 1.61, and 6.43 for DECK-RP. Our method's performance achieved a level comparable to DECK-RP, an advanced, novel algorithm for recognizing 3-D protein-RNA structures. Our algorithm showed better performance, particularly when the testing set had 64 cases rather than 38. Furthermore, our method yielded a higher performance when being measured by the hit count.

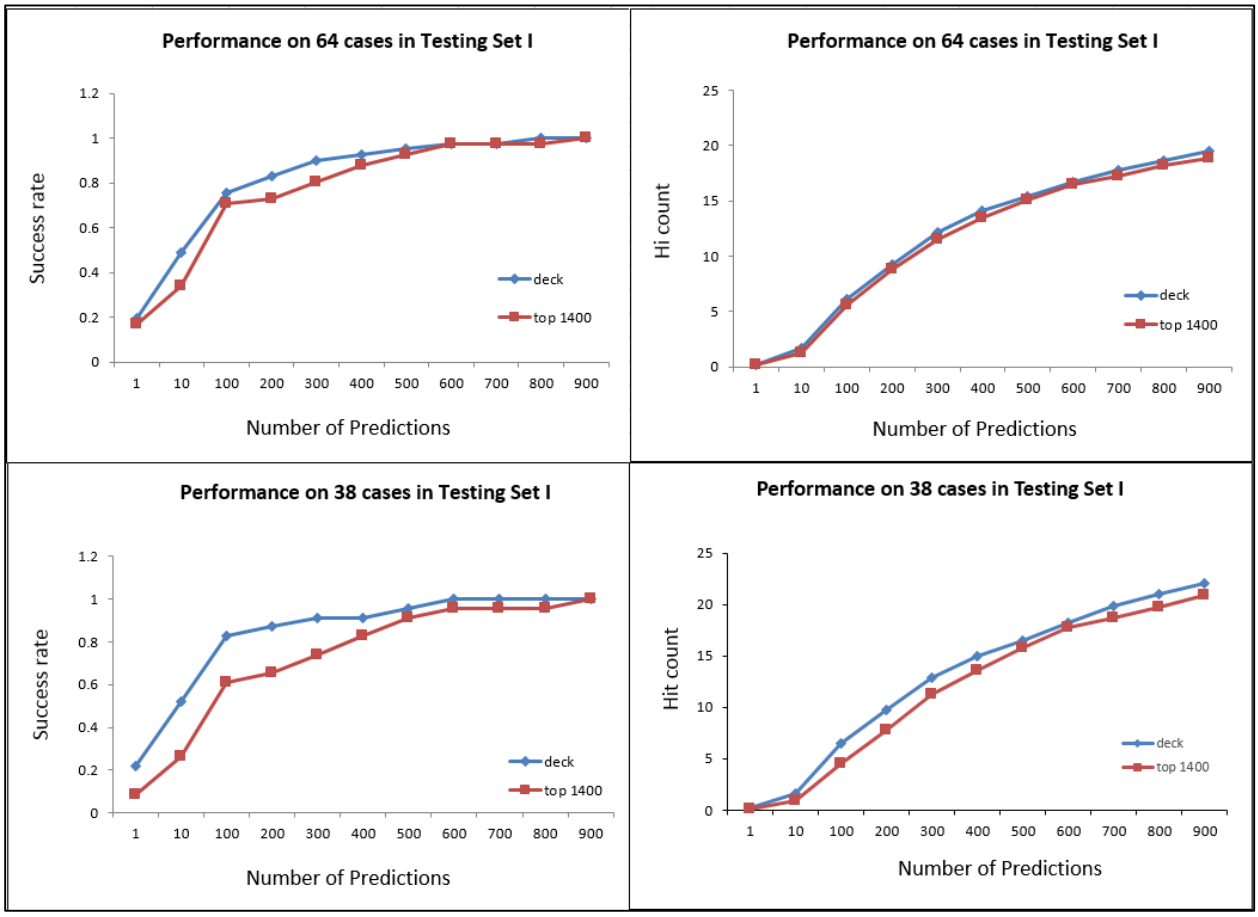

Figure 3. Success rate and hit count comparisons over the entire 64 cases of Testing Set I (top two graphs) and 38 cases of Testing Set I (bottom two graphs). Comparison is between the proposed algorithm, using top 1400 CSs of 3-node size, and DECK-RP

