PROTEIN FUNCTION PREDICTION BY INTEGRATING SEQUENCE, STRUCTURE AND BINDING AFFINITY INFORMATION

Huiying Zhao

Submitted to the faculty of the University Graduate School
in partial fulfillment of the requirements
for the degree
Doctor of Philosophy
in the School of Informatics,
Indiana University

August 2013

Accepted by the Faculty of Indiana University, in partial
fulfillment of the requirements for the degree of Doctor of Philosophy.

Yunlong Liu, PhD, Chair

Doctoral Committee

Samy Meroueh, PhD

Sarath Chandra Janga, PhD

May 3, 2013

Yaoqi Zhou, PhD, Advisor

Dedicated to my parents.

Acknowledgements

I would never have been able to finish my dissertation without the guidance of my committee members, help from friends, and support from my family.

I would like to express my deepest gratitude to my advisor, Prof Yaoqi Zhou, for his excellent guidance, caring, patience, and providing me with an excellent atmosphere for doing research.

To my committee members Profs Yunlong Liu, Samy Meroueh, and Sarath Chandra Janga for their encouraging words, thoughtful criticisms, and time and attentions during busy semesters.

To my colleagues for sharing their enthusiasm for and comments on my work: Ms Zhixiu Li, and Drs Jian Zhan, Yuedong Yang, Tuo Zhang, Liang Dai, Eshel Faraggi, Wenchang Xiang, Shesheng Zhang, Beisi Xu, Jihua Wang, Md Tamjdul Hogue, et al.

Finally to my family for their love, support and understanding during the long years of my education.

Huiying Zhao

PROTEIN FUNCTION PREDICTION BY INTEGRATING SEQUENCE, STRUCTURE AND BINDING AFFINITY INFORMATION

Proteins are nano-machines that work inside every living organism. Functional disruption of one or several proteins is the cause for many diseases. However, the functions for most proteins are yet to be annotated because inexpensive sequencing techniques dramatically speed up discovery of new protein sequences (265 million and counting) and experimental examinations of every protein in all its possible functional categories are simply impractical. Thus, it is necessary to develop computational function-prediction tools that complement and guide experimental studies. In this study, we developed a series of predictors for highly accurate prediction of proteins with DNA-binding, RNA-binding and carbohydrate-binding capability. These predictors are a template-based technique that combines sequence and structural information with predicted binding affinity. Both sequence and structure-based approaches were developed. Results indicate the importance of binding affinity prediction for improving sensitivity and precision of function prediction. Application of these methods to the human genome and structure genome targets demonstrated its usefulness in annotating proteins of unknown functions and discovering moon-lighting proteins with DNA, RNA, or carbohydrate binding function. In addition, we also investigated disruption of protein functions by naturally occurring genetic variations due to insertions and deletions (INDELS). We found that protein structures are the most critical features in recognising disease-causing non-frame shifting INDELs. The predictors for function predictions are available at http://sparks-lab.org/spot, and the predictor for classification of non-frame shifting INDELs is available at http://sparks-lab.org/ddig.

Yunlong Liu, PhD, Chair

Contents

List of Tables

List of Figures

Chapter 1 Introduction

1.1 Proteins and their functions

1.1.1 Proteins

Proteins are large biological molecules consisting of amino acids. They play a vast array of functions within living organisms. Proteins are different from each other by their sequences and three-dimensional structural properties. A protein sequence is a series of letters that describe the amino acid composition of protein. Currently, there are two major direct methods, mass spectrometry [1] and Edman degradation [2], for determination of protein sequences. It is also possible to utilize next generation sequencing technique to obtain the DNA/mRNA sequence that codes the protein sequence.

Proteins perform their functions with help of their molecular structures. Protein structures can be divided into four levels: primary structure, secondary structure, tertiary structure and quaternary structure. Primary structure refers to linear amino-acid sequence of the polypeptide chain. The primary structure is held together by covalent peptide bonds, which are formed during the process of protein biosynthesis or translation. Protein secondary structure refers to regular protein backbone sub-structure. There are three main types of secondary structures: alpha helix, beta strand, and coil [3]. Both alpha helix and beta sheet represent conformations that connect hydrogen bond donors with acceptors in the peptide backbone. Tertiary structure refers to three-dimensional (3D) structure of a single protein molecule. The 3D structure of a protein is formed by protein folding process. In this process, a polypeptide folds into its characteristic and functional 3D structures from a random coil. The folding process is driven by non-specific hydrophobic interactions and hydrogen bonds. During protein folding, protein structure becomes stable when the structure reaches global minimum of free energy. Quaternary structure is made of multiple subunits of 3D structures. Protein structures are often referred as structural domains to distinguish

from intrinsically disordered regions. A structure domain is an element of the overall structure of a protein. Protein domains can evolve, function and exist independently of the rest part of the protein. One protein may contain several domains, and each domain can perform multiple functions.

1.1.2 Protein function through binding

Proteins are one of the most important molecular machines in the living organism. Proteins contain half the dry weight of an Esherichia coli cell [4]. Most of the biological processes are related with protein activity. Protein functions include enzyme catalysis, interaction with other molecules, supporting materials, etc. Among these functions, the interaction with other molecules are contributed by their ability to bind with molecule partners. The residues in a protein that bind with other molecule are called as binding sites. The binding ability of a protein is mainly determined by the binding sites on protein surface [5].

Proteins can bind to DNA and form protein-DNA complexes (DBP) [6]. These proteins are composed of DNA-binding domains and have binding affinity for either single or double stranded DNA. DNA-binding proteins play essential roles in transcription, regulation, replication, packaging repair and rearrangement. For example, transcription factors modulate the process of transcription; nucleases cleave DNA molecules; and histones are involved in chromosome packaging and transcription in the cell nucleus.

RNA-binding proteins (RBP) are another class of important proteins through binding to RNA in cells and forming ribonucleoprotein complexes. RNA-binding proteins are important in translation regulation and post-transcriptional processing of pre-mRNA including RNA splicing, editing and polyadenylation. They play critical roles in the biogenesis, stability, transport and cellular localization [7, 8]. RNA-binding proteins can specifically recognize their RNA targets by complementary shapes. Three most widely studied RNA-binding domains include double-stranded RNA-binding motif (dsRBM), RNA-recognition motif (RRM) and zinc fingers.

Carbohydrate-binding proteins (CBPs) are functional proteins that recognize cell-surface carbohydrates. CBPs are important for immune systems. For example, viruses can use carbohydrates to attach themselves to the host cell during infection. On the other hand, host CBPs can also recognize these carbohydrates and prevent virus invasion. Therefore, CBPs have been employed as potential drug targets in pathogens.

Proteins can also bind to other partners. For example, iron-binding proteins are important in metabolism. Their binding with iron can inhibit microbial growth. Furthermore, proteins can bind to other proteins to regulate enzymatic activity, control progression through the cell cycle and allow the assembly of large protein complexes.

1.2 Annotation of protein functions

1.2.1 Experimental approaches for detection of protein functions

There are many studies to detect protein-DNA interaction experimentally. Recent strategies relied on sophisticated mass spectrometry technologies. Washburn and Fournier published their work on identification of DBPs by pulldown experiments in conjunction with multi-dimensional protein identification technology (MudPIT) [9, 10]. Other standard methods include EMSA, DNAasa I footprinting, exonuclease III footprinting, southwestern blotting and others [11]. However, experimental approaches face many challenges. For example, both EMSA and DNase I footprinting methods are usually combined together to improve experimental accuracy [12]. Unfortunately, many DNA-binding proteins can only be detected by one type of assay. Thus, the detection is not guaranteed for those proteins which can only be recognized by one assay.

Similar to identifications of DBPs, most frequently used methods for RBPs are protein microarray [13] and mass spectrometry [14, 15]. Protein microarray and RNA probes have been used to identify a limited number of RBPs. As an alternative to in vitro approaches, stable isotope labeling by amino acids in cell culture and mass spectrometry were applied to identify the interaction between protein and RNA [16]. More recently, a fluorescence-based quantitative method has been developed to monitor mRNA-protein interactions, and 300 new RDPs were uncovered [17].

For experimentally detecting CBPs, there are three most commonly used approaches: X-ray [18], NMR [19, 20] and fluorescence spectroscopy [19] studies [21].

1.2.2 Computational approaches for prediction of protein functions

While experimental techniques for determining protein functions are less likely to produce false positives, they are time consuming and expensive. More importantly, the number of protein sequences are exponentially increasing with the development of next generation sequencing technology. There is a widening gap between the number of proteins with annotated functions and the number of protein with known sequences. Meanwhile, the structure genome project generated a large number of structures without known function. Therefore, it is necessary to develop effective computational approaches for predicting protein functions from their structures or sequences.

Historically, commonly used approaches for prediction of protein functions are based on sequence/structure homology [22–26]. The assumption is that similar sequence/structure encodes similar function. However, this assumption is only partially true for highly homologous proteins, while most proteins don't have homologous proteins with known functions. Thus, it is necessary to develop an alternative approach for more sensitive protein function detection.

Currently, the most widely-used methods for prediction of protein functions are machine-learning based methods, which usually employ sequence or structure features of proteins to train classifiers for protein function prediction. For example, several sequence-based classifiers for DBP/RBP prediction were based on support-vector machine (SVM) [27, 28]. Common features in these predictors include amino acid composition, solvent accessible surface, hydrophobicity, conjoint triad [29], position specific scoring matrices (PSSM), and interface propensities [30]. There is only one published method for prediction of CBPs from sequence. This method employed sequence patterns and frequencies of three neighboring amino acids as input features for SVM.

Although machine learning-based methods have achieved reasonable accuracies in prediction of protein functions, they have several limitations. First, their performance decrease significantly when they are applied to real large scale database because the methods are typically trained on datasets with a small, equal number of positive and negative cases. Furthermore, machine-learning based methods can only provide binary prediction without information of 3D complex structures. Methods for predicting binding sites are separate from those methods for predicting functions. A more recent approach is to utilize protein template structure. Such template-based methods perform structure comparison to determine target function. For targets having sequence information only, structure prediction tools were employed. For each structurally similar template protein, a model complex structure can be generated by modeling the target protein structure (template-based predicted structure in absence of experimental structure) and its binding partner from the template complex. For these model complex structures, binding affinity will be predicted, and only those having high binding affinity will be kept. Thus, a template-based method considers not only the structural similarity but also the interaction strength between the target protein and its potential binding partner. Moreover, the template-based method is able to predict binding residues and complex structures in addition to binary function prediction.

1.3 Prediction of protein functions by a template-based method

The first template-based method was developed for predicting DNA-binding proteins [31] from structure. This method was later improved by replacing the contact-based energy function to DDNA3 [32], a more accurate all-atom, DFIRE [33] -derived energy function. This approach was extended to the prediction of RNA-binding proteins from structure [34]. In addition, the template-based method using sequence only has also been developed. In this method, the target structure was predicted by recognizing correct structural templates from proteins with known structures in PDB. The confidence of prediction was evaluated by sequence to structure matching Z-score

[35, 36]. Several techniques utilized by the template-based approaches are described as following.

1.3.1 Structure comparison

Structure comparison is a useful method for detecting proteins with similar functions in the absence of sequence similarity. Different from sequence comparison, structure comparison employs structure alignment and attempts to establish the homology between two protein structures from their shapes and 3D conformations. This procedure relies on protein tertiary structures. Structure alignment is useful for prediction of protein functions because protein structures are more conserved than their sequences [37], and many proteins with similar functions may converge to similar structure during evolution. Therefore, structure alignment has been an active research area for more than 30 years. Currently, there are more than 50 published computational methods [38, 39].

Critical difference between various structure alignment methods is the scoring function that measures structural similarity. Structure similarity is often evaluated by root-mean-square deviation (RMSD). The RMSD between two aligned structures indicates their divergence from one another. However, RMSD is strongly dependent on protein size and radius of gyration, and very sensitive to poorly aligned local regions [40]. Zhang and Skolnick developed TM-score to remove the dependence of structure similarity score on protein sizes, and later applied to structure alignment [41]. The score is based on LG-score with an empirical size-dependent d_0 $[= 1.24(L-13)^{1/3} - 1.8]$. However, this score assumes that proteins are globular and aligned in a predetermined size L.

To further remove the size dependence, SP-align was developed by us [42]. This method was proposed by introducing an effective alignment length that avoids the need to pre-specify a length for normalization. The function is defined as

$$\mathrm{SP-score} = \frac{1}{3L^{1-\alpha}} \mathrm{Max} \left[\sum_{r_{ij}<2d_0} \left(\frac{1}{1 + r_{ij}^2/d_0^2} - 0.2 \right) \right] \tag{1.1}$$

, where d_{ij} is the distance between C_α atoms of two aligned residues, d_0 was chosen 4.0 Å somewhat in between 3.5 Å in MaxSub and 5 Å in LG score, α is a to-be-determined parameter for removing the dependence on protein length L, a constant of 0.2 is used for a smooth cutoff for SP-score at $d_{ij} = 2d_0$, and a factor of 1/3 is used to scale the threshold for fold discrimination to around 0.5. The new score (SP-score) with its alignment method (SP-align) was tested in structure classification and prediction of nucleic-acid binding proteins with comparison to several established methods: DALI, CE, and TMalign. The comparison indicates that SP-align consistently improves over other methods.

1.3.2 Structure prediction

Structure prediction attempts to predict protein structure from a given query sequence. The most reliable structure-prediction technique is to match with existing known structure templates. Such template-based modeling becomes increasingly powerful because most popular structural folds are known [43,44]. However, it is still challenging to recognize structurally similar templates as revealed from the critical assessment of structure prediction (CASP). Past CASP experiments highlighted the importance of post treatment of models predicted by individual fold-recognition methods through the use of consensus predictions. Recently developed new methods include combining fragment and template comparison [45], utilizing non-linear scoring function from conditional random field model and profile entropy [46], employing predicted torsion angles and combined use of profile-profile alignment and pairwise and solvation potentials [47,48].

One common issue in the above methods is that matching predicted 1D profiles of query sequence with actual profiles of templates is based on simple matrices, without accounting for the probability of errors in predicted 1D structural properties. SPARKS-X [49] introduced energy terms based on estimating the matching probability between target and template. This method also takes advantage of recently improved torsion angle predictor, SPINE-X [50] in prediction of secondary structure. The

matching score calculation of SPARKS-X was described as Eq. 1.2.

$$\begin{aligned} S(i,j) = \; & -\tfrac{1}{200}[F^{\mathrm{seq}}_{\mathrm{query}}(i) \cdot M^{\mathrm{seq}}_{\mathrm{template}}(j) + F^{\mathrm{seq}}_{\mathrm{template}}(j) \cdot M^{\mathrm{seq}}_{\mathrm{query}}(i)] \\ & + w_1 E(SS_t(i)|SS_q(j), C_{SS,q}(j)) \\ & + \textstyle\sum_{k=2}^{4} w_k E(\Delta^k_{ij}|C_{k,q}(j)) + s_{\mathrm{shift}}. \end{aligned} \tag{1.2}$$

with weight parameters (w_k) and a constant shift s_{shift}. The first term in Eq. (1.2) is the profile-profile comparison between the sequence profile from the query sequence, $M^{\mathrm{seq}}_{\mathrm{template}}(j)$ and $M^{\mathrm{seq}}_{\mathrm{query}}(i)$ are the sequence-derived log-odd profile of the template sequence and that of query sequence, respectively. These sequence profiles are constructed by three iterations of PSIBLAST searching (E value cutoff of 0.001) against non-redundant (NR) sequence database, which was filtered to remove low-complexity regions, transmembrane regions, and coiled-coil segments. The second term in Eq. (1.2) measures the difference between the predicted secondary structure and the actual secondary structure of the template. The third term in Eq. (1.2) measures the difference Δ^k_{ij} between two other predicted 1D structural properties of the query sequence and the actual properties of the template [real-value torsion angles (ϕ/ψ) and real-value solvent accessibility].

SPARKS-X was tested on several benchmarks and compared to other automatic servers. All the results indicate that SPARKS-X is one of the best single-method fold-recognition servers. Given the robust performance of SPARKS-X, it was employed as a structure prediction tool for predicting protein functions.

1.3.3 Energy function for calculation of Binding affinity

An energy function describes physical interactions between a protein and its binding partner. A knowledge-based energy function is obtained from statistical analysis of structures. Different knowledge-based energy functions are mainly different from their definitions of a reference state. The DFIRE energy function (Eq. 2.1) defines the reference state based on ideal gas mixture (r^{α}) with $\alpha < 2$ to account for the

finite-size effect [33]. Several knowledge-based energy functions were developed for protein-DNA interactions. For example, a residue base-level energy function was proposed to calculate the protein-DNA interaction [51]; atom-level energy functions were developed by extending the DFIRE to protein-DNA binding affinity calculation [52]. The DFIRE energy function was further improved by adding a volume fraction correction [32, 53]. Similarly, an energy function for protein-RNA interaction [34, 36] and protein-carbohydrate interaction (In preparation) were derived. A DFIRE-based potential satisfies the following equation:

$$\overline{u}_{i,j}^{\mathrm{DFIRE}}(r) = \begin{cases} -RT \ln \frac{N_{obs}(i,j,r)}{(\frac{r}{r_{cut}})^{\alpha}(\frac{\Delta r}{\Delta r_{cut}})N_{obs}(i,j,r_{cut})}, & r < r_{cut}, \\ 0, & r \geq r_{cut}, \end{cases} \tag{1.3}$$

where R is the gas constant, $T = 300K$, $\alpha = 1.61$, $N_{obs}(i, j, r)$ is the number of ij pairs within the spherical shell at distance r observed in a given structure database, r_{cut} is the cutoff distance, Δr_{cut} is the bin width at r_{cut}. The value of $\alpha(1.61)$ was determined by the best fit of r^{α} to the actual distance-dependent number of ideal-gas points in finite protein-size spheres.

1.4 Overview of the dissertation

As described above, a template-based approach is a powerful and reliable approach for prediction of protein functions. This dissertation mainly focuses on development of template-based approaches for prediction of DNA-binding proteins, RNA-binding proteins, and carbohydrate-binding proteins. How to fully utilize protein structural information is a critical point for template-based approaches. In addition to protein function prediction, we also predict function disruption due to insertions and deletions of bases in the human genome.

This dissertation can be divided into four parts. The first part is prediction of DNA-binding proteins based on structures (chapter 2) and sequences (chapter 3). The second part contains four chapters that includes the prediction of RBPs from structure (chapter 4) and sequence (chapter 5), application of sequence-based prediction

method of RBPs to the human genome (chapter 6), and the review of current status of RBPs prediction from low to the highest resolution (chapter 7). The third part is the prediction of CBPs from their structures (chapter 8). The final part is the classification of disease-related non-frame shifting insertion/deletions of bases in the human genome (chapter 9).

Chapter 2 Structure-based prediction of DNA-binding proteins by structural alignment and a volume-fraction corrected DFIRE-based energy function

Abstract

Motivation: Template-based prediction of DNA-binding proteins requires not only structural similarity between target and template structures but also prediction of binding affinity between the target and DNA to ensure binding. Here, we propose to predict protein-DNA binding affinity by introducing a new volume-fraction correction to a statistical energy function based on a distance-scaled finite ideal-gas reference state (DFIRE).

Results: We showed that this energy function together with the structural alignment program TM-align achieves the Matthews correlation coefficient (MCC) of 0.76 with an accuracy of 98%, a precision of 93%, and a sensitivity of 64%, for predicting DNA binding proteins in a benchmark of 179 DNA-binding proteins and 3797 non-binding proteins. The MCC value is substantially higher than the best MCC value of 0.69 given by previous methods. Application of this method to 2235 structural genomics targets uncovered 37 as DNA-binding proteins, 27(73%) of which are putatively DNA-binding and only 1 (3%) protein whose annotated functions do not contain DNA-binding while the remaining proteins have unknown function. The method provides a highly accurate and sensititive technique for structure-based prediction of DNA-binding proteins.

Availability: The method is a port of the SPOT (Structure-based function -Prediction On-line Tools) package available at http://sparks-lab.org/spot

2.1 Introduction

DNA-binding proteins are proteins that make specific binding to either single or double stranded DNA. They play an essential role in transcription regulation, replication,

packaging, repair and rearrangement. With completion of many genome projects and many more in progress, more and more proteins are discovered with unknown function [54]. The structures for some of those function-unknown proteins are solved because of structural genomics projects [55]. Functional annotations of these proteins are particularly challenging because the goal of structural genomics is to cover the sequence space of proteins so that homology modeling becomes a reliable tool for structure prediction of any proteins and, thus, many targets in structural genomics have low sequence identity to the proteins with known function. Therefore it is necessary to develop computational tools that utilize not only sequence but also structural information for function prediction [25, 31, 56–59].

Many methods have been developed for structure-based prediction of DNA-binding proteins. These include function prediction through homology comparison and structural comparison [22–26, 60]. Others explore sequence and structural features of DNA-binding and non-binding proteins with sophisticated machine-learning methods such as neural network [56, 61–63], logistic regression [64], and support vector machines [22, 27, 63, 65, 66].

Recently, Gao and Skolnick proposed a new two-step approach, called DBD-Hunter [31], for structure-based prediction of DNA-binding proteins. In DBD-Hunter, the structure of a target protein is first structurally aligned to known protein-DNA complexes and the aligned complex structures are used to build the complex structures between DNA and the target protein. The predicted complex structures are, then, employed for judging DNA binding or not by structural similarity scores (TM-Score) and predicted protein-DNA binding affinities. TM-align [52] and a contact-based statistical energy function are employed in the first and second steps of DBD-Hunter, respectively. DBD-Hunter is found to substantially improve over the methods based on sequence comparison only (PSI-BLAST), structural alignment only (TM-align), and a logistic regression technique [67].

In this study, we investigate if one can further improve the prediction of DNA-binding proteins by employing a different statistical energy function for

predicting binding affinity. Our knowledge-based energy function is distance-dependent and built on a distance-scaled finite ideal gas reference (DFIRE) state originally developed for proteins [33, 68, 69] and extended to protein-DNA interactions [52, 53]. Here, we introduce a new volume-fraction correction for the DFIRE energy function in extracting protein-DNA statistical energy function from protein-DNA complex structures. This volume fraction correction term, unlike previously introduced one [53], is atom-type dependent to better account for the fact that protein and DNA atom types are unmixable and occupy in physically separated volumes. In addition to introduction of a new energy function, we further optimize protein-DNA binding affinity by performing DNA mutation. These two techniques lead to a highly accurate and sensitive tool for structure-based prediction of DNA-binding proteins.

2.2 Methods

2.2.1 Datasets

We employed the datasets compiled by Gao and Skolnick [31]. One positive and one negative datasets for training are 179 DNA-binding proteins (DB179) and 3797 non DNA-binding proteins (NB3797), respectively. These structures were obtained based on 35% sequence identity cutoff, a resolution of 3Å or better, a minimum length of 40 residues for proteins, 6 base pairs for DNA, and 5 residues interacting with DNA (within 4.5Å of the DNA molecule). As in [31], we use significantly larger number of non DNA-binding proteins in order to reduce false positive rate because DNA-binding proteins are only small fraction of all proteins. APO and HOLO testing datasets are made of 104 DNA-binding proteins whose structures are determined in the absence and presence of DNA, respectively. A maximum of 35% sequence identity was also employed in selecting these 104 proteins. For APO/HOLO datasets, 93 APO-DB179 pairs and 92 HOLO-DB179 pairs have sequence identity $>$35%. These pairs are excluded from target-template pairs during testing.. An additional test set of 1697 proteins (the SG1697 set) was compiled from structural genome targets with a sequence identity cutoff at 90% by Gao and Skolnick from the Jan 2008 PDB release. We further

updated the release on November 2009 and obtained 2235 chains(the SG2235 set). This was done by queried "structural genomic" words in the PDB databank, resulting in 2447 PDB entries. These PDB entries were divided into protein chains and clustered by the CD-HIT [70]. For the clusters that contain a protein chain in SG1679, we chose the protein chain as the representation. For other clusters, we randomly chose one protein chain. There are 538 additional proteins and a total of 2235 protein chains.

To provide an additional test set and examine the effect of a larger database of DNA-binding proteins, we have also updated DNA-binding proteins from DB179 to DB250. This updated data set of DNA-binding proteins is selected from PDB released on December 2009 based on the same criteria that produced DB179. After removing the chains with high sequence identity (>35%) with any chain contained in DB179 and with each other, we obtained 71 additional protein-DNA complexes. This leads to an additional test dataset DB71 and an expanded training set DB250 (DB179+DB71).

2.2.2 Knowledge-based energy function

We employ a knowledge-based energy function to predict the binding affinity of a protein-DNA complex. We have developed a knowledge-based energy function for proteins based on the distance-scaled finite ideal-gas reference state (DFIRE) that satisfies the following equation [33]:

$$\overline{u}_{i,j}^{\mathrm{DFIRE}}(r) = \begin{cases} -RT \ln \frac{N_{obs}(i,j,r)}{(\frac{r}{r_{cut}})^{\alpha}(\frac{\Delta r}{\Delta r_{cut}})N_{obs}(i,j,r_{cut})}, & r < r_{cut}, \\ 0, & r \geq r_{cut}, \end{cases} \tag{2.1}$$

where R is the gas constant, $T = 300K$, $\alpha = 1.61$, $N_{obs}(i,j,r)$ is the number of ij pairs within the spherical shell at distance r observed in a given structure database, r_{cut} is the cutoff distance, Δr_{cut} is the bin width at r_{cut}. The value of $\alpha(1.61)$ was determined by the best fit of r^{α} to the actual distance-dependent number of ideal-gas points in finite protein-size spheres.

Eq. (2.1) for proteins was initially applied to protein-DNA interactions unmodified with 19 atom types for both proteins and DNA (DDNA) [52]. In DDNA2

[53], a low count correction is made to$N_{obs}(i,j,r)$:

$$N_{obs}^{lc}(i,j,r) = N_{obs}(i,j,r) + \frac{75\sum_{i,j} N_{ij}^{Protein-DNA}(r)}{\sum_{i,j,r} N_{ij}^{Protein-DNA}(r)} \tag{2.2}$$

In addition, we employed residue/base specific atom types with a distance-dependent volume-fraction correction defined as $f^v(r) = \frac{\sum_{i,j} N_{ij}^{\mathrm{Protein-DNA}}(r)}{\sum_{i,j} N_{ij}^{\mathrm{All}}(r)}$. This volume fraction correction was made to take into account the fact that DNA and protein atoms with residue/base specific atom types do not mix with each other. However, we found that DDNA2 is unable to go beyond existing techniques for predicting DNA-binding proteins. To further improve DDNA2, we introduce atom-type dependent volume fractions: $f_i^v(r) = \frac{\sum_j N_{ij}^{\mathrm{Protein-DNA}}(r)}{\sum_j N_{ij}^{\mathrm{All}}(r)}$. Our final equation for the statistical energy function is

$$\overline{u}_{i,j}^{\mathrm{DDNA3}}(r) = \begin{cases} -\eta \ln \frac{N_{obs}(i,j,r)}{\left(\frac{f_i^v(r) f_j^v(r)}{f_i^v(r_{cut}) f_j^v(r_{cut})}\right)^{\beta} \frac{r^{\alpha}\Delta r}{r_{cut}^{\alpha}\Delta r_{cut}} N_{obs}^{lc}(i,j,r_{cut})}, & r < r_{cut}, \\ 0, & r \geq r_{cut}, \end{cases} \tag{2.3}$$

where we have introduced a parameter β. Physically, β should be around 1/2 so that volume fraction is counted once. We will employ it as an adjustable parameter here for the same reason that makes α less than 2: proteins are finite in size. As in DDNA2, we will use residue/base specific atom types (167 atom types for proteins and 82 for DNA) and r_{cut}=15Å, Δr=0.5Å. We also set the factor η arbitrarily to 0.01 to control the magnitude of the energy score. For convenience, we shall label the volume-fraction corrected DFIRE as DDNA3.

2.2.3 Training of the method for predicting DNA-binding proteins

DB179 is used to generate the DDNA3 statistical energy function Eq. (2.3). To avoid overfiting, we employed the leave-one-out scheme to train DDNA3 statistical energy function. A target protein is chosen from DB179/NB3797. The TM-align program is employed to make a structural alignment between this target protein with a protein in DB179 (except itself if it is in DB179). If the alignment score (TM-score) is

greater than a threshold, the proposed complex structure between the target protein and DNA is obtained by replacing the template protein from its protein-DNA complex structure. The binding affinity between DNA and the target protein is evaluated by the DDNA3 energy function Eq. (2.3). Instead of using template DNA sequences, we perform exhaustive mutations of DNA base pairs to search for the highest binding affinity. DNA bases are paired by X3DNA software package [71]. The conformation of mutated bases are built using default bond length, bond angle and dihedral angle parameters as defined in AMBER98 forcefield [72]. A DNA base, if does not have a corresponding pairing base, is not mutated. If the highest binding affinity is greater than an optimized threshold, the target protein is considered as a DNA binding protein. The method described above has two important differences from DBD-hunter: the use of our distance-dependent energy function and the search for the strongest binding DNA fragment.

2.2.4 Evaluation of the method for predicting DNA-binding proteins

The measures of the method performance are: Sensitivity [SN=TP/(TP+FN)], Specificity [SP=TN/(TN+FP)], Accuracy [AC=(TP+TN)/(TP+FN+TN+FP)], and Precision [PR=TP/(TP+FP)]. In addition, we employed a Matthews correlation coefficient:

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \tag{2.4}$$

Here TP, TN, FP, and FN refer to true positives, true negatives, false positives, and false negatives, respectively.

2.3 Results

2.3.1 Training based on DB179/NB3797 (DDNA3)

We have optimized volume-fraction exponent β, TM-score and binding affinity thresholds to achieve the highest MCC values. Optimization is performed by a

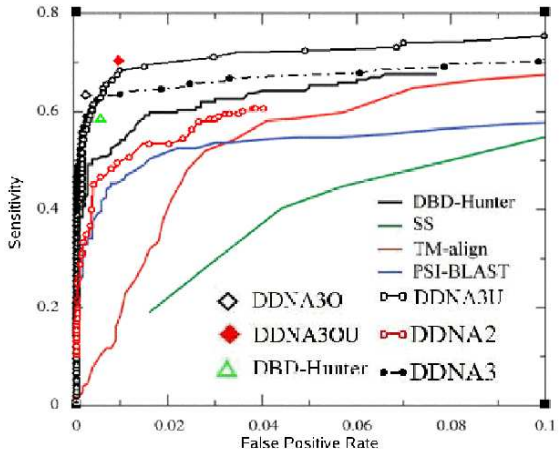

Fig. 2.1: Sensitivity versus false positive rate, given by DDNA3 (Filled black circles) and DDNA2 (Open red circles) reveals the importance of an appropriate reference state for method performance in predicting DNA binding proteins. The results of other methods are adapted from [31]. DDNA3U (open black circles) is the sensitivity versus false positive rate given by DDNA3 based on updated DB250 dataset. TM-Score dependent energy-score thresholds lead to DDNA3O (Open Diamond) and DDNA3OU (Red filled diamond), compared to optimized DBD-Hunter (Open green triangle).

grid-based search. The grids for β and TM-score are 0.02 and 0.01, respectively. For the binding affinity threshold, the lowest energy of each aligned complex under different TM-score thresholds is calculated and these energy values are considered sequentially as the energy threshold. We found that the highest MCC is 0.73 for β=0.4, the structural similarity threshold of 0.60 and the energy threshold of -11.6. The corresponding accuracy, precision and sensitivity are 98%, 91%, and 60%, respectively. The effect of a knowledge-based energy function can be revealed by replacing DDNA3 with DDNA2. The optimized MCC value (Structural similarity threshold of 0.53 and energy threshold of -4.2) is 0.61. (Note, there is no β parameter in DDNA2.) The corresponding accuracy, precision, and sensitivity are 97%, 85%, and 55%, respectively. It is clear that the reference state of a statistical energy function has a significant impact on the performance in predicting DNA-binding proteins. The largest improvement is 6% improvement in precision, the fraction of correct prediction in all prediction. The overall performance of DDNA3 significantly improves over that of DBD-Hunter which has a MCC of 0.64, 98% accuracy, 84% precision and 55% sensitivity, respectively.

Table 2.1: Optimized TM-score-dependent energy thresholds based on DB179 and NB3797 (DDNA3O)

TM-score Range	Energy Threshold	ΔTP	TP	ΔFP	FP	Max MCC
0.74-1.00	-9.87	53	53	3	3	0.52
0.62-0.74	-13.95	52	105	4	7	0.73
0.58-0.62	-16.50	3	108	1	8	0.74
0.55-0.58	-18.64	4	112	0	8	0.76
0.52-0.55	-29.10	2	114	0	8	0.76

Fig. 2.1 shows sensitivity as a function of false positive rate. Our results were obtained by fixing structural similarity threshold and varying the energy threshold. It is clear that DDNA3 yields a substantially higher sensitivity than either DDNA2 or DBD-Hunter for a given false positive rate.

The predicted binding complexes can be employed to examine predicted DNA binding residues. An amino-acid residue is considered as a DNA-binding residue if any heavy atom of that residue is less than 4.5Å away from any heavy atom of a DNA base. Predicted binding residues from template-based modeling can be compared to actual binding residues. For the training set (179 DB and 3797 NB proteins), there are 108 predicted DB proteins with 11 false positives. For these 108 predicted complexes, specificity, accuracy, precision, sensitivity and MCC of predicting DNA binding residues are 94%, 89%, 74%, 68%, and 0.64, respectively. For a comparison, DDNA2 has predicted 99 DB proteins and the corresponding performance in predicting DNA binding residues are 93%, 88%, 75%, 67%, and 0.63, respectively. These performances are similar to a specificity of 93%, an accuracy of 90%, a precision of 71% and a sensitivity of 72% achieved by DBD-hunter. Similar performance in predicting DNA-binding residues is due to the same structural alignment (TM-align) method used in the first step by the three methods.

2.3.2 TM-Score dependent energy threshold (DDNA3O)

Obviously, one threshold for energy and one for structural similarity (TM-Score) are too simple to capture the complex relation between structure and binding. For

example, one expects that the binding-energy requirement should be stronger for less similar structures but weaker for highly similar structures between template and query. This has led Gao and Skolnick to develop TM-Score dependent energy thresholds (9 energy thresholds for 9 TM-Score bins ranging from 0.40 to 1.0 to maximize MCC value in each bin), and they finally set a minimum TM-score cutoff at 0.55 for maximum MCC. Here, we slightly changed the way to calculate MCC by including those predicted positive(TP/FP) in higher TM-score region. The results are shown in Table 2.1. By this way, the cutoff of TM-score is extended to 0.52 rather than 0.55 as Gao's way, and the number of TP increase 2 without increasing FP. We followed their method and optimized 9 parameters for the MCC value at each TM-Score bin separately for the same dataset (DB179 and NB3797). We further found that the top four bins in the table with negative prediction for TM-score$<$0.55 generate the highest MCC value of 0.76 for the entire dataset. To distinguish this further optimized method, we labeled it as DDNA3O. DDNA3O yields a MCC value of 0.76 with the corresponding sensitivity of 0.64 and specificity of 0.998. By comparison, the corresponding optimized DBD-Hunter with the same dataset has a MCC value of 0.69 with the corresponding sensitivity of 0.58 and specificity of 0.995 while the DDNA3 has a MCC value of 0.73 with sensitivity of 0.60 and specificity of 0.997. Thus, most significant improvement from DDNA3 to DDNA3O is significant increase in sensitivity (from 60% to 64%) also with reduction in rate of false positives (from 11/3797 to 8/3797).

There are 114 complexes predicted as DNA-binding proteins by DDNA3O. For these 114 complexes, predicted DNA-binding residues are compared with native complexes. The specificity, accuracy, precision, sensitivity and MCC are 95%, 90%, 77%, 69% and 0.67, respectively. These do not change significantly from DDNA3 because of same complex structures generated by TM-align. The slight difference is caused by 2 reasons. First, in different potential energy functions, different proteins are predicted as binding; Secondly, protein may choose different templates.

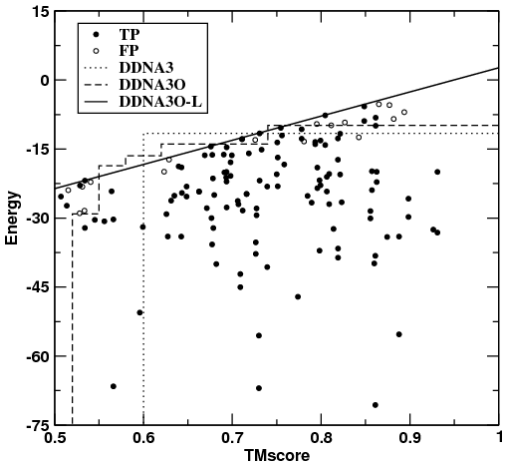

Fig. 2.2: Energy threhold versus TM-score, given by DDNA3O-L(filled line) and DDNA3O (slashed line). All protein located behind the line is predicted as positive. Only TP(filled circles) and FP (open circles) by DDNA3O-L are shown. For protein with multiple matching templates, only template with highest TM-score is used.

We found the energy threshold is increasing along with TM-score threshold. To show the relation between energy and TM-score, we changed to a new way to optimize the energy threshold by linear relation with TM-score $E_{cut} = \gamma \cdot TMscore + e0$, where γ and $e0$ are two parameters for training to maximize MCC. The highest MCC is 0.76 when $\gamma = 52.5$ and $e0 = -49.85$ with the TM-score cutoff at 0.5, where there is higher sensitity 67%(120/179) but also with more number of false positive (17). This method is labeled as DDNA3O-L. As shown in Fig. 2.2, most of true positive points by this method are far below the boundary, with a few left mixed with false positive points. Relatively all false positive positive points are gathering around the boundary. Certaily, a high-order equation can discriminate the points better, however, limited to the number of samples, it's hard to overcome the over-training problem. Also DDNA3 and DDNA3O gives a reasonable boundary. To limit the rate of false positive in the prediction, we will still use DDNA3O for all future applications.

2.3.3 Test by the APO104/HOLO104 datasets

The methods trained above (DDNA3 and DDNA3O) are applied to predict DNA binding proteins of APO104/HOLO104 datasets. The numbers of positive prediction are 50 by DDNA3 and 53 by DDNA3O (out of 104) for the APO sets, and 61 by DDNA3 and 62 by DDNA3O (out of 104) for the HOLO sets, respectively. That is, using monomer structures, rather than the complex structures, leads to a reduction of

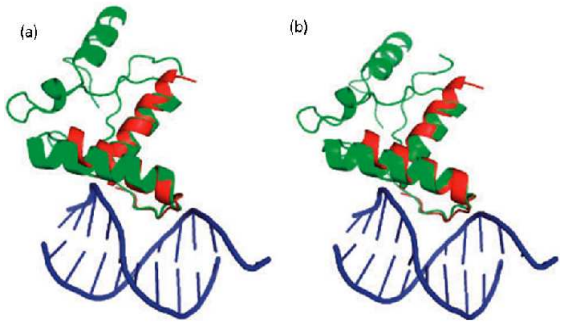

Fig. 2.3: (a) Structural comparison between APO target protein 1mjkA (green) and template protein 1ea4A(red) . The TM-score between them is 0.79 and the interaction energy between 1mjkA and template DNA is -20.9. (b) Structural comparison between HOLO target protein 1mjmA(green) and template protein (1ea4A). The TM-score between them is 0.76 and the interaction energy between 1mjmA and template DNA is -20.6.

11% in sensitivity (from 59% for the HOLO to 48% for the APO set) by DDNA3 and 9% by DDNA3O (from 60% to 51%). The corresponding sensitivity values for DDNA2 are 43.3% (45/104) and 53.8% (56/104) for the APO and HOLO sets, respectively. The performance of DBD-Hunter (47% for the APO and 55% for the HOLO sets) is somewhat in between DDNA2 and DDNA3. The test confirms a significant increase in sensitivity by DDNA3O over by DDNA3 for the APO set, in particular.