Performances for the proposed method with Testing Set II are shown in Figure 4. As with Testing Set I, we conducted an experiment not only with all 72 cases, but also on 37 cases which remained after removing 35 complexes that already existed in our training set. For the 72 cases

of Testing Set II, the success rates at N_p = 1, 10, and 100 were 32.00%, 56.00%, and 82.00% for the CS method, respectively; and 34.00%, 50.00%, and 86.00% for DECK-RP, respectively. The hit counts at N_p = 1, 10, and 100 were 0.32, 2.32, and 7.42 for the CS method, respectively; and 0.34, 2.38, and 7.78 for DECK-RP, respectively. For the 37 remaining cases, the success rates for the 3 prediction numbers were 30.77%, 50.00%, and 76.92% for the CS method, respectively; and 30.77%, 42.31%, and 84.62% for DECK-RP, respectively. The hit counts were 0.31, 2.31, and 6.69 for the CS approach, respectively; and 0.31, 2.04, and 7.31 for DECK-RP, respectively. We saw that our algorithm shows excellent results for Testing Set II, giving the same success rate as DECK-RP when the prediction number was 1 and even outperforming DECK-RP at N_p = 10. As we know, a high success rate and hit count for low prediction numbers are preferred in many situations. Due to the small size of the testing set, our algorithm may not exhibit its full discrimination ability.

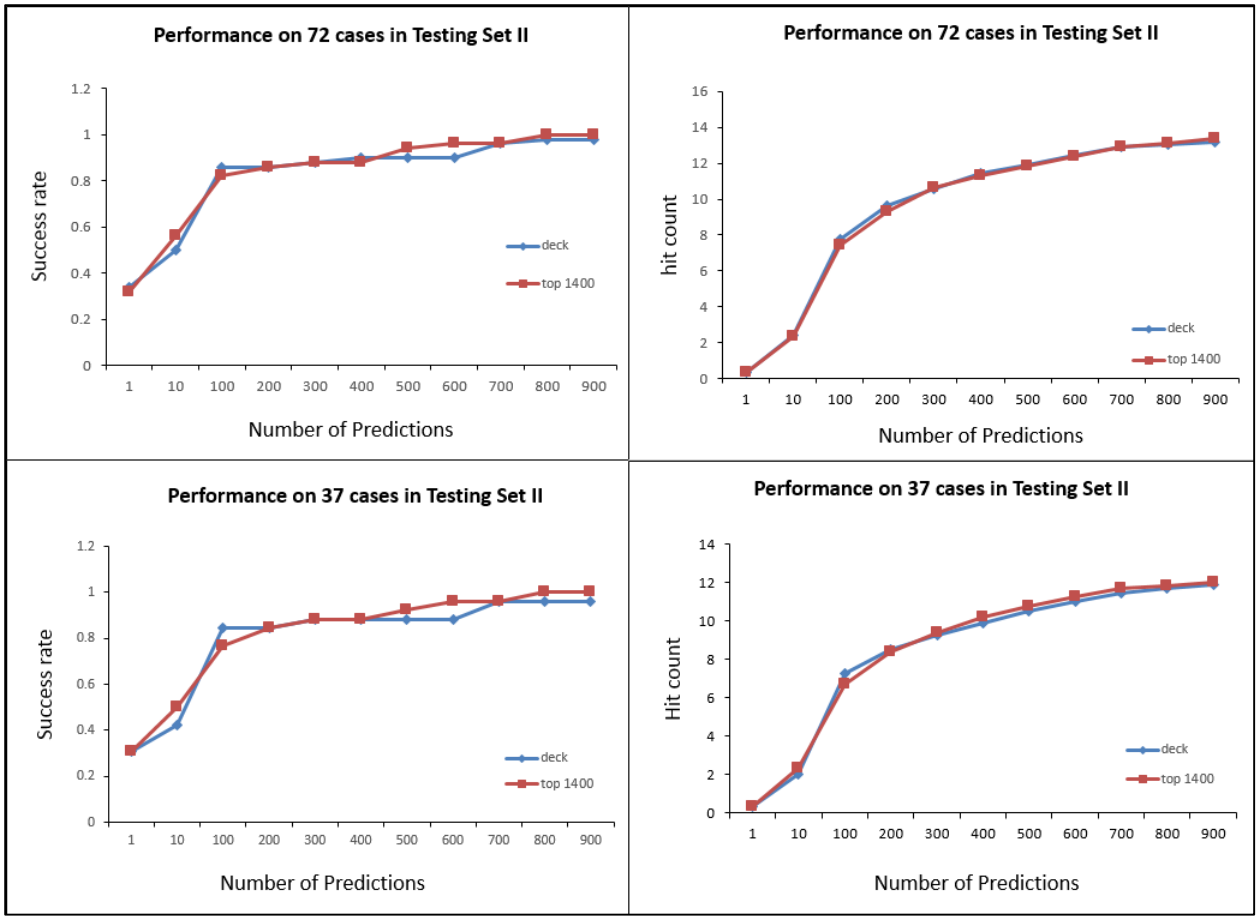

Figure 4. Success rate and hit count comparisons over the entire 72 cases of Testing Set II (top two graphs) and 37 cases of Testing Set II (bottom two graphs). Comparison is between the proposed algorithm, using top 1400 CSs of 3-node size, and DECK-RP

In order to analyze the mechanism of the CS algorithm, we also compared the performances of various CSs. Figures 5 and 6 show how some 3-node CSs, ranging from the top 500 to the top 1,400, perform at N_p = 100 with Testing Sets I and II, respectively. Theoretically, the more CSs that are involved with distinguishing near-native decoys from docking ones, the more accurate the discrimination will be. This observation is also proven by the two figures because both success rates and hit counts, overall, increase when more CSs are used. However, it is also worth noting that the performance is not perfectly proportional to the number of CSs used. For example, the top 700 CSs form local maxima in most graphs of Figures 5 and 6. The reason

can either be because the testing data are not sufficient enough or because using more CSs may bring more noise.

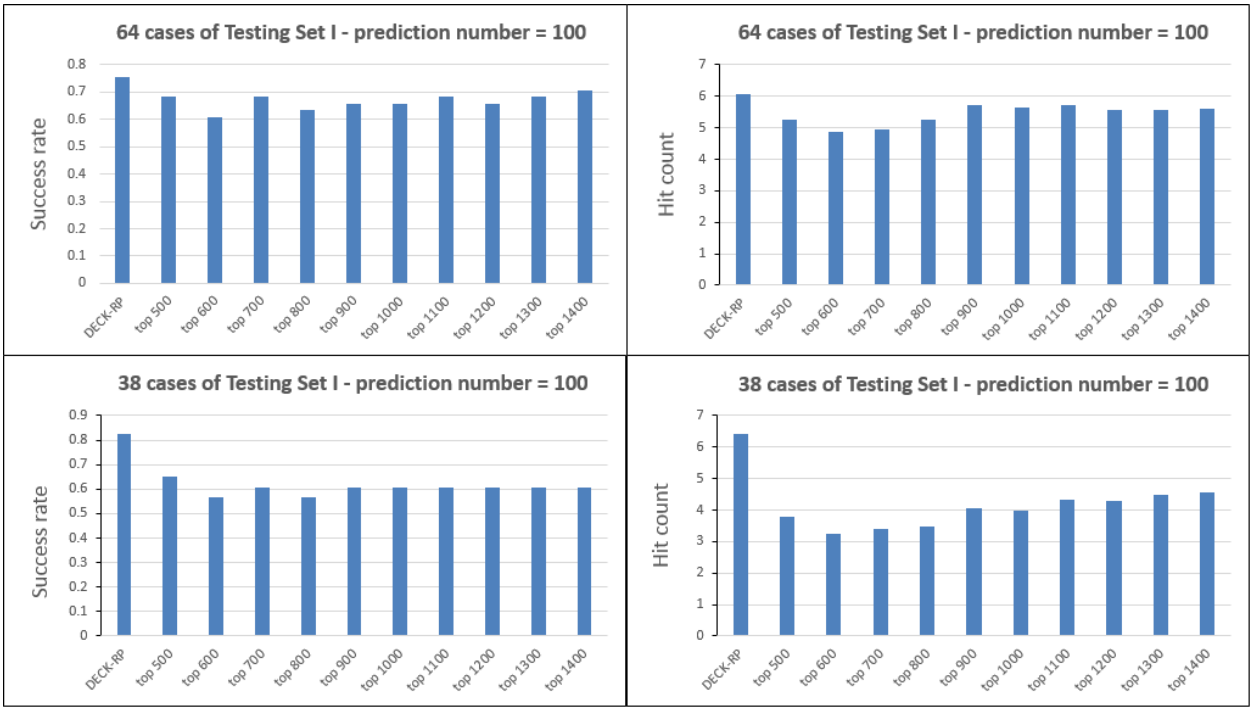

Figure 5. Performance comparison of the proposed method using difference CSs over 64 cases (top two graphs) and 38 cases (bottom two graphs) of Testing Set I