A more detailed analysis on predictions made by DDNA3O shows that there is an overlap of 49 predictions between the APO and HOLO sets. Fig. 2.3 shows one example of the test on target proteins 1mjkA (contained in APO104) and 1mjmA (contained in HOLO104). 1mjkA and 1mjmA are the structure of the same methionine repressor protein in the absence and presence of DNA fragment, respectively. There is a small conformational change before and after DNA binding (TM-Score between the two is 0.93). This small conformational change apparently does not prohibit the successful match to the same template protein 1ea4A with strong binding affinity.

On the other hand, there are 12 correctly predicted HOLO targets but incorrectly predicted APO targets as shown in Table 2.2. The difference is caused by significant local conformational change in binding regions (high TM-align score but low binding affinity). An example (1le8A in HOLO and corresponding 1f43A in APO) is shown in Fig. 2.4a where significant change in binding regions (from red in APO to green

Table 2.2: Targets predicted as DNA-binding on HOLO set but not on APO set.

APO[a]	HOLO[b]	TMP[c]	Seqid[d]	HOLO TMP[e]	HOLO EN[f]	APO EN[f]	AP TMP[g]	HOLO APO [h]
1nfa_	1a02N	1hjbC	82	0.67	-25.70	-1.1	0.53	0.64
1uklC	1am9A	1nlwB	70	0.82	-24.99	-6.5	0.84	0.86
1rxr_[i]	1by4A	1kb4A	83	0.90	-29.57	-20.5	0.81	0.80
1es8A	1dfmA	2bamB	88	0.68	-30.68	14.1	0.64	0.89
1jyfA	1efaA	1rzrA	100	0.90	-12.97	-1.6	0.89	0.96
1i11A	1gt0D	1cktA	52	0.78	-26.68	-9.5	0.73	0.74
1ev7A	1iawA	1cf7A	97	0.55	-23.51	-20.0	0.53	0.82
1q39A	1k3wA	2f5pA	90	0.82	-20.67	-18.4	0.48	0.55
1f43A	1le8A	1fjlA	100	0.88	-19.47	-7.5	0.58	0.64
1bgt_	1sxpA	1y6fA	93	0.75	-19.17	-2.0	0.78	0.98
1mi7R	1trrA	1gdtA	89	0.68	-21.58	-15.0	0.38	0.52
2audA	1tx3A	4rveB	96	0.56	-24.53	-20.2	0.54	0.95

[a]. Targets from APO set; [b]. Targets from HOLO set; [c]. Template; [d]. Sequence Identity between APO and HOLO target calculated by bl2seq in blast2.2; [e]. TM-score between HOLO target and template protein; [f]. Energy value between template-target complex; [g]. TM-score between APO target and template protein; [h]. TM-score between HOLO target and APO target. [i]. template used for HOLO is unable to be used for APO because of $>$35% sequence ID.

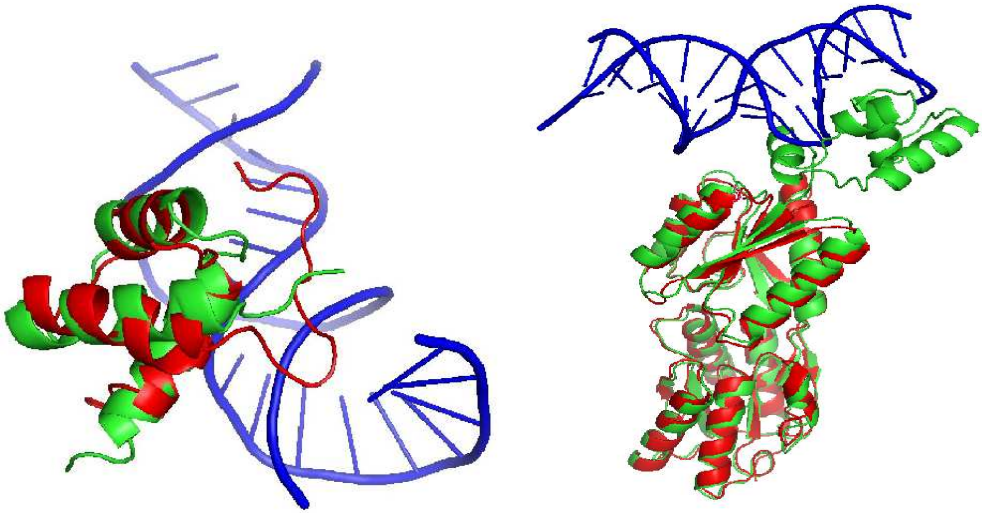

Fig. 2.4: (a) Structural comparison between APO target 1f43A and HOLO target 1le8A. Red: fragment of binding domain of 1f43A. Green: fragment of binding domain of 1le8A. Orange: template DNA of 2bamB. (b) Structural comparison between APO target 1jyfA (red) and HOLO target 1efaA (green). Orange: template DNA of 1rzrA.

in HOLO) leads to incorrect prediction despite insignificant structural change in nonbinding regions of the protein. In another more extreme case (Fig. 2.4b), disordered region in APO structure (1jyfA) changes to ordered binding domain in HOLO structure (1efaA).

Another cause of incorrect prediction in APO and correct prediction in HOLO is large overall structural change. The large overall structure changes lead to poor structural alignment to templates so that their TM-scores are lower than the threshold. For example, despite 90% sequence identity, TM-score between 1q39A in APO and 1k3w in HOLO structures is only 0.48 and leads to the poor alignment of APO structure to template (best is 0.48 in TM-score). We also discovered a technical reason for an APO target (1rxr_). We are unable to use the template employed for the corresponding HOLO target because the sequence identity between the template and its respective APO target is slightly higher than 35%.

There are also 3 targets identified as DNA binding proteins correctly in the APO set but not in the HOLO set. All 3 (1llzA, 1bf5A and 1esgA) are just outside of arbitrary boundaries generated by optimization. This highlights the empirical nature of the proposed approach.

One can further examine the performance of DDNA3O in predicting binding residues. We found that the specificity, accuracy, precision, sensitivity and MCC for predicting binding residues are 94%, 90%, 69%, 64%, 0.59 for the APO set and 95%, 90%, 75%, 67%, 0.63 for the HOLO set, respectively. The performance for the HOLO set is close to the results for training set (93%, 89%, 76%, 66%, and 0.64 for specificity, accuracy, precision, sensitivity and MCC, respectively). This highlights the robustness of DDNA3O.

2.3.4 Test by the DB71 dataset

The additional 71 proteins contained in the updated protein/DNA complex structural dataset (DB71) offer a challenging test set. DDNA3 (DDNA3O) predicts 34 (39) out of 71 proteins as DNA binding proteins. Thus, the sensitivity is 34/71(48%) by

DDNA3 and 55% by DDNA3O. DDNA3O continues to make significant improvement in sensitivity over DDNA3. This 55% sensitivity is 5% lower than the sensitivity of 60% for the HOLO dataset but is higher that the sensitivity of 51% for the APO dataset. This suggests that more than 50% new complex structures are recognizable by DDNA3O with DB179 as templates for protein-DNA complexes for all the sets tested (APO, HOLO, and DB71).

2.3.5 The effect of a larger, updated dataset of DNA-binding proteins (DDNA3U)

To examine the effect of a larger dataset of DNA-binding proteins, we use DB250 and NB3797 as the training set. We found that for this larger, updated dataset, the highest MCC is 0.75 with the same or similar values for three parameters (β=0.4, TM-score threshold of 0.55 and energy threshold of -13.7) as DDNA3. This result highlights the stability of trained parameters with a 40% increase in DNA-binding proteins. The corresponding accuracy, precision and sensitivity are 97%, 87%, and 67%, respectively. In particular, 45 out of 71 additional proteins outside DB179 are recognized as DNA binding by DB250-trained DDNA3 (DDNA3U), the same proteins recognized by DB179-trained DDNA3 (DDNA3) for which 71 proteins are employed as an independent test set.

Application of this newly trained method to APO104 and HOLO104 sets leads to 52(50%) and 64(62%) predicted DNA binding proteins, respectively. That is, a 40% expansion of DNA-binding proteins (from 179 to 250) leads to about 2% improvement in sensitivity. For 52 successfully predicted APO targets, the specificity, accuracy, precision, sensitivity and MCC for predicted binding residues are 94%, 90%, 66%, 63%, 0.58, respectively. The corresponding values for 64 successfully predicted HOLO targets are 95%, 90%, 74%, 67%, 0.63, respectively. However, as Fig. 2.1 indicates, newly trained DDNA3 (labeled as DDNA3U) yields higher sensitivity only when false positive rate $>$0.005. That is, at a lower false positive rate, a larger template database in fact decreases sensitivity and precision.

Table 2.3: Structural Genomics targets (SG1697) predicated as DNA-binding proteins by DBD-Hunter, DDNA3, and DDNA3O.

Method	Prediction	Putative	Other Function	Unknown
DDNA3	32	19	3	10
DDNA3O	27	19	1	7
DBD-Hunter	37	18	3	16
Overlap*	19	15	0	4

$*$Overlap between DBD-Hunter and DDNA3O

Here, by applying TM-Score dependent energy thresholds to the updated DB250/NB3797 databases, MCC hasn't been changed much. This is caused by the increase of number of false positive (from 26 to 34), although with more number of true positive (from 167 to 176). Because we are interested in predicting DNA binding proteins with very low false positive rate ($<$0.005), we will employ the methods (DDNA3 and DDNA3O) trained by DB179 to structural genomics targets.

To further examine the possibility of overfitting in DDNA3U, we perform a ten-fold cross-validation tests on the DB250/NB3797. That is, all the binding and non-binding sets are randomly divided into 10 folds. Each time, one fold is chosen as the test set while the other 9 folds are employed for all training: the statistics of potential energy function, the structure templates for protein-DNA binding, and re-training of the parameters. The test is repeated for 10 times. The method performance is analyzed by 1000 times of bootstrap resampling [73]. We found that the average MCC value is 0.70$\pm$0.02 with the accuracy of 97%, the precision of 88% and the sensitivity of 58%, respectively. It is clear that the only significant change from the leave-one-out results is the reduction of sensitivity from 65% to 58%. This is likely caused by the reduced number of templates in the ten-fold cross-validation. Indeed, if 249 templates are permitted to use, the average MCC value is 0.72$\pm$0.02. Thus, our results are reasonably robust with different trainining.

2.3.6 Application to Structural Genomics Targets

As shown in Table 2.3, application of DDNA3 leads to 32 DNA-binding proteins from SG1697. Among them, 19 out of 32 proteins (59%) are putative DNA binding proteins,

3 out of 32 proteins (10%) are annotated to having other functions while others (31%) have unknown function. DDNA3O decrease the prediction of DNA binding proteins from 30 to 27 without change on the number of putative DNA binding proteins (19) and a decreased number of proteins with other annotated function from 3 to 1. This result further confirms the improvement of DDNA3O over DDNA3. By comparison, DBD-Hunter predicts 37 DNA-binding proteins. Among the 37 proteins, there are 18 (48.6%) putative DNA binding proteins, 3 (8.1%) with other putative functions, and 16 (43.2%) with unknown function. All the putative functions are according to NCBI database.

The overlap between predicted proteins by DDNA3O and DBD-Hunter is only 19 proteins, 15(79%) of which are putative DNA binding proteins. The large fraction of putative DNA binding proteins in overlapped predictions highlights significant improvement in confidence of prediction when a consensus prediction is made. Meanwhile, only 70% proteins predicted by DDNA3O overlap with those by DBD-Hunter highlights that the energy function plays a significant role in prediction. There are 4 putative DNA binding proteins (1ug2A, 1y9bA, 2cqxA and 2fb1A) predicted by DDNA3O but missed by DBD-Hunter. Similarly, there are 3 putative DNA binding proteins (2hytA, 2iaiA and 2od5A) predicted by DBD-Hunter but missed by DDNA3O. The complete list of predicted DNA-binding proteins is shown in Table 2.4. Table 2.4 includes 10 additional predicted proteins from SG2235, 8 of which are putative DNA binding proteins. That is, 80% of predicted proteins from SG2235 are putative DNA binding proteins. This result confirms the prediction quality of the proposed DDNA3O technique.

2.4 Discussion

We have developed a highly accurate method (DDNA3O) to predict DNA binding proteins. This is accomplished by developing a new statistical energy function for predicting DNA-binding proteins. We found that introducing an atom-type dependent volume fraction correction and DNA mutation in the DFIRE statistical energy function

Table 2.4: Targets are predicted as DNA-binding proteins by DDNA3O from SG1697 and SG2235 with function based on GO annotations.

Target	Template	TM-score	Energy	Putative Function
2keyA[d]	1p7dB	0.58	-22.19	DB
2khvA[d]	1p7dB	0.72	-30.06	DB
2kobA[d]	1p7dB	0.75	-26.52	DB[a]
3cecA[d]	3croL	0.75	-21.67	DB
3edpA[d]	1sfuA	0.74	-13.42	DB
3frwF[d]	1trrG	0.77	-23.04	DB
3ic7A[d]	1cf7A	0.61	-17.48	DB
3ikbA[d]	4sknE	0.62	-16.54	DB
3iuvA[d]	1jt0A	0.77	-14.97	UK[b]
3ke2A[d]	1gdtA	0.58	-18.58	UK
1iuyA	1f4kB	0.61	-19.25	NB[c]
1s7oA	1gdtA	0.67	-14.37	DB
1sfxA	1u8rJ	0.72	-24.89	DB
1ug2A	1fjlA	0.58	-17.92	DB
1wi9A	1repC	0.62	-17.50	UK
1x58A	1w0tA	0.87	-24.86	DB
1y9bA	1ea4A	0.67	-22.76	DB
1z7uA	1u8rJ	0.66	-14.75	DB
1zelA	1cgpA	0.56	-20.67	UK
2cqxA	1akhA	0.69	-17.87	DB
2da4A	1akhA	0.74	-27.67	DB
2e1oA	1akhA	0.87	-18.37	DB
2eshA	1f4kB	0.67	-17.10	DB
2esnA	1u8rJ	0.62	-21.74	DB
2ethA	1u8rJ	0.71	-20.94	DB
2f2eA	1u8rJ	0.71	-14.07	DB
2fb1A	2as5F	0.62	-14.47	DB
2fyxA	2a6oB	0.78	-18.83	DB
2g7uA	1u8rJ	0.70	-15.83	DB
2jn6A	1gdtA	0.70	-17.11	DB
2jtvA	2ex5A	0.61	-21.07	UK
2nx4A	1jt0A	0.76	-16.34	DB
2qvoA	1z9cF	0.80	-10.19	UK
3b73A	1z9cF	0.68	-23.89	UK
3bddA	1u8rJ	0.76	-21.56	DB
3bhwA	1fokA	0.58	-19.04	UK
3bz6A	1u8rJ	0.73	-17.02	UK

[a]. Targets are annotated as protein which has putative functions related with DNA binding in PDB. [b]. It is unknown whether a target has putative functions related with DNA binding. [c]. Nonbinding to DNA according to GO annotation. [d]. Targets in SG2235

leads to a significant improvement in the performance in predicting DNA-binding proteins (MCC= 0.76 for DB179/NB3797 by DDNA3O). This is a significant improvement from MCC of 0.69 given by optimized DBD-Hunter. Application of DDNA3O to structural genome targets confirms the accuracy of the proposed method with 73% potentially correct prediction of DNA-binding proteins (annotated as putative DNA-binding), 3% potentially false positives (function annotated but not DNA-binding) and the rest unknown.

For DDNA3, the effect of DNA mutation is small for improving the MCC value of the training set (from 0.72 to 0.73) but is significant for improving the sensitivity from 46/104 (44%) to 50/104 (48%) of the APO test set. We further find that the mutation leads to no significant improvement in sequence identity between template DNA sequence and wild-type DNA sequence. The sequence identities to wild-type DNA sequences before and after mutation are both close to the random value of 25%. One possible reason is the absence of structural refinement for protein during mutation. This result also suggests that DDNA3 is not yet specific enough to identify binding DNA bases.

In principle, exhaustive mutations of DNA base pairs can lead to significant increase in computing time for a long DNA segment. However, because our energy function does not consider base-base interaction by assuming a rigid DNA structure before and after binding, the computing requirement for the exhaustive mutations of DNA base pairs is only four times more than that without base mutations.

One potential concern is insufficient statistics due to the small number of complex structures for deriving the DDNA3 energy function. We have addressed this question by employing the leave-one-out (for both DB179 and DB250 sets) and ten-fold cross-validation (for the DB250 set) techniques. The consistency between different training and test sets provides the confidence about the energy functions obtained.

Another concern is potential overfitting due to 5 threshold parameters in DDNA3O because of the small number of true positives for each TM-Score bins (Table 1). This concern is reduced somewhat as the energy threshold mostly satisfies

the expectation that less similar structures (low TM-Scores) requires higher energy thresholds. Moreover, there is a consistent improvement in sensitivity from training (DB179) to test (APO/HOLO104, DB71, and structural genomics targets). This consisteny makes the improvement statistically significant. However, one certainly can not completely remove the concern of overfitting. More studies as larger data set becomes available are certainly needed.

One advantage of the proposed structure-based prediction method is the prediction of protein-DNA complex structures. The predicted complex structures allow prediction of DNA binding residues. High specificity and accuracy ($>$90%) are achieved for binding residue prediction even for the APO structures (protein structures in the absence of DNA).

The success of DDNA3O is limited by the availability of protein-DNA complexes as templates. A 40% expansion of template databases from 179 to 250 proteins leads to significant improvement in sensitivity if false positive rate$>$0.005 (Fig. 2.1) but also slightly decreases sensitivity if false positive rate$<$0.005. Thus, there is a clear need to further improve the energy function that discriminates binding from nonbinding proteins. The rigid-body approximation employed here likely has limited the performance of DDNA3O. Introducing flexibility to DNA and proteins to DDNA3 is in progress.

Chapter 3 Sequence-based prediction of DNA-binding proteins by fold recognition and calculated binding affinity

Abstract

Structure-based methods are limited because they require structure data as input. For fully understanding the mechanism of protein-DNA interaction, a specialized method for prediction of DBPs from sequence is necessary. Here, we propose to predict DBPs from sequence level by integrating structure prediction program HHM with binding affinity calucation program (DFIRE).

This method was benchmarked on a database with 179 DNA-binding proteins(DBP) and 3797 non-DNA-binding proteins(NDBPs). The final results indicate structure prediction program together with energy function can achieve the MCC 0.77 with an accuracy of 98%, precision 94% and sensitivity 65%. These results are significantly higher than the best MCC value 0.68 from DBD-Threader. This method was applicated on 20270 human genome targets, and discovered 1975 DBPs. Amonge these proteins, 1612 (56%) are annotated as DBPs by GO. The newly developed method is accurate and sensitive in prediciton of CBPs from sequence.

3.1 Introduction

Completion of thousands of genome projects has led to an explosive increase in number of proteins with unknown functions. The comprehensive Uniprot database [74] contains 107 protein sequences and, yet, less than 5% of these sequences have annotated functions from Gene Ontology Annotation database [75]. This gap between sequences and annotations is widening rapidly as inexpensive and more efficient next generation sequencing techniques become available. Experimentally identifying function for millions of proteins is obviously impractical. Thus, it is necessary to develop effective bioinformatics tools for initial functional annotations.

One important function of proteins is DNA-binding that plays an essential role in transcription regulation, replication, packaging, repair and rearrangement. Function prediction of DNA-binding can be classified into three levels of resolution (low, medium and high). A low-resolution prediction is a simple two-state prediction whether or not a protein will bind to DNA. A medium resolution prediction is to predict the region in a protein that binds with DNA (DNA-binding residues or DNA-binding interface regions). A high-resolution prediction is to predict the complex structure between DNA and a target protein of unknown function.

Most existing methods have been focused on low-resolution two-state prediction [22, 27, 28, 42, 56, 62, 67, 76–80, 80–84] and medium-resolution prediction of binding residues [56, 63, 77, 85–89, 89–99].The majority of these techniques are based on machine-learning techniques ranging from neutral networks, random forest, decision trees to support vector machines that are trained on the features derived from sequence (sequence-based) and structure (structure-based). A structure-based technique attempts to infer functions from known protein structures. Both sequence-based [27, 28, 78, 79, 81, 82, 84, 100] and structure-based [22, 56, 62, 67, 77, 80, 83, 101]prediction of DNA-binding proteins were developed. The same is true for sequence-based binding residue prediction [27, 86, 88, 94, 96, 98–100, 102–104].

An alternative approach to above machine-learning techniques is to take advantage of known protein-DNA complex structures. This can be accomplished by structural comparison between a DNA-binding template and a target protein structure [68, 85, 92, 93].For example, we demonstrated that a size-independent, structural alignment method SPalign makes a significant improvement over several other commonly used tools to locate functionally similar structures [68]. If the structure of a target protein is unknown, homology modeling [105, 106] has been employed. Gao and Skolnick further illustrated the importance of combining the predicted structure (through structural alignment [31]or threading [35]) with binding prediction for detecting DNA-binding proteins. One important aspect of this approach is its ability to predict the complex structure between the target protein and template

DNA. This high-resolution function prediction at atomic details allows an improved understanding of binding mechanism and an integration with low and medium-level prediction of DNA-binding proteins and DNA-binding residues.

This work will focus on improving the high-resolution function prediction. The DBD-Threader method developed by Gao and Sholnich [35] first employed the threading technique called PROSPECTOR [107] to predict structures based on known DNA-binding domains. Confidently predicted complex structures are then confirmed for DNA-binding by utilizing a pairwise knowledge-based, contact energy function [31] . The method has achieved the Mathew correlation coefficient (MCC) of 0.68 for the two-state prediction of DNA-binding proteins by using a database of 179 DNA-binding domains (DB179) and 3797 non-DNA-binding domains (DB3797).

In this work, we approach this function prediction problem with different methods for protein-structure prediction and binding prediction. Instead of a contact-based energy function employed in DBD-Threader [35] , we will employ a statistical energy function based on a distance-scaled ideal-gas reference state (DFIRE) [33]extended for protein-DNA interactions [32, 52, 53]. This DDNA energy function is found useful in developing a highly accurate structure-based technique called SPOT-Struc (DNA) that achieves the MCC value of 0.76 for the same database of DB179 and NB3797, employed by DBD-Threader. In addition to energy functions, we will examine two fold-recognition techniques to enable a sequence-based prediction as DBD-Threader. One is a method based on hidden Markov model (HHM) called HHblits [108]. The other is our in-house built technique called SPARKS X [49]. Both methods are among the top performers in critical assessment of protein structure prediction techniques (CASP 9) [49, 109]. This development of SPOT-Seq for DNA-binding proteins is inspired by the success of prediction of RNA-binding proteins by integrating SPARKS for structure prediction and DFIRE for binding prediction [36] and its successful application to human genome [Zhao et al. submitted].

3.2 Methods

3.2.1 Dataset

Gao-Skolnick domain datasets (DB179 and NB3797). These two datasets were compiled by Gao and Skolnick that contains 179 DNA-binding protein domains and 3797 non-DNA binding protein domains. These two sets are developed by collecting the proteins with a resolution of 3 or better, a minimum length of 40 amino acid residues per protein and at least 6 base pairs of DNA and five residues interacting with DNA. The redundant data among two sets are excluded by using 35% sequence identity cutoff. DB179 is used as a template library.

Test set of RNA-binding proteins (RB174). RB174 is a dataset made of 174 high-resolution RNA-binding proteins (whole chains), collected by us in developing SPOT-Seq (RNA) based on a 25% cutoff. We will employ RB174 to examine if the proposed method can separate DNA from RNA-binding proteins.

Independent test dataset (DB82). An independent test set was developed by including the DNA-binding proteins released after December 2009. The protein chains were divided into SCOP domains, and the redundant data was removed by using sequence identity cutoff 30%. We further excluded the proteins that have sequence identity higher than 30% with any proteins in DB179. Finally, we generated an independent test dataset with 82 protein domains and chains.

3.2.2 Function prediction protocol

The prediction protocol proposed here is the same as SPOT-seq (RNA) developed by us [36] , except that 1) the template library is made of known protein-DNA complex structures and 2) HHBlits [108], in addition to SPARKS-X [49] is used in structure prediction. Briefly, HHBlits or SPARKS X is firstly employed to match a target sequence to the template structures in the template library. If a significant match is found based on a Z-score, that is based on the alignment score, relative to the average alignment score for all binding and non-binding proteins in the dataset. The top

matched template(s) will be used to construct model protein-DNA complex structure(s) by copying the query sequence to the template complex structure(s) according to the alignment result from SPARKS X/HHblits while keeping the template RNA intact. The model complex structures are then employed to estimate the binding affinity between the target protein (main-chain only) and the template DNA by utilizing DDNA3 [32]. The target protein is classified as DNA-binding if the binding affinity is higher than a threshold. Thus, there are only two parameters to be optimized: sequence-structure alignment Z-score and the binding energy value.

3.2.3 Other Methods

PSI-BLAST was applied for prediction of DBPs by searching homology sequences from NCBI non-redundant sequence library for four iterations. A target is classified as DBPs, if it has at least one template with E-value lower than an optimized threshold. All templates with sequence identity $>$ 30% with the target sequence are excluded. The E-value threshold is optimized by maximizing the MCC value. PSI-BLAST was downloaded from NCBI. HHblits is a fold-recognition technique that extracts homologous sequences of targets from template library by Hidden-Markov models (HMM). The HHM matrices of targets and templates were built by searching against the Uniprot database. Probability of match was calculated by comparing the HHM matrix of a target to the HHM matrix of a template. We define a target sequence as a DBP if probability of match is higher than a threshold. The threshold is optimized by maximizing the MCC value.

3.3 Results

3.3.1 Low-reolution two-state prediction

Leave-one-out cross validation (Gao-Skolnick Domain-level datasets): This work is accomplished by removing all templates with $>$ 30% sequence identity to the target. The results were obtained by taking one chain sequence from DB179 or NB3797 and predicting whether it binds or does not bind to DNA. Figure 3.1

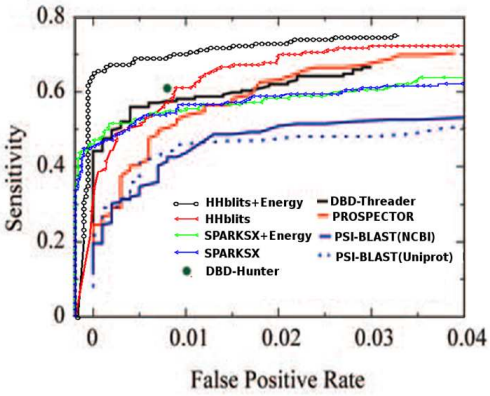

Fig. 3.1: Performance of various methods of DBP prediction for the Gao-Skolnick domain datasets.

Table 3.1: Method comparison for prediction of DNA-binding proteins

Method	SN(%)	PR(%)	SP(%)	ACC	MCC
Structure based					
DBD-Hunter	61	79	92	-	0.681
DDNA3	60	91	99	98	0.73
Sequence based					
PSI-BLAST(NCBI)	49	64	87	-	0.540
PSI-BLAST(Uniprot)b	43	75	93	-	0.553
PROSPECTORb	53	74	91	-	0.609
HHblits	61	69	99	97	0.639
SPARKS X	45	95	99	97	0.647
SPARKS X+Energy	53	84	99	97	0.652
DBD-Threaderb	56	86	96	-	0.680
HHblits+Energy	65	94	99	98	0.771

and Table 3.1 compared both structure and sequence-based methods where results of DBD-Hunter, PSI-BLAST (NCBI), PSI-BLAST(uniprot), PROSPECTOR, and DBD-Threader were obtained from Ref. [35] . We obtained the results of SPARKS X, HHBlits, SPARKS X+Energy and HHBlits+Energy for the same datasets. For sequence-based fold/homology-recognition techniques, SPARKS X yields the highest MCC value (0.647), followed by HHblits (0.639), PROSPECTOR (0.609), and PSI-BLAST (0.553 or 0.540). Adding the energy function to fold recognition leads to a small improvement over SPARKS X (MCC from .647 to 0.652) but a large improvement over PROSPECTOR (MCC from 0.609 to 0.681) and over HHblits (MCC from 0.639 to 0.771). In particular, the best performing HHblits + Energy leads to a sensitivity of 65% and precision of 94%. Such performance is better than the best structure-based technique (DDNA3) with a MCC value of 0.73.

Table 3.2: Detecting DBPs in 19 fold shared by DNA-binding (DB179) and non-binding (NB-3797) proteins

Fold	Fold Name	Dataset (bd/nb)	HHblits (bd/nb)	HHblits+ Energy(bd/nb)
a.38	HLH	5/1	5/0	5/0
a.74	Cyclin	4/10	1/2	1/2
c.52	Restriction endonuclease	14/4	3/0	4/0
a.4	DNA/RNA-binding 3-helical bundle	50/11	23/0	25/0
a.6	Putative DNA-binding domain	2/2	2/0	2/0
c.66	S-adenosyl-L-methionine-dependent methyltransferases	4/19	4/15	3/0
c.62	vWA	2/10	2/0	2/0
g.39	Glucocorticoid receptor	2/12	1/0	1/0
c.37	P-loop-containing-nucleoside-triphosphate hydrolases	5/87	2/5	2/0
d.151	DNase I	2/2	2/2	1/2
a.60	SAM domain	7/1	4/0	5/0
d.95	Homing endonuclease	6/1	2/0	3/0
c.55	Ribonuclease H motif	8/35	2/0	1/0
b.82	Double-stranded beta-helix	1/37	0/0	1/0
c.53	Resolvase	1/5	1/0	1/0
h.1	Parallel coiled-coil	5/43	2/0	2/0
d.129	TBP-like	3/13	0/0	1/0
d.218	Nucleotidyltransferase	1/8	1/0	1/0
Total		122/301	57/24	61/4

Separating DNA-binding from non-DNA-binding in the same SCOP fold. One crucial test for predicting DNA-binding function is the ability of a method to classify DBPs from non-DBPs within the same structural fold. We analyzed 19 SCOP folds shared by DNA-binding (DB179) and non-DNA-binding proteins (NB3797). As shown in Table 3.2, after incorporating the DDNA energy function for DBP prediction, the number of true positives increases from 57 to 61 and false positives decreases from 24 to 4. Thus, removal of false positives is the key factor for large improvement by the energy function.

Separating RNA-binding proteins from DNA-binding proteins: As the RNA-protein interaction shares features with DNA-binding proteins (both are positively charged, for examples), it is important to examine if the proposed method can separate DBPs from RBPs. We tested the HHblits+energy method with the thresholds optimized by DB179+NB3797 datasets on the RBP dataset (RB174). It predicts 5 proteins as

DBPs. Two of the five (1zbiB and 1hysA) are highly homologous (sequence identity ¿70%) to the templates (1zblB and 1r0aA, respectively). As expected, 1zbiB and 1r0aA are two proteins related with both RNA- and DNA-binding functions. 1zbiB is a protein with 12-Mer RnaDNA hybrid and 1r0aA involves the function related with RNA-dependent DNA polymerase. Two of the three remaining proteins (2qk9A and 1ooaA) are known DNA-binding. 2qk9A is Human RNase H catalytic domain that complexed with both RNA and DNA [110] and 1ooaA contains Rel homology domain (RHD) and DNA binding site [111]. The only remaining protein (PDB ID 2jluA) is dengue virus 4Ns3 helicase in complex with ssRNA [112] . This helicase was found to function on both RNA and DNA templates [113]. Thus, there is zero false positive in DNA-binding prediction.

3.3.2 Medium Resolution Prediction of DNA-binding residues

The complex structures predicted from our method allow us to infer amino-acid residues involved in DNA-binding. We define an amino-acid residue as a DNA-binding residue if any heavy atoms of the residue are less than 4.5 away from any heavy atoms of a DNA base. The accuracy of binding-residue prediction is examined on 116 true positive proteins from DB179. The final average values of MCC, precision and accuracy of the prediction are 0.55, 66%, and accuracy 89%, respectively. Fig. 3.2(a) displays MCC values of DNA-binding residues for predicted DBPs along with their corresponding probability of match for predicted structures. Here, the probability of match was clustered into 29 bins and the MCC value is represented by the median value in each bin. It is clear that the high the probability of match can lead the high MCC value, and the correlation coefficient is 0.40.

We employed SPARKS-X to predict binding residues of the 116 targets. The SPARKS-X was used by matching sequences of the targets to their corresponding templates searched by HHblits. The final prediction achieves a MCC 0.54, a precision 63%, and an accuracy 88%. The relationship between the MCC and Z-score is described by Fig. 3.2 (b). Fig. 3.2 includes 116 pointes that were generated by the MCC value

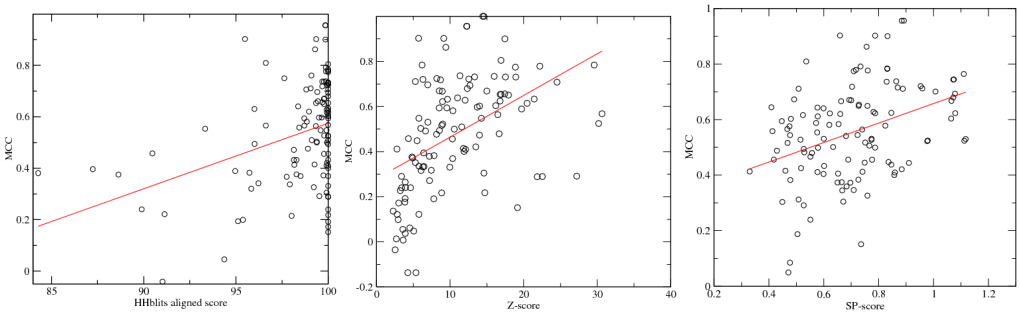

Fig. 3.2: The MCC values for predicting DNA-binding residues as function of HHblits matching probability, Z-s core from SPARKSX and SP-score scoring the similarity between predicted structures and native structu res. (a) Average MCC and matching probability in 29 bins, and the correlation coefficient is 0.40. (b) MCC and Z-score of 116 targets, the correlation coefficient is 0.50. (c) MCC and SP-score of 116 targets, the correlation coefficient is 0.38.

on Y axis and Z-score on X axis. The correlation coefficient between these two values is 0. 50. The high correlation between the predictions on the binding residues and on structure indicates that SPARKS-X is more reliable in prediction of binding residues.

3.3.3 High Resolution Prediction of DNA-binding Complex Structures

The quality of predicted DNA-binding complex structures is examined by the structural alignment SPalign [42] that makes a size-independent comparison between native structures and predicted structures. For 116 correctly predicted targets, the average SPscore is 0.65 (two structures are considered as in the same fold if SPscore¿0.5 [42]).The structure similarity can also be evaluated by the fraction of aligned residues with a root mean-squared distance (RMSD) between two compared structures less than 4. We found that the medium value is 67%.

As an example, Fig. 8.3 compared the predicted binding sites with native binding sites, and the predicted structures with the native structures. For the target (1yfjD, DAM), the predicted (light grey) and actual DNA (orange) location in a similar to the real position, the predicted binding sites (cyan) is also close to the native binding region(yellow) . The MCC value for the predicted binding residues is 0.60. The sequence identity between the target and the template (2g1pA, dam) is 24%.

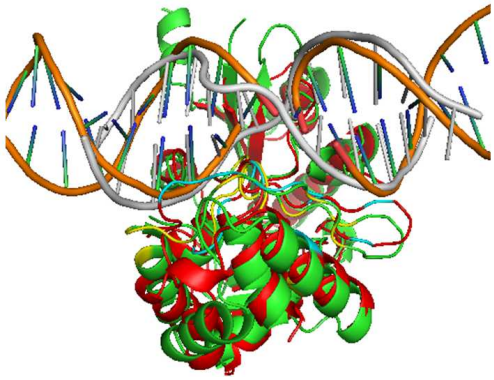

Fig. 3.3: Comparison of predicted (red) and native structures (green) of target 1yfjD (DAM). Native structure and DNA are represented by green and orange, respectively. The predicted structure and DNA are denot ed by color red and grey. The predicted binding sites and native binding sites are in cyan and yellow colors, respectively.

3.3.4 Independent test

We further tested the performance of SPOT-seq (DNA) by detecting the DNA-binding proteins from DB82. Among them, 42 (51%) proteins are correctly predicted as DNA-binding proteins by using the thresholds, matching probability 84% and energy -8.6. We further inferred the binding residues from the predicted complex structures, and compared them with native ones. For 42 correctly predicted DBPs, the MCC 0.64 can be achieved.

3.3.5 Experimental Validation on human TFs

To demonstrate the SPOT-seq DNA is a reliable tool for discover protein-DNA interaction, we tested it on the proteins that were experimentally confirmed as DBPs in the study of protein-DNA profiles [114]. In this study, the researchers characterized the sequence-specificity of 201 TFs, and 136 of them have no binding sites listed in TRANSFAC but confirmed as DBPs by CHIP experiments in this study. Among 201 proteins, we predicted 117 (58%) as DBPs, and 69 (51%) of them are from 136 novel DBPs. From 117 predicted DBPs, 76 are predicted as DBPs by templates with NCBI annoated transcription factor function.

Table 3.3: Number of annotated and predicted DBPs in human genome

Function	#of annotated	#of predicted	Recovery rate
DNA binding	1508	915	61%
TF	1153	684	59%
Others	222	13	6%
Total	2883	1612	56%

3.3.6 Application to human genome

Our approach was applied on detecting DBPs from human genome. The human genome with 20270 proteins was downloaded in 2010 from Uniprot. We applied Gene Ontology (GO) as a tool to annotate the proteins from Human genome. The DNA-related GO annotation can be divided into 8 protein activities, DNA-binding, transcription factor, . Here, the annotation DNA-binding means the annotation 2883 out of 20270 proteins are annotated as DNA-binding by Gene Ontology (GO) database with keywords, DNA binding, transcription factor and others(DNA replication, DNA repair , DNA recombination , DNA helicase activity and keywords related with DNA-binding in biological process). The numbers of proteins in each category of key words are listed in Table 3.3. The newly developed method predicted 1975 out of 20270 proteins as DBPs by using two thresholds [energy: -8.6 and align score: 84.0] as cutoff. 1612 predicted DBPs are also annotated as DBPs by GO. That is, our method recovered 56% (1612/2883) annotated DBPs. The remained predicted DBPs include 104 unknown function proteins (not annotated by GO) and 259 proteins annotated with other functions. The prediction recover rate of targets with keywords of DNA-binding/transcription factor is close to the recovery rate 65% in training dataset. However, the proteins with other functions related with DNA-binding are lower. That is because we define the protein with other DNA-binding functions by using the GO annotation not only related molecular function but also with biological process and cell activity in order to all possible DBPs in human genome. Thus, the proteins in this category could be not directly related with DNA-binding functions.