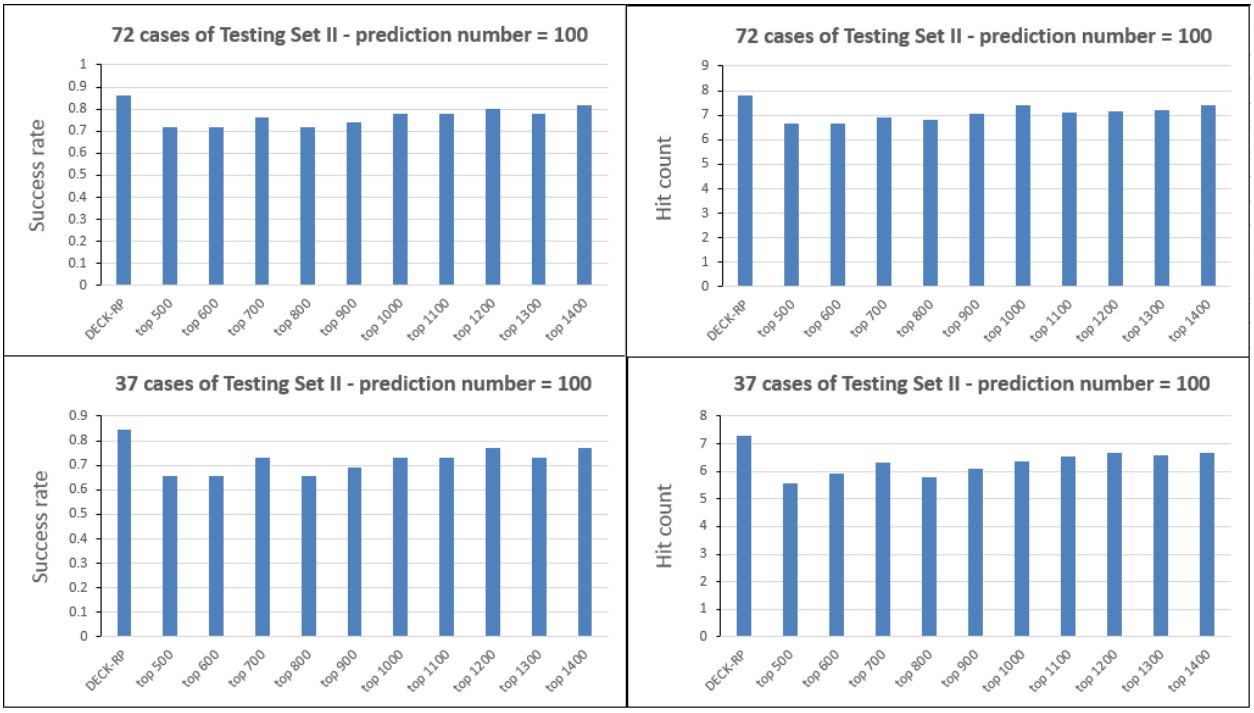

Figure 6. Performance comparison of the proposed method using difference CSs over 72 cases (top two graphs) and 37 cases (bottom two graphs) of Testing Set II

3.4. Summary

By taking advantage of the binding-site, repetitive sub-graph patterns, we proposed a graph-mining approach that can not only reveal the graph patterns that occur frequently at the protein-RNA binding sites, but can also predict 3-D protein-RNA complexes. We extracted interface areas from the protein surface and then built a graph for it, where the graph node represented an interface residue and an edge formed between residues that are close enough to each other. For all pairs of binding-site graphs in the training set, we searched for certain-size CSs and removed redundant ones. Then, the t-test method was applied to rank the distinct CSs. The proposed method was evaluated in terms of two discrimination abilities: recognizing RNA-binding sites and distinguishing near-native decoys from docking decoys. The experimental results were highly competitive when tested on recently used benchmarks. Compared with other state-of-the-art methods, the proposed CS algorithm was easily understood and was simple while

being effective at the same time. Our method was also suitable for problems relevant to protein-DNA interactions.

4. IMPROVEMENT FOR THE CS APPROACH

There are three main tasks that we will consider as our potential future work: the first task is to improve the accuracy of our CS method for discriminating native-complex decoys and docking decoys by using the graph kernel technique; the second task is to use the CS method to locate the binding-site position on the protein surface; and the last task is to incorporate the binding-site properties to make the CSs more effective. These tasks are described, in detail, in the following sections.

4.1. Improving the Common Sub-Graph Method with a Graph Kernel

Conformation of the protein-nucleic acids complex can change in water or other solvents, causing structure changes for the binding sites on proteins. Thus, instead of enumerating the exact matching of CSs on the protein surface, we need to design a more mismatch-tolerant way to accommodate the structural changes. Inspired by the work of Alvarez and Yan (2012) who invented a graph-kernel method to predict protein functions, we plan to borrow their idea and apply it to our CS method in order to improve performance. Given a query protein with unknown function, their work encrypts the protein into a graph where each node represents a cluster of amino acids and is labeled with a vector containing information about the amino acid composition. Then, the graph is fed to the SVM as a kernel to compare its similarity with other encoded graphs that have known functions in the database. The function of the graph most similar to the query graph is predicted to be the one same as the query graph. Because the graph-kernel method calculates the similarity between two graphs while our algorithm checks for the exact existence of a CS, we need to make some adjustments. Given a protein-nucleic acid decoy, instead of checking whether a CS occurs in the interface area of the decoy, we can calculate the similarity between this CS and each sub-graph of the same size on the decoy's interface area. We

may also need to set a threshold for similarity. In this way, the CS method may become more mistake tolerant and robust.

4.2. Locating the Binding-Site Using Common Sub-Graphs

One strength of the proposed CS method is that it reveals the graph patterns that occur frequently at protein-RNA binding sites. However, this method does not directly tell where the binding site is on the protein surface. To solve this problem, we can extend our work by searching where the selected CSs occur. Motivated by Kasahara et al.'s (2010) work that is based on a knowledge-based approach using fragment-fragment contacting pairs to predict ligand-binding sites on proteins, we plan to cluster the repetitively occurring CSs and rank the "hot spots" on protein surface. The "hot spot" with the highest score can be predicted to be the binding site.

4.3. Extracting Common Sub-Graphs with More Effectiveness

Many studies investigated the properties of RNA-binding site regions on proteins and found that RNA-binding sites have unique characteristics. For example, Iwakiri et al. (2012) classified the protein's interface areas into three shapes, dented, intermediate, or protruded, and calculated the amino-acid compositions for each shape type. Iwakiri et al. (2012) also discovered the relationships between protein surface shapes and the contacting nucleotides. Huang et al. (2013) observed that positively charged amino acids prefer to appear at RNA-protein interfaces, and some other studies revealed that protein-RNA interfaces have an abundant occurrence of arginine-rich patterns (Jones et al., 2001; Kim et al., 2003, 2006; Li et al., 2008). In addition, several papers (Bahadur et al., 2008; Perez-Cano et al., 2010; Terribilini et al., 2006) also analyzed the residue preference at the interface. Sparked by these studies, we intend to exploit the proclivities of the residues and/or shapes at the protein interface to help build CSs with

stronger discriminatory ability. In addition to the aforementioned improvements, there are a couple other tasks we may do: the first thing is to conduct experiments to test the CS method's ability to discriminate between native protein-RNA complexes and near-native decoys; the second task is to design our own protein-RNA docking procedure and to integrate it with our CS method.

BIBLIOGRAPHY

Ahmad, S. and Sarai, A., 2005. PSSM-based prediction of DNA binding sites in proteins. BMC Bioinformatics, 6, 33.

Alamanova, D., et al., 2010. Creating PWMs of transcription factors using 3D structure-based computation of protein-DNA free binding energies. BMC Bioinformatics 11: 225.

Allers, J. and Shamoo, Y., 2001. Structure-based analysis of protein-RNA interactions using the program ENTANGLE. J. Mol. Biol. 311: 75-86.

Altschul, S., et al., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucl. Acids Res. 25: 3389-3402.

Alvarez, M. and Yan, C., 2012. A new protein graph model for function prediction. Journal Computational Biology and Chemistry. 37:6-10.

Bacardit, J., et al., 2009. Automated alphabet reduction for protein dataset. BMC Bioinformatics, 10(1): 6.

Bahadur, R.P., et al., 2008. Dissecting protein-RNA recognition sites. Nucl. Acids Res. 36: 2705-2716.

Blanchi, V., et al., 2012. Identification of binding pocket in protein structures using a knowledge-based potential derived from local structural similarities. BMC Bioinformatics 13(Suppl 4): S17.

Brenke, R., et al., 2009. Fragment-based identification of druggable ‘hot spots’ of proteins using Fourier domain correlation techniques. Bioinformatics 25: 621-627.

Cai, Y. and Lin, S., 2003. Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence. Biochimica et Biophysica Acta (BBA) 1648: 127-133.

Carson, M. B., et al., 2010. NAPS: a residue-level nucleic acid binding prediction server. Nucl. Acids Res. 38: W431-W435.

Chang, C. and Lin, C., 2011. LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology 2, 27:1-27:27. http://www.csie.ntu.edu.tw/~cjlin/libsvm.

Chen, P. and Li, J., 2010. Sequence-based identification of interface residues by an intergrative profile combining hydrophobic and evolutionary information. BMC Bioinformatics, 11: 402.

Chen, X.W. and Jeong, J.C., 2009. Sequence-based prediction of protein interaction sites with an integrative method. Bioinformatics, 25: 585-591.

Chen, Y., et al., 2004. A new hydrogen-bonding potential for the design of protein-RNA interactions predicts specific contacts and discriminates decoys. Nucl. Acids Res., 32: 5147-5162.

Chen, Y., et al., 2013. Identifying RNA-binding residues based on evolutionary conserved structural and energetic features. Nucl. Acids Res., Advance Access.