For 1612 predicted and annotated targets, 371 of them have experimentally obtained structures according to Uniprot annotation. Among them, 28 targets are

Table 3.4: Structure similarity between predicted and native structures of novel DBPs

Target	Template	Seq Identity(%)a	Native structure	SP-score
P38919	2p6rA	16.5	2j0qA	0.797
Q96LI5	1dewB	15.5	3ngnA	0.781
O95718	1kb4A	8.2	1lo1A	0.780
Q13206	1z63A	15.4	2pl3A	0.737
Q9H0S4	1z63A	21.2	3berA	0.729
P32019	1dewB	10.8	3n9vA	0.716
Q9Y2R4	1z63A	17.2	3dkpA	0.714
Q13838	1z63A	17.7	1t5iA	0.710
Q86TM3	1z63A	16.3	3iuyA	0.705
Q9NVP1	2p6rA	17.0	3ly5A	0.703
Q14240	1z63A	19.2	3borA	0.693
O75909	1c9bE	10.3	2i53A	0.683
Q9NRR6	1dewB	14.6	2xswA	0.677
P60842	1z63A	20.5	2g9nA	0.675
Q9UJV9	2p6rA	19.2	2p6nA	0.671
Q9UHL0	1z63A	21.9	2rb4A	0.664
P53370	1rrqA	20.1	3h95A	0.646
P26196	2p6rA	16.1	2waxA	0.641
Q9UMR2	2p6rA	17.9	3ewsA	0.621
Q86W50	2ibsA	17.3	2h00A	0.613

obtained their predicted structures by choosing the most significant template as the one having the same PDB ID as the native structures. For the remained targets, 131 (35%) targets have the predicted structures with the SP-score [42] higher than 0.6 comparing to the native structures. The average SP-score is 0.52. This result is expected since the annotated DBPs with experimental structures have big chance to find their protein folds from template library. We also examined 366 novel discovered DBPs, and found 74 of them with experimental structures. 20 out of 74 (27%) are with the predicted structures having SP-score higher than 0.6 comparing to their native structures and they are shown in Table 3.4. The DNA-binding function of targets has high probability to be further validated experimentally since structural similarity can be used as a criterion to distinguish DBPs [32].

We further analyzed the results of predicted DBPs from human genome by employing DAVID database as another protein function annotation tool. We found that 49 (13%) out of 363 predicted but not annotated targets are annotated as DBPs by DAVID database. The remained 363 novel targets are inputted into KEGG database

to find their involved pathways. Among them, 72 targets have their related disease pathways, and 1 target, DDX58 (O95786), is involved in Cytosolic DNA-sensing pathways and annotated as related with Nucleotide-binding function by Uniprot database. The remained 66 targets are related with 233 diseases. 19 out of 71 targets are involved in the disease of congenital disorders of DNA repair systems.

Chapter 4 Template-based Prediction of RNA-binding Domains and RNA-binding Sites and Application to Structural Genomics Targets

Abstract

Identifying RNA-binding proteins and RNA-binding sites is an important first step toward mechanistic understanding of many key cellular processes. RNA-binding proteins and RNA-binding sites are often predicted separately by employing machine-learning methods with sequence and/or structure-based features to separate RNA-binding from nonbinding proteins or amino-acid residues. Here, we propose an approach that simultaneously identifies RNA-binding proteins and binding regions based on structural alignment to known protein-RNA complex structures followed by binding assessment with a distance-dependent knowledge-based energy function. We showed the importance of using a Z-score to measure relative structural similarity and dividing structures into domains to improve the sensitivity of detecting RNA-binding proteins. This method achieves an accuracy of 98% and a precision of 87% for predicting RNA binding proteins and an accuracy of 93% and a precision of 76% for predicting RNA binding amino-acid residues for a large benchmark of 212 RNA binding and 6761 non-RNA binding domains (leave-one-out cross validation). Additional tests revealed that the method only makes one false-positive prediction out of 213 DNA-binding proteins and correctly identified close to one third of 75 unbound (APO) RNA-binding domains with an accuracy of 93% and a precision of 64% for predicted binding residues. Application of this method to 2076 structural genomics targets predicted 15 targets as RNA-binding proteins, 13 (87%) of which are putatively RNA-binding with the remaining two having unknown function. The method is implemented as a part of the SPOT (Structure-based function-Prediction On-line Tools) package available at http://sparks-lab.org/spot

4.1 Introduction

RNA-binding proteins (RBPs) make specific binding with RNAs and play an important role in translation regulation and post-transcriptional processing of pre-mRNA including RNA splicing, editing, and polyadenylation [7]. Interactions between proteins and RNA influence the structure of RNA and play an critical role in their biogenesis, stability, function, transport and cellular localization. RNA and proteins are stably bound together as Ribonucleoprotein (RNP) complexes throughout journey from synthesis to degradation in a temporal and spatial manner [8]. Proteomic studies in human further showed that RBPs are associated with cell cycle checkpoint defects, genomic instability and cancer [115]. Thus, a comprehensive, mechanistic understanding of a wide variety of cellular processes requires the identification of RNA-binding proteins and RNA-binding sites.

Identifying RNA-binding proteins and binding residues is often treated as two separate problems. Several classifiers dedicated for predicting RBPs are developed by employed support-vector machines (SVM) [27–29, 116]. In some studies [27, 116], homologous sequences were not excluded from training or testing. Performance for most methods was not measured by standard measure of a receiver operating characteristic (ROC) curve or the Matthews Correlation Coefficient (MCC). The only reported MCC value for RBP classification is 0.53 for a sequence-based SVM classifier (5-fold cross validation on 134 RNA binding and 134 nonbinding proteins) [30] and 0.72 for a structure-based SVM classifier for a dataset of 76 RNA binding proteins and 246 non-nucleic-acid binding proteins (leave-one-out test) [117]. The latter, however, is unable to distinguish RNA binding proteins from DNA binding proteins.

Separately, RNA-binding residues are predicted by employing sequence-based [30, 118–124] and structure-based [117, 125–129] information. Sequence-based predictors have employed a number of machine-learning or statistical techniques such as neural-network [118], SVM [30, 121–124], and a naive Bayes classifier [119, 120]. Structure-based predictions, on the other hands, relied on patches built on electrostatics,

evolution and geometric information [117,125], accessible surface and contact network topology based on SVM and naive Bayes classifiers [126], linear-regression analysis of structural neighboring information combined with sequence profiles [127], secondary structure, solvent accessibility, sidechain environment, interaction propensity and other features with a random forest method [128], and a simple propensity-based technique [129]. The best reported MCC values are between 0.47-0.51 [30, 127, 128] for both sequence and structure-based techniques.

One issue facing binding-site prediction is that it will predict RNA binding sites even for the proteins that do not bind RNA. In this work, we will predict RBP and RNA binding site within a single method. This method is based on a recently developed approach [31,32] that was successfully employed for identifying DNA-binding proteins and binding sites. In this approach, protein structures in known protein-DNA complex structures are employed as templates and structurally aligned to the target protein structure. If structural similarity between the target structure and a template is observed, the predicted protein-DNA binding complex structure is confirmed by the prediction of protein-DNA binding affinity.

Here, we will extend this structure-based approach by developing a distance-dependent knowledge-based energy function for protein-RNA interactions. Only a few knowledge-based energy function for protein-RNA interactions have been developed so far [130, 131]. Here, we will build the statistical energy function based on a distance-scaled, finite, ideal gas reference (DFIRE) state, initially developed for proteins [33,68,69] and subsequently extended to protein-DNA interactions [32,52,53]. This new energy function, together with a measure of relative structural similarity by Z-score makes an accurate domain-based prediction of RNA-binding proteins and binding residues. The Mathews correlation coefficients for RNA binding domains and RNA-binding residues are 0.56 and 0.71, respectively, for a large benchmark of 212 RNA-binding and 6761 non-RNA-binding domains. The accuracy of the new technique is further validated on 213 DNA-binding domains (negatives) and 75 unbound APO

structures (positives) and applied to uncover RNA-binding proteins from structural genomics targets.

4.2 Methods

4.2.1 Datasets

RB250: Template library of RNA-binding domains. A template library was built by querying the PDB (July 2009 release) to retrieve all protein-RNA complex structures determined by X-ray (resolution better than 3.0Å). The resulting 419 complex structures were split into chains and the chains are further divided into domains by using an automatic domain parser program called DDOMAIN [132] (with the parameter set that mimics SCOP annotation [133]). These domains were further clustered with a sequence-identity cutoff of 95% with BLASTClust [134]. One representative was randomly selected from each cluster. There is a total of 250 representative domain structures with at least 40 amino acides long and at least 5 residues contacting with 5 or more RNA bases. A protein residue and a RNA base are considered in contact if the shortest distance between any pair of heavy atoms from them is within 4.5Å. These representative structures (RB250) form the template library for predicting RNA-binding proteins and binding sites.

RB212: Non-redundant RNA-binding domains. We further obtain a non-redundant RNA-binding domains by using BLASTClust [134] at a 25% sequence identity cutoff. There is a total of 212 domains (the RB212 set).

NB6761: Non-RNA-binding data set. A non-redundant set of 8770 protein structures was obtained by using PISCES [135] with a 30% global sequence identity cutoff, a resolution better than 3Å and a chain length cutoff of 40 amino-acid residues. We removed those chains whose function is associated with RNA-binding and whose PDB records contain the key words ”RIBOSOMAL”, ”UNKNOWN FUNCTION” and ”RNA” by searching in the title. The remaining 6699 chains were divided into domains with DDOMAIN [132] and clustered with a sequence identity cutoff 25% by BLASTClust [134]. One representative was randomly selected from each cluster. The

final dataset contains 6761 protein domains that do not binding RNA (NB6761). We emphasize that DNA-binding proteins are not excluded from this dataset.

APO75/HOLO75 dataset. To examine the effect of binding induced conformational changes on the accuracy of predicting RNA-binding proteins, we established a dataset with both bound (HOLO) and unbound (APO) structures. We started with the set of bound structures (RB250) and performed BLAST [134] search for the sequences homologous to the sequences in RB250. We selected those homologous sequences whose protein structures do not contain RNA. These unbound APO structures are partitioned into domains by using the DDOMAIN program [132]. An all-against-all sequence alignment between the APO domain set and the HOLO domain set from RB250 was performed by employing the ALIGN0 program from the FASTA2 package [136]. The alignment yielded 869 pairs with sequence identity above 45% that are further culled by excluding redundant sequences with a identity cutoff of 30% and removing the structure with lower resolution. The final set contains 75 APO domains whose sequence identity ranges from 45% to 100% to their corresponding HOLO domains. The majority (56 out of 75 pairs) are more than 85% sequence identity. The APO and their corresponding HOLO domain sets are labeled as APO75 and HOLO75, respectively.

DB213: DNA-binding protein database. To examine the ability to distinguish RNA-binding and DNA-binding proteins, we also obtained a DNA-binding protein dataset composed by 179 DNA-binding structures [31]. These DNA-binding structures were divided into domains by DDOMAIN and clustered by BLASTClust [134] sets. The clustered 232 domains were further reduced with a sequence identity cutoff of 25% to produce the final dataset of 213 DNA-binding domains (DB213).

SG2076: Structural Genomics targets. A set of 2076 domains is obtained from previously collected 2235 structural genomics targets [32] by domain parsing (DDOMAIN) and clustering (BLASTClust) with a sequence identity cutoff of 35%.

RNA binding domain supperfamily(RBD). RBD(RNA binding domain) or RRM(the RNA-recognition motif) is the most abundant RNA-binding domain in eukaryotes

[137]. For this domain,the mode of protein and RNA interaction is varible. This domain can modulate its fold to recognize many RNAs and proteins to achive multiple biological function [138]. The dataset RRM was built to test the performance of our method on annotation of RNA-binding proteins of RRM supperfamily. The dataset is obtained from SCOP supperfamly database. RRM supperfamily is devided into 5 families:Canonical RRM, Non-canonical RRM, Splicing factor U2AF subunits, Smg-4/UPF3 and GUCT, which respectively contain 171 PDB, 4 PDB, 1 PDB and 1 PDB. These PDBs are split into chains and then divided into 292 domains. 280 domains blongs to canonical RRM family, 9 domains are inclued into non-canonical family, and others are contained into splicing factor U2AF subunits, smg-4/UPF3 and GUCT families, respectively.

4.2.2 Knowledge-based energy function

We employed exactly the same volume-fraction corrected DFIRE energy function that generated DDNA3 [32] to produce an DRNA energy function for protein-RNA interaction $\overline{u}_{i,j}^{\mathrm{DRNA}}$.

$$\overline{u}_{i,j}^{\mathrm{DRNA}}(r) = \begin{cases} -\eta \ln \frac{N_{obs}(i,j,r)}{\left(\frac{f_i^v(r) f_j^v(r)}{f_i^v(r_{cut}) f_j^v(r_{cut})}\right)^{\beta} \frac{r^{\alpha}\Delta r}{r_{cut}^{\alpha}\Delta r_{cut}} N_{obs}^{lc}(i,j,r_{cut})}, & r < r_{cut}, \\ 0, & r \geq r_{cut}, \end{cases} \tag{4.1}$$

where the volume-fraction factor $f_i^v(r) = \frac{\sum_j N_{obs}^{\mathrm{Protein-RNA}}(i,j,r)}{\sum_j N_{obs}^{\mathrm{All}}(i,j,r)}$, $N_{obs}(i,j,r)$ is the number of pairs of atoms i and j within the spherical shell at distance r observed in a given structure database, r_{cut} is the interaction cutoff distance, Δr_{cut} is the bin width at r_{cut}, the value of α (1.61) was determined by the best fit of r^{α} to the actual distance-dependent number of ideal-gas points in finite protein-size spheres, the value of β (0.4) was optimized for protein-DNA interactions [32]. We employ residue/base specific atom types with a total 253 atom types (167 for protein and 86 for RNA). We cutoff interactions at 15Å (r_{cut}) with a bin width of 0.5Å(Δr) as for the protein-DNA

interaction [32]. We also set the factor η arbitrarily to 0.01 to control the magnitude of the energy score. The RB250 set was used to train the statistical energy function (i.e. to calculate $N_{obs}(i, j, r)$). To avoid overfitting, we employed the leave-one-out scheme to training multiple statistical energy functions for different targets. For each target, we exclude all template proteins whose sequence identity to the target protein is higher than 30%.

4.2.3 Prediction protocol

The protocol for predicting RNA-binding proteins and binding sites is as follows. First, the target structure is scanned against those templates with sequence identity lower than 30% in the template library (RB250) by using the structural alignment program TM-align [139]. If the structural similarity score is higher than a threshold, the protein-RNA complex structure is predicted by replacing the template structure with the aligned target structure. Two structural similarity scores are employed: one is based on the raw TM-Score and the other one is based on Z-score (see results). If the lowest binding energy between the target protein and template RNA is lower than a threshold and the structure similarity is higher than a threshold, the target is predicted as a RNA-binding protein and its RNA binding site can be predicted from the predicted protein-RNA complex structure. If no matching template is found to satisfy these two thresholds, this target is predicted as a non-RNA binding protein.

4.2.4 Performance Evaluation

The performance of the proposed method is measured by sensitivity [$SN = TP/(TP + FN)$], specificity [$SP = TN/(TN + FP)$], accuracy [$AC = (TP + TN)/(TP + FN + TN + FP)$], and precision [$PR = TP/(TP + FP)$]. In addition, we calculate a Matthews correlation coefficient given by

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \tag{4.2}$$

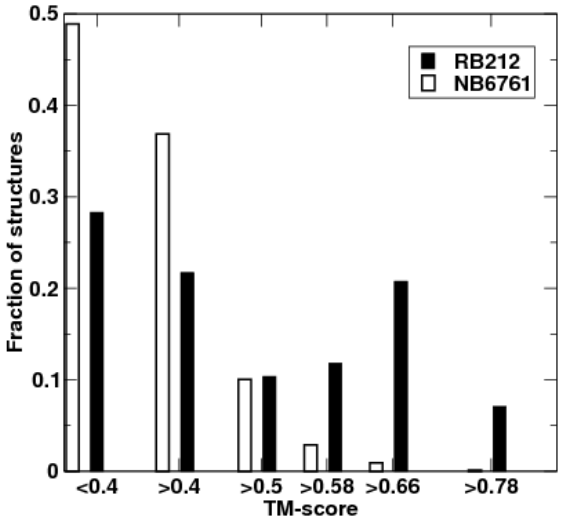

Fig. 4.1: Distribution of the top TM-score-ranked templates on RB212/NB6761

Here TP, TN, FP and FN refer to true positives, true negatives, false positives and false negatives, respectively. This performance measure is applied to both binding-protein prediction and binding-residue prediction.

4.3 Results

4.3.1 Using structural similarity measured by TM-Score for discrimination

We first examine the ability of the structural similarity measured by TM-Score from TM-align [139] for discriminating RNA-binding proteins from non-binding proteins. TM-Score is 1 for 100% structural similarity and around 0.2 between two random protein structures. Fig. 4.1 shows the fraction of the target domains (binding or nonbinding proteins) as a function of the highest TM-Score from its alignment to the templates in the RB250 set, generated by the leave-one-out scheme. 48% binding targets (from RB212) but only 14% nonbinding targets (from NB6761) have a TM-Score of more than 0.5 with at least one binding template. When the threshold of TM-Score is 0.58, 40% binding targets but only 3% nonbinding targets have a hit to a binding template. Increasing the TM-Score threshold further reduces the fraction of non-RNA-binding domains relative to that of RNA-binding domains. However, the highest MCC value is only 0.29 at the TM-score threshold of 0.72. Thus, the structural

similarity based on TM-Score alone has a weak ability to discriminate RNA-binding proteins from non-binding proteins.

4.3.2 Using relative structural similarity measured by Z-Score for discrimination

The structural similarity measured by TM-score between two protein domains with significantly different sizes is normalized by the average size. This structural similarity will be small if the smaller target has a nearly perfect match to only a small portion of the larger template (the binding region). To help remediate this situation, we introduce a relative structural similarity based on Z-score. For a given target whose TM-Score is greater than 0.4 with a binding template, the Z-score of this target is defined as follows:

$$\text{Z-score} = \frac{\text{TM-Score}_{\text{tT}} - \sum_i \text{TM-Score}_{Ti}/n}{\sqrt{\sigma}} \tag{4.3}$$

where TM-Score$_{\text{tT}}$ is the structural similarity score between the target t and a RNA-binding template T, TM-Score$_{Ti}$ is the structural similarity score between the template T and a reference structure i, n is the number of reference structures, and σ are the standard deviation of TM-Score$_{Ti}$. Here, we use the mixed binding and nonbinding proteins (RB212 and NB6761) as the reference structures and choose only top TM-Score ranked structures ($n = 6300$) and exclude the structure pairs TM-Score higher than 0.7 to avoid noises from irrelevant or high homologous structures. TM-Score$_{Ti}$ and σ for each binding template can be pre-calculated and stored.

Fig. 4.2 displays the fraction of target structures as a function of the highest Z-score from its structural alignment to binding templates. 42% binding targets (from RB212) but only 2.5% nonbinding targets (from NB6761) have a Z-score of more than 1 with at least one binding template. When the Z-score threshold is 2, 20% binding targets but only 0.01% (11) nonbinding targets have a hit to a binding template. Increasing the Z-score threshold further reduces the fraction of non-RNA-binding domains relative to that of RNA-binding domains. The highest MCC value is 0.48 at the Z-score threshold of 1.4. Thus, the relative structural similarity based on Z-score alone is a substantially better than TM-Score to discriminate RNA-binding proteins from non-binding proteins.

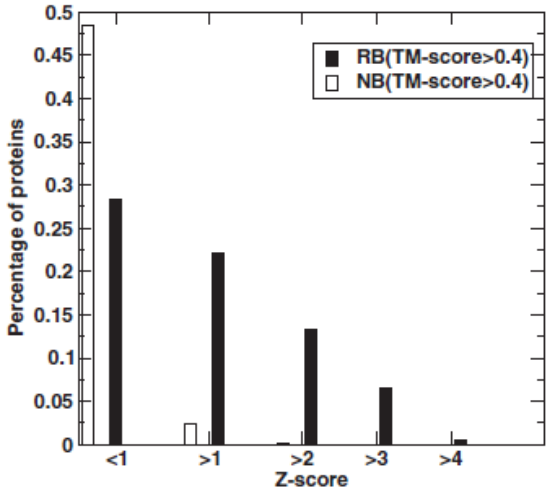

Fig. 4.2: Distribution of the top Z-score-ranked templates on RB212/NB6761

4.3.3 Combined with the DRNA binding energy score for discrimination

To further improve the discriminative power, we calculate the DRNA binding energy [Eq. (1)] based on the predicted complex structure generated from structural alignment of the target with the binding template. Using the leave-one-out scheme on RB212/NB6761, we have optimized TM-Score and binding affinity thresholds to achieve the highest MCC value by a simple grid-based search. The grid for TM-score is 0.01. For the binding affinity threshold, we obtained the lowest energy in all predicted complex structures under different TM-score thresholds for a given target. These energy values are considered sequentially as the energy threshold. The highest MCC is 0.49 for the TM-score threshold of 0.60 and the energy threshold of -15.3. The corresponding accuracy, precision, and sensitivity are 98%, 77%, and 32%, respectively.