Chen, Y.C. and Lim, C., 2008. Predicting RNA-binding sites from the protein structure based on electrostatics, evolution and geometry. Nucl. Acid Res., 36, e29.

Cheng, CW., et al., 2008. Predicting RNA-binding sites of proteins using support vector machines and evolutionary information. BMC Bioinformatics 9 (Suppl 12): S6.

Chikhi, R., et al., 2010. Real-time ligand binding pocket database search using local surface descriptors. Proteins: Structure, Function, and Bioinformatics 78, 2007-2028.

Cordella, L. P., et al., 2001. An Improved Algorithm for Matching Large Graphs. 3rd IAPR-TC15 Workshop on Graph-based Representations in Pattern Recognition, Cuen: 149-159.

Cordella, L. P., et al., 2004. A (Sub) graph Isomorphism Algorithm for Matching Large Graphs. IEEE Transactions on Pattern Analysis and Machine Intelligence, 26: 1367-1372.

Csardi, G. and Nepusz, T., 2006. The igraph software package for complex network research. http://igraph.org.

Denessiouk, K.A. and Johnson, M.S., 2000. When fold is not important: A common structural framework for adenine and AMP binding in 12 unrelated protein families. Proteins: Struct. Funct. Genet. 38: 310-326.

Denessiouk, K.A., et al., 2001. Adenine recognition: A motif present in ATP-, CoA-, NAD-, NADP-, and FAD-dependent proteins. Proteins: Struct. Funct. Genet. 44: 282-291.

Dominguez, C., et al., 2003. HADDOCK: a protein-protein docking approach based on biochemical-or biophysical information. J. Am. Chem. Soc., 125: 1731-1737.

Gabb, H., et al., 1997. Modelling protein docking using shape complementarity, electrostatics and biochemical information. J. Mol. Biol. 272: 106-120.

Gallet, X., et al., 2000. A fast method to predict protein interaction sites from sequences. J Mol Biol. 302:917-26.

Gao, M. and Skolnick J., 2008. A knowledge-based method for the prediction of DNA-protein interactions. Nucl. Acids Res. 36: 3978-3992.

Gromiha, MM., et al., 2004. Intermolecular and intramolecular readout mechanism in protein-DNA recognition. J. Mol. Biol. 337, 285-294.

Guo, F. and Wang, L., 2012. Computing the protein binding sites. BMC Bioinformatics 13(Suppl 10):S2.

Hall, K.B., 2002. RNA-protein interactions. Curr. Opin. Struct. Biol. 12: 283-288.

Han, L., et al., 2004. Prediction of RNA-binding proteins from primary sequence by a support vector machine approach. RNA 10: 355-368.

Huang, S. Y. and Zou, X., 2013. A nonredundant structure dataset for benchmarking protein-RNA computational docking. J Comput Chem 34: 311-318.

Huang, S. Y. and Zou, X., 2014. A knowledge-based scoring function for protein-RNA interactions derived from a statistical mechanics-based iterative method. Nucl. Acids Res., Advance Access.

Huang, Y., et al., 2013. A novel protocol for three-dimensional structure prediction of RNA-protein complexes. Sci Rep 3, 1887.

Ito, J., et al., 2012. PDB-scale analysis of known and putative ligand-binding sites with structural sketches. Proteins: Structure, Function, and Bioinformatics 80: 747-763.

Iwakiri, J., et al., 2012. Dissecting the protein-RNA interfaces: the role of protein surface shapes and RNA secondary structures in protein-RNA recognition. Nucl. Acids Res., Advance Access.

Jeong, E., et al., 2004. A neural network method for identification of RNA-interacting residues in protein. Genome Informatics International Conference on Genome Informatics 15: 105-116.

Jones, S., et al., 2001. Prediction-RNA interactions: A structural analysis. Nucl. Acids Res. 29: 943-954.

Jones, S., et al., 2003. Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. Nucl. Acids Res 31: 7189-7198.

Jones, S., et al., 2003. Using electrostatic potentials to predict DNA-binding sites on DNA-binding proteins. Nucl. Acids Res., 31(24): 7189-7198.

Kasahara, K., et al., 2010. Ligand-binding site prediction of proteins based on known fragment-fragment interactions. Bioinformatics, 26: 1493-1499.

Katchalski-Katzir, E. et al., 1992. Molecular surface recognition: determination of geometric fit between proteins and their ligands by correlation techniques. Proc Natl Acad Sci U S A. 89(6):2195-9.

Kim, H., et al., 2003. Computational analysis of hydrogen bonds in protein-RNA complexes for interaction patterns. FEBS Lett. 552: 231-239.

Kim, O., et al., 2007. Amino acid residue doublet propensity in the protine-RNA interface and its application to RNA interface prediction. Nucl. Acids Res. 34: 6450-6460.