Similarly, we can combine Z-score with the DRNA energy score for RNA-binding discrimination. With a grid of 0.1 for the Z-score threshold, we found that the highest MCC is 0.57 with the Z-score threshold of 1.2 and the energy threshold of -9.9. The corresponding accuracy, precision, and sensitivity are 98%, 91%, 36%, respectively. It is clear that combining Z-score and binding affinity score substantially improves precision (10%) and sensitivity (5%) without changing the accuracy (98%) over combining TM-score and binding affinity.

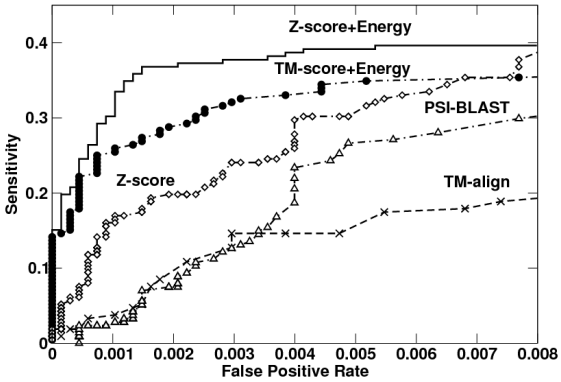

Fig. 4.3: Sensitivity verus false positive rate, given by TM-align(plus), PSIBLAST(open triangle), TM-score combining with the DRNA energy score (closed circle), Z-score (open diamond), and Z-score combining with the DRNA energy score (solid line).

4.3.4 Methods Comparison

To further benchmark the performance of our approach, the ROC curves given by various methods are shown in Fig. 9.2. PSI-BLAST [134] was performed with 4 iterations of searching against NCBI non-redundant protein sequence library. A target is identified as a RNA-binding protein by PSI-BLAST if it has at least one template from RB250 with an E-value higher than a specific threshold (excluding all templates with 30% or higher sequence identity to the targets). The highest MCC of PSIBLAST is 0.41 with accuracy 97%, precision 54% and sensitivity 33%. This MCC value is higher than the method based on TM-align but lower than the method based on Z-score alone (0.48). The combination of Z-score with energy is the most effective in detecting RNA-binding proteins. The combined technique can achieve a reasonable sensitivity at a very low false positive rate.

4.3.5 Test on APO75/HOLO75 datasets

The trained method (combined Z-score and binding affinity) is further benchmarked on APO75/HOLO75 datasets. For a given target, any template with sequence identity $>$30% was excluded from the template library. The number of positive predictions are 31 for the APO set, and 32 for the HOLO set, respectively. These numbers correspond to a sensitivity of 41% for APO75 and 42% for HOLO75, compared with the value of 37% (78/212) observed in RB212. That is, using monomeric unbound structures leads to 1% reduction of sensitivity.

Table 4.1: Targets are predicted as RNA-binding on HOLO set but not on APO set.

HOLO[a] /APO[b]	TM_{HA}[c]	SeqID[d]	TMP[e]	TM_{H}[f]	Z_{HT}[g]	E_{H}[h]	TM_{AT}[i]	Z_{AT}[j]	E_{A}[k]
2atwA2 /1hh2P3	0.95	47.9	2asbA3	0.66	1.4	-17.4	0.57	0.98	-14.7
1uvlA /1hi8B	0.98	96.2	2r7xA	0.43	1.2	-27.9	0.42	1.1	-25.9
2j03S /1ovyA	0.56	54.3	1jj2M	0.60	1.2	-59.3	0.46	1.1	-37.3

[a]. Targets from HOLO set; [b]. Targets from APO set; [c]. TM-Score between HOLO and APO targets; [d]. Sequence Identity between APO and HOLO target calculated by bl2seq in blast2.2; [e]. Template for HOLO target [f]. TM-score between template and HOLO target; [g]. Z-score between HOLO target and template; [h]. Binding energy of template RNA-HOLO target complex; [i]. TM-score of APO target and template; [j]. Z-score of APO target and template; [k]. Binding energy of template RNA-APO target complex;

A more detailed analysis on predicted results shows that there is an overlap of 28 predicted positive results between the APO and HOLO sets. These predictions agree because RNA binding only leads to minor conformational changes in these cases. There are 3 correctly predicted HOLO targets but incorrectly predicted APO targets as shown in Table 4.1.Three APO targets (some even with only small structural changes due to binding) have strong protein-RNA binding (lower than the energy threshold) but with borderline Z-score values ($0.98-1.1$ versus 1.2, the Z-score threshold). The result suggests the need to further improve structural similarity measure. Furthermore, there are 2 correctly predicted APO targets but missed by HOLO targets prediction.One target 2bggB2 has Z-score 2.4 much higher than threshold 1.2 but with a borderline energy (-9.8 vs. -9.9, the energy threshold). Another HOLO target 1ec6A is missed which is caused by technical reason because the sequence identity between the target and the template is higher than 30%.

4.3.6 Binding sites prediction

The predicted binding complexes can be employed to infer the RNA binding residues. We define an amino-acid residue as a RNA-binding residue if any heavy atom of that residue is less than 4.5Å away from any heavy atom of a RNA base. Predicted binding residues from template-based modeling can be compared to actual binding residues. For 77 predicted RNA-binding proteins from RB212, we achieved 75% in sensitivity, 96% in specificity, 93% in accuracy, 76% in precision, and 0.72 for the MCC value.

For predicted HOLO targets, we achieved 56% in sensitivity, 96% in specificity, 92% in accuracy, 65% in precision, and 0.56 for the MCC value. For predicted APO targets, we achieved 55% in sensitivity, 97% in specificity, 92% in accuracy, 64% in precision, and 0.56 for the MCC value.

4.3.7 Discriminate against DNA-binding proteins

We further examine the ability of our method to separate DNA-binding from RNA-binding proteins because they share common structural features [117]. We apply our approach to the set of 213 DNA-binding domains. Only four (1sfuA,1h38D2, 1zblB and 1p7hN) out 213 targets are recognized as RNA-binding proteins. Two of these three targets (1h38D2 and 1zblB) are annotated as DNA/RNA binding proteins [140, 141]

4.3.8 Application to RRM superfamily

Appliation of this method was preformed on prediction of RNA-binding proteins from RRM superfamily. The trained thresholds (Z-score 1.2 and energy -9.9) was used. 250 (250/290) canonical family are predicted as RNA-binding. All of these 250 domains are RNA-binding domains.4 out of 9 non-canonical family are RNA-binding domains,which are not recognized by our method. Other 5 domains are leucine-rich repeat domains(LRR), which is required in cis to the RNP domains for CTE RNA binding [142, 143]. The remained domains that blong to Splicing factor U2AF subunits, Smg-4/UPF3 and GUCT are predicted correctly.

4.3.9 Application to structural genomics targets

This method was applied to 2076 structural genomics domains of unknown function. Based on the same thresholds (Z-score of 1.2 and energy of -9.9) that yielded the highest MCC on the leave-one-out benchmark test of RB212/NB6761, we predict a total of 25 targets as RNA-binding proteins (Table 4.2). Among them, 22 out of 25 (88%) targets are putative RNA-binding proteins according to NCBI annotations. One target

Table 4.2: Structural Genomics targets (SG2076) predicated as RNA-binding proteins

Target	Template	TM-score	Z-score	Energy	Putative Function
1vhyA1	2rfkA2	0.56	1.5	-14.0	RB[a]
1nnhA	1asyA2	0.78	2.8	-13.5	RB
1nzjA	1gaxA1	0.49	1.2	-16.8	RB
2oceA5	2ix1A4	0.65	1.4	-12.2	UK[b]
2f96A	2a1rB	0.57	1.4	-13.5	RB
2cphA	1fxlA2	0.70	1.3	-17.9	RB
3cymA1	2a1rB	0.56	1.3	-11.9	RB
1tuaA1	1ec6A	0.68	1.4	-11.5	RB
2q07A2	1r3eA2	0.67	2.1	-10.9	RB
1yvcA	2bh2A1	0.72	1.8	-13.5	RB
1t5yA2	1r3eA2	0.77	2.8	-15.3	RB
3go5A2	2ix1A4	0.68	1.5	-13.7	RB
2k52A	2ix1A4	0.63	1.3	-12.4	RB
1zkpA	2fk6A	0.78	2.3	-15.9	RB
1x40A	2f8kA	0.62	1.3	-10.8	UB[c]
2ogkD	1jj2D	0.62	1.8	-25.5	RB
2cpfA	1fxlA2	0.74	1.5	-12.0	RB
1yezA	2bh2A1	0.69	1.6	-14.9	RB
2e5hA	1fxlA2	0.74	1.5	-13.3	RB
3frnA3	1jj2J	0.51	1.2	-20.4	UK
2jz2A	1jj2P	0.59	1.3	-33.5	UK
3ir9A	1rlgB	0.56	1.2	-11.5	UK
3hp7A1	1h3eA2	0.63	1.4	-12.5	RB
1wi6A	1fxlA2	0.70	1.3	-17.6	RB
1wdtA4	1fjgI	0.55	1.4	-29.7	RB

[a]. Targets are annotated as having putative functions related to RNA binding in the NCBI database. [b]. Function unknown. [c]. Non-RNA binding

1x40A has phosphorylation site and may have the putative function related with protein binding. The function of the remaining two proteins is unknown.

4.4 Discussion

In this study, we developed a new approach to predict RNA-binding proteins and binding sites simultaneously. This approach is based a similar, successful approach employed for predicting protein-DNA binding proteins with structural alignment to known complex structures followed by evaluation of binding affinity [31, 32]. The main distinction in this paper is the employment of Z-score, rather than TM-Score to measure structural similarity and development of a statistical energy function for protein-RNA interaction based on a volume-fraction-corrected DFIRE reference state [32]. The proposed technique is able to identify RNA-binding proteins with low sequence homology ($<$30% sequence identity) but have high structural similarity in binding regions to known RNA-binding proteins. More importantly, the majority of HOLO structures (28 in 32) detected for RNA-binding continues to be classified as RNA-binding when APO structures are employed. The reduction of sensitivity in detecting binding proteins from 75 HOLO to APO structures is 2% (from 41% and 43%). Furthermore, its successful application to structural genomics targets (23 out of 25 predictions are annotated as putative RNA-binding proteins) confirms the usefulness of the proposed method. This method is applicated to recognize RNA-binding proteins from RRM supperfamily. The result indicates that this method has the strong ability to detect the proteins with canonical binding domains but is weak on the recognition the proteins with non-canonical binding domains. Since this method is template-dependent, the fold of non-canonical domains is novel and fails to find the template with the similar fold. The structural comparison results show that the TM-score and Z-score of these domains are ranged from $0.40-0.59$ and $0.11-0.87$, respectively.

The employment of Z-score, rather than TM-Score, to measure structural similarity is because the TM-score for aligning two protein structures with significantly different sizes strongly depends on how the TM-score is normalized. Z-score provides a

simple way of removing size dependence through a normalization of standard deviation of TM-scores against reference structures of mixing RNA-binding and nonbinding proteins. Z-score alone yields a respectable MCC value of 0.48 and its combination with the DRNA energy function leads to the MCC value of 0.57. By comparison, TM-score alone only achieves a MCC value of 0.29. We have chosen about 9/10 top-ranked TM-scores (6300 values) and removed the TM-scores larger than 0.7 to calculate average and standard deviation of TM-score for a given template. This was an optimized value in order to reduce noises from irrelevant random reference and high homologous structures. The MCC value reduces somewhat to 0.52 if all structures (RB212+NB6761) are employed as reference structures in calculating Z-score.

Another change in RNA-binding protein prediction from DNA-binding protein prediction is the use of binding domains as templates. We found that if whole chains are employed as templates and targets (i.e. the datasets of RB176 and NB5667), the highest MCC values are 0.39 for the combined use of TM-Score and DRNA energy score and 0.47 for the combined use of Z-Score and DRNA energy score. The latter has an accuracy of 98%, a precision of 87%, and, a sensitivity of 26%. Compared to the domain-based prediction, the employment of domains leads to 9% improvement in sensitivity without changing accuracy and precision. This result is consistent the fact that other methods such as phylogenic analysis and protein modeling work best for single domains [144].

It is difficult to make an exact comparison with existing machine-learning based techniques because we have used a significantly large database of non RNA-binding proteins for training and leave-one-out cross validation. This mimics the realistic situation that RNA-binding proteins are only a small fraction of all proteins. Existing machine learning techniques are typically trained on equal or similar number of RNA-binding and non-binding proteins. It is possible that these methods would have substantially higher false positive rates when they were applied to a significantly larger set of non-binding proteins most of which are unseen by machine learning techniques. Nevertheless, we have achieved a comparable MCC value of 0.57 with the

largest nonredundant set of RNA-binding proteins and nonbinding proteins (including RNA-binding ones), compared to 0.53 for a sequence-based classifier (5-fold cross validation on 134 RNA-binding and 134 non-RNA binding proteins) [30] and 0.72 for structure-based classifier for a database of 76 RNA binding proteins and 246 non-nucleic acid binding proteins, leave-one-out test) but the latter is unable to separate RNA from DNA binding proteins [117].

One advantage of the proposed structure-based method is simultaneous prediction of protein-RNA complex structures. The predicted complex structures allow prediction of RNA binding residues. High specificity and accuracy($>$90%) are achieved for binding residue prediction even for the APO structures. Our MCC values for binding site prediction range from 0.71 for leave-one-out cross validation, 0.56 for HOLO targets and APO targets. These results can be compared to the best reported MCC values between 0.47-0.51 for sequence and structure-based binding site prediction [30, 127, 128].

One potential concern is insufficient statistics due to the small number of complex structures for deriving the DRNA energy function. However, a smaller dataset of 179 protein-DNA complexes was employed for obtaining the DDNA3 energy function for protein-DNA interaction and its robustness is found via various tests [32]. Here, we have addressed this question by employing the leave-one-out (for RB212 sets) technique. The consistency between the leave-one-out and APO/HOLO test sets provides the confidence about the energy functions obtained.

One possible way to improve our prediction is to introduce an energy threshold that is dependent on structural similarity threshold because one expects that the binding-energy requirement should be stronger for less similar structures but weaker for highly similar structures between template and query. Previously, we found that introducing a TM-Score dependent energy threshold makes significant and consistent improvement in predicting DNA-binding proteins [32]. Here, we found that introducing TM-Score dependent energy threshold does lead to an increase of the MCC value from 0.49 to 0.52. However, an Z-Score dependent energy threshold leads to no significant

change (0.5690 versus 0.5694). Thus, we employed two independent (Z-score and energy) thresholds only in this work.

The success of our proposed technique is limited by the availability of protein-RNA complexes as templates. It cannot predict RNA-binding proteins with novel structures or binding modes that are not included in the template library. We have used DB250 based on 90% sequence-identity cutoff as template library for the purpose of maximizing available templates. The low sensitivity (32-39%) in various tests is likely in part due to lack of structurally matching templates. On the other hand, binding induced conformational changes suggest that the rigid-body approximation employed here likely has limited the performance of DRNA to discriminate the binding from nonbinding proteins. How to improve our method by incorporating protein flexibility is a challenging problem to be addressed.

Compared to our corresponding method for DNA binding proteins, the present work indicates that RNA-binding proteins are more difficult to predict. In particular, sensitivity is more than 50% for predicting DNA-binding proteins, compared to about 35% for RNA-binding proteins. This is likely due to highly flexible and diverse RNA structures [145] compared to DNA structures. More diverse RNA structures will lead to more diverse protein structures to bind them. The latter will be more difficult to detect by structural alignment to a limited number of existing RNA-binding template structures that is similar to the number of available template structures for protein-DNA interactions.

Chapter 5 Highly accurate and high-resolution function prediction of RNA binding proteins by fold recognition and binding affinity prediction

Abstract

A full understanding of the mechanism of post-transcriptional regulation requires more than simple two-state prediction (binding or not binding) for RNA binding proteins. here we report a sequence-based technique dedicated for predicting complex structures of protein and RNA by combining fold recognition with binding affinity prediction. The method not only provides a highly accurate complex structure prediction (77% of residues are within 4 RMSD from native in average for the independent test set) but also achieves the best performing two-state binding or non-binding prediction with an accuracy of 98%, precision of 84% and Mathews correlation coefficient (Mcc) of 0.62. Moreover, it predicts binding residues with an accuracy of 84%, precision of 66% and Mcc value of 0.51. in addition, it has a success rate of 77% in predicting RNA binding types (mRNA, tRNA or rRNA). We further demonstrate that it makes more than 10% improvement either in precision or sensitivity than PSi-BLAST, Remmert2012 and our previously developed structure-based technique. This method expects to be useful for highly accurate genome-scale, high-resolution prediction of RNA-binding proteins and their complex structures

5.1 Introduction

Significant new interest in RNA-binding proteins (RBPs) are resulted from the discovery and characterization of microRNAs in post-transcriptional regulation and the implication of RBPs in many human diseases including HIV/AIDS, cancer, and neurodegenerative disorders [115]. RBPs are encoded in large number (thousands) because their diversity appears to increase during evolution of post-transcriptional machinery and the increase in number of introns. Despite of their importance, many of these RBPs are yet to be uncovered and/or characterized. Computational

prediction methods are therefore essential as initial steps for function annotation and characterization.

Function prediction for RBPs can be roughly classified into four levels of resolutions with different levels of details (low, medium, high, and the highest). The first low level of prediction is a simple two-state classification of binding or non-binding to RNA. The next medium level is the location of RNA binding residues of RBPs. A high resolution prediction is to predict the RNA type that the RBP would bind. This prediction would provide further deeper understanding of the RBP function. The highest resolution will involve the prediction of the actual binding RNA sequence and its binding complex structure with the predicted RBP.

Most computational methods developed so far attempted to detect the sequence homologous and/or evolutionary relationship between un-characterized and characterized proteins [146, 147]. The principle of these methods is that homologous sequences have the same biological function. However, less than half of identified proteins are anotated even with help of sequence homology [148]. Morover, many proteins have hidden function of RNA binding [14, 15]. Thus, it is necessary to develop sequence-based techniques that can detect function similarity in the absence of high sequence homology to known RBPs.

Several sequence-based classifiers for RBP prediction are based on support-vector machines (SVM) and limited to the low resolution prediction of binding or non-binding proteins [27–30, 116, 149]. Early studies [27, 116] did not exclude homologous sequences from training or testing. Moreover, all these techniques were trained and tested in a balanced set with equal number of positive (RBP) and negative (Non-RBP) data sets [28–30, 149]. The reported Mathews correlation coefficient value for RBP classification is 0.53 for a sequence-based SVM classifier (5-fold cross validation on 134 RNA binding and 134 non-binding proteins) [30] and 0.72 for a structure-based SVM classifier for a dataset of 76 RNA binding proteins and 246 non-nucleic-acid binding proteins (leave-one-out test) [117]. The performance of these techniques for

such a balance set likely becomes worse when applied to a real world situation where RBP is about 15% of all proteins.

Other methods make medium resolution prediction of RNA binding residues (or binding sites) directly based on either sequence-based [30, 94, 102, 118–120, 122–124] or structure-based [117, 125–129] information. The best reported values for Mathews correlation coefficient are between 0.47-0.51 [30, 127, 128]. One issue associated with these techniques is that they will predict RNA binding sites even for the proteins that do not bind RNA.

This work is inspired by our structure-based prediction of DNA and RNA binding proteins [SPOT-Struc (DNA) [32], SPOT-Struc (RNA) [34]]. We found that structural alignment to known protein-RNA complexes coupled with binding assessment with a statistical energy function based on distance-scaled finite ideal gas reference (DFIRE) state yields a highly accurate (98%) prediction of RBPs with a reasonable sensitivity of 36% and Mathews correlation coefficient (MCC) of 0.57 for a large benchmark of 212 RNA binding and 6761 non-RNA binding domains (leave-one-out cross validation). Its applications on additional APO and HOLO benchmarks and structural genomics targets yielded consistent accuracy and/or sensitivity.

This structure-based technique, however, has a limited application because the structures for the majority of proteins are unknown. The success of this structure-based technique motivates us to develop a sequence-based technique by coupling structure prediction with binding prediction, an approach proven successful for protein-DNA binding prediction [35]. Here we perform structure prediction by using the latest version of our fold recognition technique called SPARKS X [49] that are among the best performing single automatic servers in several critical assessment of structure prediction (CASP) meetings (CASP 6, CASP 7 and CASP 9 [49]). While many template-based structure prediction methods exist, the coupling between fold recognition and binding affinity prediction provides the first dedicated high-resolution function prediction for RBPs.

The new technique, called SPOT-Seq, is initially trained and validated (leave-one-out) on a dataset of 174 RNA-binding protein chains and 5778 non-binding protein chains, so that it can compare to other methods. SPOT-Seq achieves the highest MCC value of 0.61, when compared to Altschul1997, the commonly used sequence-to-profile homology search technique [134] (MCC=0.48), Remmert2012, a profile-profile fold-recognition technique based on the hidden Markov model [108] (MCC=0.50), SPARKS X fold recognition method [49] (MCC=0.57), and the structure-based prediction technique (SPOT-Struc, MCC=0.56). More than 10% improvement in either sensitivity or precision or both are observed. Further expansion of test and training sets (431 RBPs) and template libray (1164 binding domains and chains) confirms the MCC of 0.61, accuracy of 98%, precision of 78%, and sensitivity of 50%.

5.2 Methods

5.2.1 Function Prediction Protocol

The method proposed here is similar to the structure-based technique called SPOT-Struc (RNA) developed by us [34] excepted that the structure is predicted by fold recognition technique called SPARKS X [49]. The flow diagram is shown in Fig. 5.1.

First, we perform fold recognition between the target sequence and templates in the template library of RBPs by SPARKS X [49]. Our template library is built on a collection of RNA binding and non-binding proteins (see below). SPARKS X [49] attempts to match the squence profile of the target sequence (generated from Altschul1997 [134] and predicted one-dimensional structural profiles (secondary structure, solvent accessibile surface area and backbone torsion angles from SPINE X [50]) to the corresponding profiles of the template structure in the library. The sequence-structure matching score is measured by Z-Score where

$$\text{Z-score} = \frac{\text{S}_i\text{-S}_{\text{mean}}}{\sigma} \tag{5.1}$$

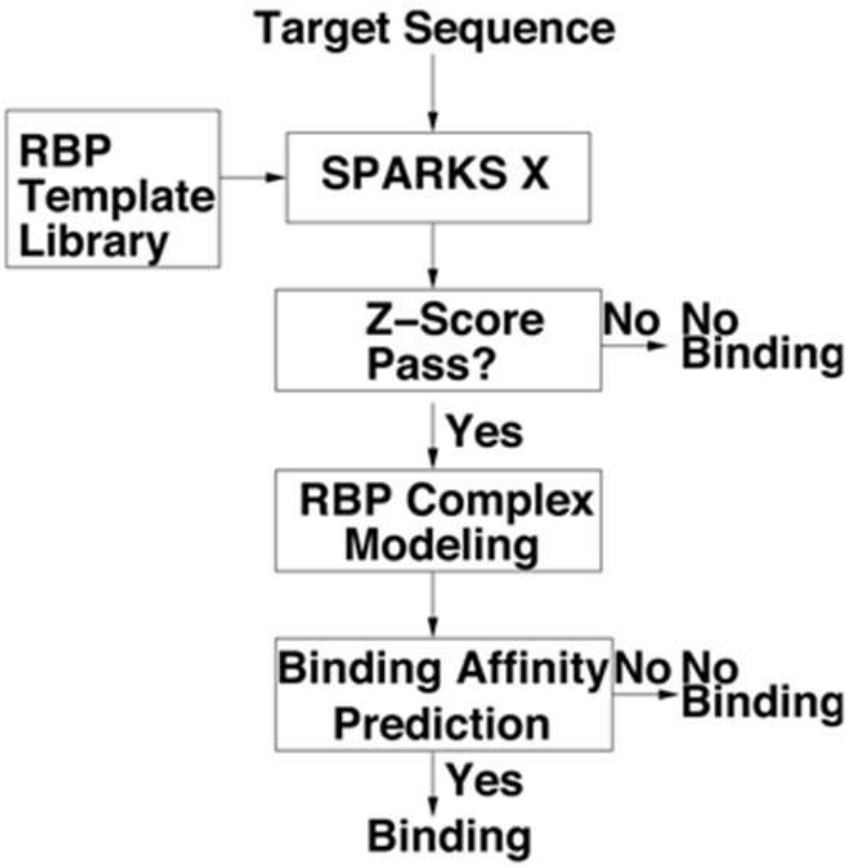

Fig. 5.1: The flow diagram of the sequence-based function prediction of RBP.

Where S_i is the alignment raw score between target and template i and S_{mean} and σ are the average raw score and the standard deviation for all templates. Typically, a Z-score of higher than 6.0 is considered as a significant template hit.

If the Z-score for any of RBP templates is higher than a threshold to be determined, a complex structure of the target protein and template RNA is built by replacing template protein sequence with target protein sequence based on the sequence-to-structure alignment from SPARKS X. For this study, the gap region is not modeled for simplicity.

Using the complex structure of model target structure and template RNA structure we can estimate the binding affinition according to a statistical energy function based on the distance-scaled finite ideal-gas reference state [33] that was extended to protein-RNA interaction (DRNA) [34]. In this work, we made no changes to the DRNA energy function. However, the binding affinity is evaluated with mainchain atoms and C_β atoms only to avoid the need to build sidechains in this initial development of the technique. If the binding affinity is higher than a to-be-determined threshold, the target protein is predicted as RNA binding and its complex structure model serves as the basis for the high-resolution prediction of RNA binding function. For convenience, we shall label our method as SPOT-Seq.

5.2.2 Template Library

For comparison, we initially employ both binding and non-binding chains are from the structure-based method SPOT-Struc. (RNA) [34]. These 225 high-resolution RNA-binding protein chains are protein-RNA complex structures (the July 2009 release). They are divided into domains according to SCOP annotations [133] or by automatic technique called DDOMAIN [132] if SCOP annotations are not available. A domain is RNA-binding if it has at least 5 amino acid residues whose heavy atoms are within 4.5Å from any heavy atoms of nucleotide functional groups. Redundancy in resulting domains is removed by using BLASTClust with 95% sequence identity cutoff [134]. This leads to 255 domains as binding templates. To increase sensitity,

both original chains and domains are included in our template library and lead to a final template library of 355 RNA-binding protein structural templates (RB-T355). Non-binding templates are from the nonbinding protein-domain sets of 6761 domains obtained previously based on 25% sequence identity cutoff [34]. We only include the original chains into the template library with a 25% sequence identity cutoff. The final number of templates after a 25% sequence identity cutoff is 5765 (NB-C5765).

5.2.3 Cross Validation Datasets

RB-C174 and NB-C5765: We built a leave-one-out cross-validation data set of RNA-binding sequences by removing redundant sequences of all sequences contained in RB-T355 with BLASTClust [134] at a sequence identity 25% cutoff. A total of 174 sequences (RB-C174) remained. Only full chains (not domains) (RB-C174 for positive and NB-C5765 for negative sets) are employed for cross validation.

5.2.4 Expanded Template Library and Independent Test Set

The above template library was based on high-resoultion X-ray structure (3Å or less) on July 2009. To examine the effect for an expanded template library and provide an independent test set, we downloaded all pdb structures that contains RNA on April 1, 2011. After removing the structures contained in the template library, we obtained 1027 complex structures that are separated into chains and domains according to SCOP or DDMAIN classifications. After removing domains with less than 60 residues or having less than 5 binding residues and redundant domains with more than 95% sequence identity by BLASTClust, we obtained 612 domains in addition to 255 domains previously obtained. Both domains and their respective chains are included in our new expanded library with a total of 1164 templates including RB-T355. We shall label this library as RB-T1164.

There are a total of 566 chains contained in the new template library. These sequences are clustered with BLASTClust at a sequence identity of 25% cutoff among themselves and the squences contained in RB-C174. This leads to an independent

test dataset of 257 chains (RB-IC257). However, this independent test set cannot be considered as representative because it contains both high and low resolution structures. Thus, we randomly divide RB-C174 and RB-IC257 into two equal sets of 216 and 215 chains, respectively (RB-C216 and RB-C215). One will be used for final training and one for final testing.

5.2.5 Performance Evaluation

The performance of the method is evaluated by sensitivity [$SN = TP/(TP + FN)$], specificity [$SP = TN/(TN + FP)$], accuracy [$AC = (TP + TN)/(TP + FN + TN + FP)$], precision [$PR = TP/(TP + FP)$], and Matthews correlation coefficient (MCC)

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \tag{5.2}$$

Here, TP, TN, FP, FN refer to true positives, true negtives, false positives and false negtives, respectively. A MCC value provides an overall assessment of the method performance with 1 for perfect agreement. One should note that sensitive can also be called as coverage of true positive prediction while precision is fraction of corrected predictions in all positive predictions.

5.2.6 Other Methods and Threshold Optimizations

PSI-BLAST is employed for searching homologous sequences by searching against the NCBI non-redundant sequence library for four iterations. If a target has at least one template from RB-T355 with an E-value lower than a to-be-determined threshold, the target is considered as a RNA-binding protein. Any templates having $>$30% sequence identity with the target sequence is removed. The threshold is optimized by maximizing the MCC value.

SPARKS X is a method without the steps for building the complex structure and prediction of binding affinity in Fig. 5.1. Z-Score threshold, optimized by maximizing the MCC value, is 7.

To assess the ability to detect RNA-binding proteins of SPARKS X, relative to other fold-recognition methods, we employed Remmert2012 as an example because it is one of the best fold-recognition techniques in CASP [108]. Remmert2012 version 1.5.1 was downloaded from http://toolkit.tuebingen.mpg.de/Remmert2012/. PSSM generated from Altschul1997 were used to search NR database to generate multiple sequence alignment and profiles. Default parameters, options and scripts were used to generate HMM profiles for both targets and template proteins. We also tested the option '-mact' and results are essentially the same. Probablity was used as a significant score in the prediction.

Two thresholds of Z-score and binding affinity for SPOT-Seq (i.e. SPARKS X+DRNA) are optimized by a grid-based search for the highest MCC value. The grid is 0.1 for Z-score. The binding affinity threshold is obtained by considering the lowest energy value at different Z-scores of a given target. For the prediction of RNA-binding proteins, the Z-score threshold is 6.6 and the energy threshold is -0.28. For the expanded template library (RB-T1164), the Z-score threshold is 7.0 and the energy threshold is -0.57, respectively. This was optimized based on the dataset of RB-C174 and NB-C5778. A larger template library leads to stricter Z-score and energy thresholds to prevent false positives, as expected. The same thresholds are applied to independent test set of RB-IC257.

5.3 Results

5.3.1 Low Resolution Two-State Prediction

Leave-one-out cross validation. Fig. 9.2 compares the performance of PSI-BLAST [134], fold recognition method Remmert2012 [108], SPARKS X [49], structure-based method SPOT-Struc (RNA) [34] and SPOT-Seq. from this work by the leave-one-out cross validation. The results are also quantitatively summaried in Table 5.1 based on

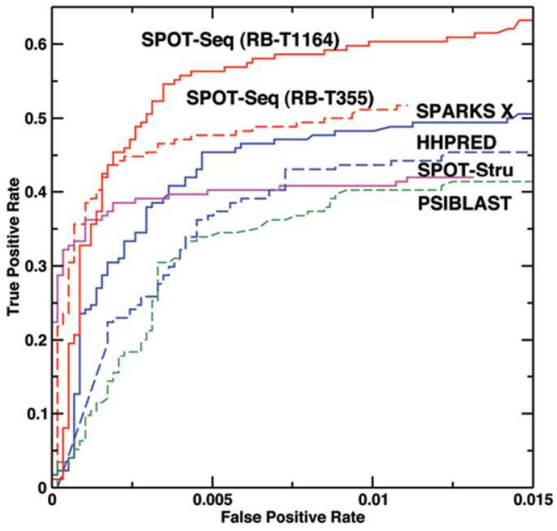

Fig. 5.2: True positive rate versus false positive rate as given by Altschul1997 (Green, dashed line), SPOT-Struc (Magenta), Remmert2012 (Blue, dashed line), SPARKS X (Blue, Solid line), and SPOT-Seq. (Red, dashed line for the RB-T355 template library and solid line for the RB-T1164 template library) for the low-resolution two-state prediction (binding vs. no binding).

Table 5.1: Methods comparison for predicting of RNA-binding proteins

Method	Sensitivity	Accuracy	Precision	MCC
PSI-BLAST[a]	33%	98%	70%	0.48
Remmert2012[a]	44%	97%	60%	0.50
SPARKS X[a]	45%	98%	75%	0.57
SPOT-Struc (RNA)[b]	35%	98%	94%	0.56
SPOT-Seq (this work)[a]	45%	98%	85%	0.61

[a] Sequence-based method. [b] Structure-based method.

thresholds optimized for the highest Mathews correlation coefficient. These results are obtained by taking one chain sequence from either RB-C174 or NB-C5765 and predicting whether it binds or does not bind to RNA. This large unbalanced dataset with 3% binding sequences is employed to mimic real situation where binding proteins are a minor portion of all proteins. Table 5.1 indicates that SPARKS X improves 12% over PSI-BLAST in sensitity and 5% in precision with similar level of accuracy. On the other hand, SPARKS X improves over Remmert2012 mostly in precision (15%) at similar level of sensitivity and accuracy. without significant changes in accuracy and precision over Remmert2012 [108]. The structure-based technique (SPOT-Struc), although has a much higher precision than the fold-recognition technique (SPARKS X) (94% versus 75%), but with a significantly lower sensitivity (35% versus 45%). This reflects the results obtained by optimizing MCC values. Introduction of binding affinity prediction further improves the precision from 75% in SPARKS X to 85% in SPOT-Seq without much change in sensitity or accuracy.

Table 5.2: Examination of 44 SCOP folds shared by both RNA-binding (RB-C174) and nonbinding (NB-C5765) proteins.

SCOP Fold ID	Dataset (RB/NB)	SPARKS X (RB/NB)	SPOT-Seq (RB/NB)
d.58	14/70	4/1	11/1
b.40	11/39	2/0	1/0
c.26	9/18	8/0	7/0
a.4	9/96	1/2	2/4
b.34	8/21	2/0	2/0
g.41	6/5	1/0	1/0
d.104	6/1	3/0	6/0
c.55	6/62	3/3	4/0
e.8	5/3	2/0	2/0
d.79	5/10	1/0	2/0
d.50	4/3	2/0	3/0
b.121	4/26	0/0	0/0
d.52	3/5	0/0	0/0
d.41	3/5	2/0	3/0
d.14	3/10	0/0	0/0
b.43	3/10	2/0	2/0
a.2	3/13	1/0	1/0
g.39	2/10	0/0	0/0
d.67	2/1	0/0	0/0
d.218	2/7	0/0	0/0
c.51	2/6	0/0	0/0
b.122	2/3	2/0	2/0
a.118	2/40	0/0	0/0
d.157	1/8	0/0	0/0
d.1	1/2	0/1	0/0
c.97	1/6	0/2	0/0
c.9	1/1	0/0	0/0
c.66	1/27	1/1	1/1
c.62	1/5	0/0	0/0
c.52	1/17	0/0	0/0
c.37	1/70	0/1	0/2
c.23	1/41	0/0	0/0
c.1	1/136	0/0	0/0
b.82	1/28	0/0	0/0
b.46	1/1	0/0	0/0
b.44	1/1	0/0	0/0
b.38	1/4	0/0	0/0
b.2	1/23	0/1	0/0
a.7	1/20	0/0	0/0
a.30	1/3	0/0	0/0
a.160	1/1	0/0	0/0
a.156	1/2	0/0	0/0
a.144	1/1	0/0	0/0
a.137	1/7	0/0	0/0
Total	134/861	37/12	50/8

Discriminating binding from non-binding within the same fold. According to the Structural Classification of Proteins (SCOP) [133], there are 44 folds shared by both RNA-binding and non-RNA-binding proteins in RB-C174 and NB-C5765. As shown in Table 5.2, the majority (849/861) non-RNA binding proteins are filtered by SPARKS X while SPOT-Seq further reduces the number of false positives from 12 to 8 and leads to a very low false positive rate of 0.9%. At the meantime, SPOT-Seq increases the true positive rate to 37% (50/134) from 28% (37/134) given by SPARKS X. The result confirms that both fold recognition technique and energy calculation contributes to the power of distiguishing the RNA-binding proteins from non-binding one even within the same fold.

5.3.2 Medium Resolution Binding-Residue Prediction

Predicted binding complexes between a target and a template RNA allow us to infer RNA binding residues for the target. We define an amino-acid residue as RNA-binding if any heavy atoms of the residue are less than 4.5 Å away from any heavy atoms of a RNA base. For a few proteins, we found that it is necessary to perform crystal symmetry operation to yield correct information on binding residues. We examine the accuracy of binding-residue prediction by focusing on true positive prediction of 78 proteins from the leave-one-out test on RB-C174/NB-C5765. Compared to native binding residues, we achieved 53% in sensitivity, 85% in accuracy, and 63% in precision. The MCC value is 0.47. This value is significantly lower than 0.72, the MCC value given by SPOT-Struc. This suggests that structural alignment allows a better detection of RNA binding regions than model complex structures, predicted by SPARKS X due to inaccuracy of models predicted. In other words, SPARKS X improves over SPOT-Struc in sensitity of detecting RNA-binding proteins (low resolution prediction) while reducing the accuracy of predicting binding regions (medium resolution prediction). Fig. 5.3 displays 78 MCC values (open circles) for the predicted binding residues as a function of Z-score. Clearly, there is a trend that higher Z-scores (high confidence in the accuracy

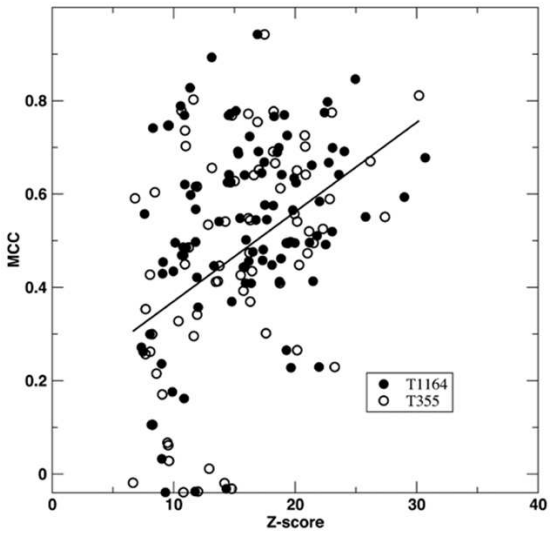

Fig. 5.3: Medium resolution prediction of RNA-binding sites. MCC values for predicted RNA-binding residues are shown as a function of fold recognition Z-scores. Results of RB-C174 tested on small and expanded template libraries of RB-T355 (open circles) and RB-T1164 (closed circles) are shown. The line from linear regression is employed to illustrate the trend.

for the model structure) leads to higher MCC values. However, there exist a few proteins with poorly predicted binding regions when Z-score <15

Fig. 5.4 shows two examples: one with a reasonable prediction of binding residues but the other with a poor prediction. For the human Rnase H1 (target 2qk9A,Fig. 5.4A), predicted (orange) and actural (magenta) RNA structures are located in similar locations, the predicted binding region (in Blue) is also close to the native binding region (in Red). The MCC value for the predicted binding residues is 0.65 with a sensitivity of 97% and an accuracy of 93%. However, the predicted and actual RNA structures for the target A. fulgidus Piwi protein (PDB ID# 1ytuB, Fig. 5.4B) are different. The native structure binds with double helix RNA and the binding residues are represented as red, but the predicted structure based on the template (3f73A) binds with a single strand RNA that only partly overlaps with native RNA strucutre. This leads to wrongly predicted binding residues (in blue). This is likely caused by the fact that predicted protein structure (green) for 1ytuB is only a part of the actual native structure.