Kinoshita, K. and Nakamura, H., 2005. Identification of the ligand binding sites on the molecular surface of proteins. Protein Sci. 14, 711-718.

Kinoshita, K., et al., 1999. Structural motif of phosphate-binding site common to various protein superfamilies: all-against-all structural comparison of protein-mononucleotide complexes. Protein Engineer 12: 11-14.

Kobayashi, N. and Go, N., 1997. ATP binding proteins with different folds share a common ATP-binding structural motif. Nat. Struct. Biol. 4: 6-7.

Konc, J. and Janezic. D., 2010. ProBis algorithm for detection of structurally similar protein binding sites by local structural alignment. Bioinformatics 26: 1160-1168.

Kono, H. and Sarai, A., 1999. Structure-based prediction of DNA target sites by regulatory proteins. Proteins: Structure, Function, and Genetics 35, 114-131.

Kumar, M., et al., 2007. Prediction of RNA binding sites in a protein using SVM and PSSM profile. Proteins: Structure, Function and Bioinformatics 71: 189-194.

Li, C., et al., 2012. A new residue-nucleotide propensity potential with structural information considered for discriminating protein-RAN docking decoys. Proteins: Structure, Function, and Bioinformatics 80:14-24.

Li, Q., et al., 2010. Improve the prediction of RNA-binding residues using structural neighbours. Protein Peptide Lett., 17: 287-296.

Liu, Z., et al., 2005. Quantitative evaluation of protein-DNA interactions using an optimized knowledge-based potential. Nucl. Acids Res 33: 546-558.

Liu, Z., et al., 2010. Prediction of protein-RNA binding sites by a random forest method with combined features. Bioinformatics 26: 1616-1622.

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

Ma, X., et al., 2011. Prediction of RNA-binding residues in proteins from primary sequence using an enriched random forest model with a novel hybrid feature. Proteins: Structure, Function, and Bioinformatics, 79: 1230-1239.

Maetschke, S.R. and Yuan, Z., 2009. Exploiting structural and topological information to improve prediction of RNA-protein binding sites. BMC Bioinformatics, 10, 341.

Murakami, Y., et al., 2010. PiRaNhA: a server for the computational prediction of RNA-binding residues in protein sequences. Nucl. Acids Res. 38: W412-W416.

Murphy, L. R., et al., 2000. Simplified amino acid alphabets for protein fold recognition and implication for folding. Protein Eng., 13: 149-152.

Perez-Cano, L. and Fernandez-Recio, J., 2010. Optimal protein-RNA area, OPRA: A propensity-based method to identify RNA-binding sites on proteins. Proteins: Structure, Function, and Bioinformatics 78: 25-35.

Perez-Cano, L., et al., 2010. Structural prediction of protein-RNA interaction by computational docking with propensity-based statistical potentials. Pac. Symp. Biocompute., 293-301.

Perez-Cano, L., et al., 2012. A protein-RNA docking benchmark (II): extended set from experimental and homology modeling data. Proteins 80: 1872-82.

Peterson, E. L., et al., 2009. Reduced amino acid alphabets exhibit an improved sensitivity and selectivity in fold assignment. Bioinformatics, 25(11): 1356-1362.

Petsalaki, E., et al., 2009. Accurate Prediction of Peptide Binding Site on Protein Surface. PLoS Comput. Bio., 5, p. e1000335.

Robertson, T.A. and Varani, G., 2007. An all-atom, distance-dependent scoring function for the prediction of protein-DNA interactions from structure. Proteins: Structure, Function, and Bioinformatics 66, 359-374.

Sael, L. and Kihara, D., 2012. Detecting local ligand-binding site similarity in nonhomologous proteins by surface patch comparison. Proteins: Structure, Function, and Bioinformatics 80: 1177-1195.

Samudrala, R. and Moult, J., 1998. An all-atom distance-dependent conditional probability discriminatory function for protein structure prediction. J Mol Biol 275: 895-916.

Schietgat, L., et al., 2013. A polynomial-time maximum common subgraph algorithm for outerplanar graphs and its application to chemoinformatics. Ann Math Artif Intell 69: 343-376.

Shao, X., et al., 2009. Predicting DNA- and RNA-binding proteins from sequences with kernel methods. J. Theor. Biol. 258: 2898-293.

Shazman, S. and Mandel-Gutfreund, Y., 2008. Classifying RNA-binding proteins based on electrostatic properties. PLoS Comput. Biol., 4, e1000146.

Siek, J. G., et al., 2001. The Boost Graph Library: User Guide and Reference Manual. http://www.boost.org/libs/graph.

Spriggs, R. V., et al., 2009. Protein function annotation from sequences: prediction of residues interacting with RNA. Bioinformatics 25: 1492-1497.

Terribilini, M., et al., 2006. Prediction of RNA binding sites in proteins from amino acid sequence. RNA 12: 1450-1462.

Terribilini, M., et al., 2007. RNABindR: a server for analyzing and predicting RNA-binding sites in proteins. Nucl. Acids Res. 35: W578-W584.

The UniProt Consortium, 2010. The Universal Protein Resource (UniProt) in 2010. Nucl. Acids Res 38, D142-D148.

Tian, B., et al., 2004. The double-strand-RNA-binding motif: Interference and much more. Nat. Rev. Mol. Cell Biol. 5: 1013-1023.

Towfic, F., et al., 2010. Struct-NB: predicting protein-RNA binding sites using structural features. Int. J. Data Min. Bioinform. 4: 21-43.

Tuszynska, I. and Bujnicki, J., 2011. DARS-RNP and QUASI-RNP: New statistical potentials for protein-RNA docking. BMC Bioinformatics, 12: 348.

Vakser, I.A. and Aflalo, C., 1994. Hydrophobic docking: a proposed enhancement to molecular recognition techniques. Proteins: Structure, Function, and Bioinformatics, 20: 320-329.

Wang, G. and Dunbrack RL, Jr., 2003. PISCES: a protein sequence culling server. Bioinformatics 19, i1589-i1591.

Wang, L. and Brown, S., 2006. BindN: a web-based tool for different prediction of DNA and RNA binding sites in amino acid sequences. Nucl. Acids Res. 34: W243-W248.

Wang, L., et al., 2010. BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features. BMC Syst. Biol. 4 Suppl 1:S3.

Wang, Y., et al., 2008. PRINTR: Prediction of RNA binding sites in proteins using SVM and profiles. Amino Acids. 35: 295-302.

Wang, Y., et al., 2013. fmcsR: Mismatch Tolerant Maximum Common Substructure Searching in R. Bioinformatics 29: 2792-2794.

Wass, MN., et al., 2010. 3DLigandSite: predicting ligand-binding sites using similar structures. Nucl. Acids Res. 38 W469-W473.

Weathers, E. A., et al., 2004. Reduced amino acid alphabet is sufficient to accurately recognize intrinsically disordered protein. FEBS Lett. 576: 348-352.

Whisstock, J.C. and Lesk, A. M., 2003. Prediction of protein function from protein sequence and structure. Quarterly Reviews of Biophysics. 36(3): 307-40.

Xie, Z.R. and Hwang, M.J., 2012. Ligand-binding site prediction using ligand-interacting and binding site-enriched protein triangles. Bioinformatics, Advance Access.

Xu, B., et al., 2009. An all-atom knowledge-based energy function for protein-DNA threading, docking decoy discrimination, and prediction of transcription-factor binding profiles. Proteins: Structure, Function, and Bioinformatics 76: 718-730.

Yang, X., et al., 2014. PBRDetector: Improved prediction of binding residues on RNA-binding protein structures using complementary feature- and template-based strategies. Proteins: Structure, Function, and Bioinformatics 82: 2455-2471.

Yang, X., et al., 2014. RBRDetector: Improved prediction of binding residues on RNA-binding protein structures using complementary feature- and template-based strategies. Proteins: Structure, Function, and Bioinformatics 82: 2455-2471.

Zhang, C., et al., 2005. A knowledge-based energy function for protein-ligand, protein-protein, and protein-DNA complexes. J Med Chem 48: 2325-2335.

Zhang, T., et al., 2010. Analysis and prediction of RNA-binding residues using sequence, evolutionary conservation, and predicted secondary structure and solvent accessibility. Curr. Protein Pept. Sci. 11(7): 609-628.

Zhao, H., et al., 2010. Structure-based prediction of DNA-binding proteins by structural alignment and a volume-fraction corrected DFIRE-based energy function. Bioinformatics 26: 1857-1863.

Zhao, H., et al., 2011. Highly accurate and high-resolution function prediction of RNA binding proteins by fold recognition and binding affinity prediction. RNA Biol. 8: 988-996.

Zhao, H., et al., 2011. Structure-based prediction of RNA-binding domains and RNA-binding sites and application to structural genomics targets. Nucl. Acids Res. 39: 3017-3025.

Zheng, S., et al., 2007. A knowledge-based potential function predicts the specificity and relative binding energy of RNA-binding proteins. FEBS J., 274: 6378-6391.

Zhou, H. and Zhou, Y., 2002. Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Science 11: 2714-2726.

Zhou, W. and Yan, H., 2010. A discriminatory function for prediction of protein-DNA interactions based on alpha shape modeling. Bioinformatics 26: 2541-2548.