5.3.3 High resolution prediction of binding RNA types.

The next resolution level of function prediction is to predict the types of RNA that binds to the target protein. We manually classified the types of RNA included in our template library, according to the annotation of DAVID [150] . In the template library (RBT-355), 272 are annotated into 5 types of RNA-binding proteins. There are 189

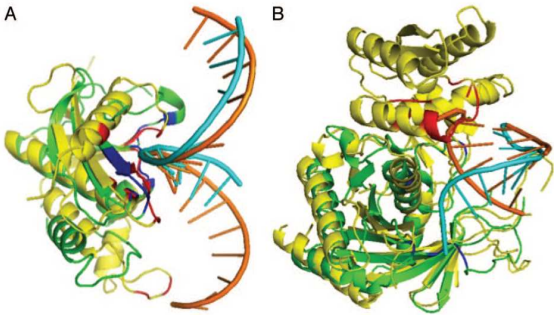

Fig. 5.4: Comparison between the predicted (green) and actual (yellow) complex structure for the target 2qk9A with RNA structures colored in cyan for predicted and orange for native RNA structure and binding regions colored in Red for native structure and Blue for predicted structure. (A) Target 2qk9A predicted with template 1zbiB (sequence identity between them is 13%). (B) target 1ytuB predicted with template 3f73A3 (sequence identity between them is 2.0%.)

Table 5.3: Mis-predicted binding types for tRNA, mRNA and rRNA-binding proteins.

Native tRNA	Pred. Type	Native mRNA	Pred. Type	Native rRNA	Pred. Type
1jj2U	rRNA	1yz9A	-	1mzpA	-
1mzpA	-	2gxbB	-	1yz9A	-
1ytyA	mRNA	2ozbA	tRNA	2bh2A	tRNA
2i82A	rRNA	2rfkA	tRNA		
3bt7A	rRNA				

binding with tRNA, 148 binding with rRNA, 47 binding with mRNA, 25 binding with synthetic RNA and 7 binding with SRP RNA. Because some RNAs have more than one function, the total number of invovled protein is less than the number of RNAs grouped according to function.

The ability of our method to predict the type of binding RNA is examined by analyzing 78 true positives (RNA-binding domains). These 78 RNA-binding domains contain 48 tRNA-binding proteins, 34 rRNA-binding proteins, 10 mRNA-binding protiens, 3 synthetic RNA-binding protein, 3 SRP RNA-binding proteins. If we use the template RNA in the predicted complex structure to predict the binding RNA type for the target protein, we achieve success rates of 90% (43/48) for tRNA, 91% (31/34) for rRNA and 70% (7/10) for mRNA. Table 5.3 listed all mis-predicted RNA types. They are betweeen tRNA, rRNA and mRNA.

5.3.4 The highest resolution: Protein-RNA Complex Structure

To examine the quality of predicted structures, we used TM-Score from TM-Align [139] to compare the native and predicted structures which is 1 for perfect agreement and about 0.2 between two random structures. For 78 correctly predicted targets, the average TM-Score is 0.73. One can also measure the structure similarity by the fraction of residues in model structure has an root-mean-squred distance (RMSD) of 4Å or less. We found that the medium value is 72%. We found that one structure for the target 2j035 (50S ribosomal protein L31) was predicted poorly (TM-Score$<$0.4 and only 22%residues has RMSD $<$ 4Å). This large error in predicted structure is caused by the non-globular shape of the native strucure (a small 59 residue protein with a radius of gyrate 23.4Å). We further found that the structural accuracy of binding regions are higher than that of whole proteins. For example, the binding regions of 15 targets have more than 95% residues with RMSD$<$5Å. By comparison, only 8 targets satisfies the same criterion for the whole protein.

As an illustrative example, Fig. 5.4A showed the predicted and actual complex structures with RNA for target 2qk9A (human Rnase H1). The template 1zbiB (Bacillus halodurans RNase H catalytic domain) was located with a Z-score of 18.0 and the binding energy of -1.62. In this example, 50% aligned residues of native structure and predicted structure has RMSD $<$ 4Å, much lower than the medium value of 72%. This is largely due to a helix near binding region in the template, but only a coil in the native structure. Yet, the binding region is reasonably accurately modeled based on the proximity of blue and red colors (a MCC value of 0.65, a sensitivity of 97% and an accuracy of 93%). This remote homologous template is identified despite of a low sequence identity of 13.3%. In this example, the conformation of RNA is also modeled correctly. For Fig. 5.4B, the only part of the target *A. fulgidus* Piwi protein (PDB ID# 1ytuB) is predicted. This part was predicted with a TM-Score of 0.75. The sequence identity between the target and template (3f73A3) is 2.0%.

5.3.5 Discrimination against DNA binding proteins

We tested the ability of SPOT-Seq for separating RNA and DNA binding proteins by applying the method to the dataset of 250 non-redundant DNA binding proteins (DBPs) collected by us previously [32] We employed the thresholds for Z-score and binding affinity obtained by optimizing the MCC value for RB-C174/NB-C5765. Only 5 out of 250 DBPs are predicted as RBPs. Among these 5 predicted RBPs, four have high sequence identity (>77%) with the templates and known for binding with both RNA and DNA. The remaining target 1sfuA (the viral Zalpha domain [151]) is a remote homolog of the template 2gxbB with a sequence idenitity of 27.1%. 1sfuA was also predicted as RNA-binding proteins in previous structure-based study [34]. This poxvirus protein is E3L protein that has a Z-alpha motif similar with ADAR1 (double-stranded RNA adenosine deaminase) which is known to bind with Z-RNA [152, 153]. Thus, there is no false positive from DNA-binding proteins.

5.3.6 Effect of the Expanded Template Library

Table 5.4 examines the effect of the expanded template library at all four levels of prediction resolutions. It is clear that expanding template library from 355 to 1164 protein domains and chains improves sensitivity from 46% to 56% at the expense of reducing precision from 85% to 81%. The effect on the ROC curve can be found in Fig. 9.2. SPOT-Seq with RB-T355 has a higher sensitivity (or true positive rate) only at a very low false positive rate (<0.2%) while SPOT-Seq with RB-T1164 has a higher sensitity at low to moderate false positive rates. The overall MCC value increases from 0.61 to 0.66 due to the expanded library.

For binding residue prediction, expanding templates improve both precision (from 63% to 69%) and sensitivity (from 53% to 60%) with change to accuracy. This leads to an improved MCC value from 0.47 to 0.53. Fig. 5.3 compared the MCC values as a function of Z-Score given by different template libraries. Expanding templates reduce the number of poorly predicted binding regions (MCC<0.2) from 10 to 7.

Table 5.4: SPOT-Seq performance for an expanded template library and an independent test

Resolution	T355[a]	T1164[a]			
Level	C174[b]	C174[b]	C257[b]	C216[b]	C215[b]
Two-state[c]					
MCC	0.61	0.67	0.60	0.62	0.62
Accuracy	98%	98%	97%	98%	98%
Precision	**86**%	81%	82%	84%	84%
Sensitivity	45%	56%	45%	48%	47%
Binding Residue[d]					
MCC	0.47	0.53	0.48	0.50	0.51
Accuracy	85%	85%	83%	84%	84%
Precision	63%	69%	63%	66%	66%
Sensitivity	53%	60%	58%	59%	60%
RNA-type[e]					
tRNA	90%	67%	67%	62%	69%
	(43/48)	(46/69)	(24/36)	(33/53)	(33/48)
mRNA	70%	82%	62%	73%	56%
	(7/10)	(9/11)	(24/39)	(16/22)	(15/27)
rRNA	91%	92%	91%	91%	96%
	(31/34)	(48/52)	(61/67)	(52/57)	(54/56)
Complex Structure[f]					
TM-Score	0.73	0.69	0.66	0.66	0.66
RMSD(<4)	72%	78%	76%	76%	77%
# (Whole)	19%	16%	17%	15%	17%
#(Binding)	10%	33%	25%	29%	25%

[a] The template sets of 355 and 1164 RBPs, respectively. [b] The target sets of C174 for training and cross validation, C257 for independent test. C174 and C257 are further randomly separated into C216 for training and cross validation and C215 for independent test. [c] Performance on low-resolution two-state prediction based on Mathews correlation coefficient and others. [d] Performance on medium-resolution prediction of RNA binding residues based on Mathews correlation coefficient and others. [e] Success rate of the high resolution prediction of bound RNA types (tRNA, mRNA and rRNA): the fraction of correctly predicted RNA binding types in actual number of proteins in that type. [f] The highest resolution of complex structure prediction based on the average strutcural similarity score (TM-Score), medium value for the percentage of aligned residues in the model structure with RMSD < 4Åfrom the native structure, percentage of targets with 95% predicted residues within RMSD <5 Åfrom the native residues for the whole protein and binding regions only.

The effect of the enlarged template library on prediction of RNA types is mixed. There is a reduction of success rate from 90% (43/48) to 67% (46/69) for tRNA, improved success rate from 70% (7/10) to 82% (9/11) and unchanged success rate [91% (31/34) versus 92% (48/52)]. This large fluctuation suggests that the dataset may be too small to assess the accuracy of RNA type prediction.

We further examined the prediciton ability on the highest resolution of protein-RNA complexes. We found that the average TM-score is reduced from 0.73 to 0.69 while the medium value for the fraction of residues increases from 72% to 78%. This somewhat conflict result reveals the difficulty to consistently assess the quality of predicted structures.

5.3.7 Independent Test on RB-IC257

Table 5.4 also displays the results of independent test on RB-IC257 based on the thresholds generated by the cross validation set of RB-C174/NB-C5765 with the template library of RB-T1164. Overall speaking, there is a somewhat reduction of performance in the two-state prediction (the MCC value reduced from 0.65 to 0.59). The most reduction is in the sensitivity from 56% to 45%. This reduction of sensitivity is somewhat expected because the RB-IC257 set contains low resolution X-ray structures and NMR structures. The performance of binding residue prediction for the independent test set is also reduced in accuracy (2%), precision (6%) and sensitivity (2%). The accuracy of predicted complex structures also decreases somewhat (TM-Score from 0.69 to 0.66 and the fraction of residues with RMSD$<$4Å from 78% to 76%. We hypothesis that the poorer performance for RB-IC257 may be because it was complited by including low resolution X-ray structures, EM structures, and NMR structures and recently solved structures.

To verify this hypothesis, we randomly divided to RB-IC257 and RB-C174 into two independent sets of RB-C216 and RB-C215. We first employed RB-C216/T1164 to train the thresholds and found that these thresholds are identical to those trained by RB-C174/T1164. Then, we tested these thresholds to RB-C215. The results are shown

in Table 5.4 . Indeed, we found that the result on RB-C216 and RB-C215 are essentially the same with MCC values for the two-state prediction at 0.61 and 0.62, respectively.

5.4 Discussions

In this paper, we describe the first technique that provides prediction of RNA binding proteins at all four levels of resolution. At the low resolution level of two-state prediction, its MCC value based on a large dataset of 216 binding proteins (or independent 215 binding proteins) and 5765 nonbinding proteins is 0.62 (0.62). This value is higher than 0.53, the best reported, sequence-based SVM classifier method (5-fold cross validation on 134 RNA binding and 134 non-binding proteins only) [30]. Its MCC values for the medium resolution prediction of RNA-binding residues [0.50 (0.51)] for RB-C216 (RB-C215) sets are for comparable to 0.47 given by the same SVM classifier [30]. More importantly, the high-resolution prediction of binding RNA types and binding complex structures are highly reliable. The success rates are 62% (69%) for tRNA, 91% (96%) for rRNA and 73% (56%) for mRNA for the same two sets, respectively. The average TM-score for predicted structures are 0.66 (0.66). One important feature of SPOT-seq is its ability to separate RNA from DNA binding proteins. It yields zero false positions when applied to 250 DNA binding proteins.

We would like to emphasize that we have purposely tested and trained SPOT-seq in entire chains of proteins, rather than protein domains. This is to mimic the real-world situation that in most cases, protein domain boundaries are unknown. SPOT-seq will allow direct identification of RNA-binding domains from the target chain as it searches for the best matching domain and/or chain from the template libarary.

SPOT-seq has one obvious limitation. It relies on the availability of protein-RNA complexes as templates. It will not be able to predict RNA-binding proteins whose structures do not have a template in the template library or when its template in the library is difficult to recognize. We have used the RB-T355 libary that includes both domains and chains with 95% sequence-identity cutoff for the purpose of maximizing available templates. The low sensitivity (46%) is in part due to lack of structurally

matching templates. Although expanding the number of templates from T355 to T1164 improves sensitivity, it reduces precision at the same time because a low resolution RBP structure will more likely make a false match to a non-binding structure. More importantly, tripling the number of templates from 355 to 1164 does not expand the structural space as much. For example, In the RB-IC257 set, there are 141 false negatives that have 52 targets with TM-score $>$0.5 to the structures in T355. The number of structurally similar templates only increases by 24 to 76 targets when the number of templates expands to 1164. It is clear that significantly more high-quality complex structures of protein-RNA are needed with the current method in order to further advance the sensitivity and precision at the same time.

The final precision of 81% based on optimized MCC values is likely a upbound when applying to a genome because our test and validation set contains significantly less binding proteins (216/5765 or 3.7%) than in a typical genome (15%). In fact, for the entire set of nonredudant set of (216+215) RBPs or 7.5% of nonbinding ones, the precision is 91% with the same number of false positive proteins. Thus, we expect that application of our method for genome-wide prediction will lead to highly accurate useful results.

Finally, one important advantage of this SPOT technique is its reasonable speed. For example, it only takes 1107 CPU hours (46 days) on a single processor PC to scan about 7380 genes in yeast genomes. We will report these results in a separate paper. A freely available, easy to use webservers is available for academic users at http://sparks-lab.org.

Chapter 6 Charting the unexplored RNA-binding protein atlas of the human genome by combining structure and binding predictions

Abstract

Detecting protein-RNA interactions is challenging both experimentally and computationally because RNAs are large in number, diverse in cellular location and function, and flexible in structure. As a result, many RNA-binding proteins (RBPs) remain to be identified. Here, we applied the RBP-prediction method SPOT-Seq to the human genome. In addition to cover 42.6% of 1,217 known RBPs annotated in the Gene Ontology (GO) database, SPOT-Seq detects 2,418 novel RBPs, 48% of which are poorly annotated in the GO database. The majority (98%) of the remaining predicted novel RBPs shared specific GO molecular function terms with known RBPs such as DNA binding and zinc ion binding. The results of SPOT-Seq were independently tested by a recent proteomic experimental discovery of 860 mRNA binding proteins (mRBPs). We achieved the coverage (or sensitivity) of 43.6% for human mRBPs, similar to 42.6% for all RBPs. In particular, 291 predicted novel proteins (in 2418) were validated by this mRBP set and the majority (70%) were predicted as mRNA binding. In a more stringent set of 315 previously unknown RBPs in 860 mRBPs that excluded homology-inferred RBPs and any proteins annotated with a keyword RNA (not just RNA binding), 19% proteins are predicted novel RBPs. This confirms the ability of SPOT-seq to go beyond homology-based bioinformatics tools and uncover truly novel RBPs. Further analysis indicates that predicted, novel RBPs play important phenotypic roles in disease pathways and their mutations can cause diseases. The dataset of 2418 predicted novel RBPs along with their predicted confidence levels and protein-RNA complex structures is available at http://sparks-lab.org for further experimental validation and hypothesis generation.

6.1 Background

A comprehensive understanding of cellular processes requires identification of RNA-binding proteins (RBP) as well as their ligands. Identification of RBPs is of significant interest because numerous studies have shown that they are key factors associated with cellular processes such as cell cycle checkpoints and genomic stability and mutations in RBPs are linked to human diseases, including cancer [115] . Recent global analysis indicates that transcripts are not only large in number, but also diverse in localization and function in cells [154–156] . This implies that underlying post-transcriptional networks are likely larger and more complex than either transcriptional networks or protein-protein interaction networks [157]. However, experimental determination of RNA-binding by every protein is inefficient and impractical, as well as technically challenging and expensive. Attempts at high-throughput biochemical approaches for identifying RBPs progress slowly and are fraught with inaccuracy [157–159]. Thus, computational methods [27–30, 34, 36, 116, 148, 149] have become a critical component for function annotation and analysis of RBPs.

Recently, we have developed a template-based technique called SPOT-Seq (RNA) that makes sequence-based prediction of RBPs [36] . In this method, a query sequence is first threaded onto the template structures of proteins by the fold recognition technique called SPARKS X [49]. The template library contains 1,164 known protein-RNA complex structures on both domain and protein chain levels (95% sequence identity or less). If one of the templates has a good match (according to Z-score) to the query, the structure for the query is predicted and a model complex structure between the predicted structure and the RNA from the template is built. The model complex structure is then employed to predict affinity for protein-RNA-binding using a knowledge-based energy function [34] . If the binding affinity is higher than a threshold, an RBP is predicted. The method achieves a precision of 84% and sensitivity of 47% for a test set of 215 RBPs and 5,765 nonbinding proteins. The

precision and sensitivity of SPOT-Seq are more than 10% higher than those given by the sequence-to-profile homology search technique PSI-BLAST [134]. More importantly, unlike some computational methods, SPOT-Seq (RNA) can distinguish DNA-binding from RNA binding (zero false positives when applied to 250 DNA binding proteins).

Here, we made a large-scale prediction of RBPs in human genome using SPOT-Seq and discovered 2,418 novel RBPs in addition to recover 519 known RBPs. Among these predicted novel RBPs, 1848 proteins possess GO annotations other than RNA-binding, more than 90% of which are associated with known RNA-binding proteins. We further showed that some of these predicted novel RBPs involve in various disease pathways and associated with disease-causing SNPs. More importantly, a large subset of predicted novel RBPs (291 proteins, 12%) are confirmed by a recently published proteomic study limited to mRNA binding proteins (mRBPs) [17]. Similar sensitivity (42.6% for annotated RBPs in human genome and 43.6% for all mRBPs from the proteomic study) confirms that SPOT-Seq can make consistent and accurate detection of RBPs.

6.2 Materials and Methods

Fold-recognition and binding-affinity based prediction by SPOT-Seq. SPOT-Seq [36] is a method that combines fold recognition and binding affinity prediction for RBP prediction. Each target sequence is aligned to the structures in a template library of 1,164 non-redundant protein-RNA complex structures (both domains and chains with 95% sequence identity cutoff) by employing the fold recognition method SPARKS X [49]. If the Z-score of the fold recognition is greater than 8.04, a model complex structure between the target protein and template RNA is built by replacing template protein sequence with target protein sequence based on the sequence-to-structure alignment generated from SPARKS X. The model complex structure is then employed to estimate binding affinity according to a statistical energy function based on the distance-scaled finite ideal-gas reference state [33] that was extended to protein-RNA interaction (DRNA) [34]. If the predicted threshold is lower than -0.57, the target

protein is predicted as RNA-binding and its complex structure model serves as the basis for the high-resolution prediction of RNA-binding function. The energy and Z-score thresholds were obtained by optimizing the Mathews correlation coefficient (MCC) based on the leave-homolog-out cross validation with a dataset of 216 RBPs and 5765 nonRNA-binding proteins.

6.3 Results

6.3.1 Application of SPOT-Seq to human genome

The human genome dataset from the Uniprot database contains 20,270 unique proteins [74] . The annotations of these genes are obtained from the GO database [160]. We broadly define a protein as a RNA-binding protein (RBP) if its annotation contains any of the keywords (RNA binding, ribosomal, ribonuclease, or ribonucleoprotein). For the protein with keywords RNA polymerase, we limited to 16 specific GO terms as RNA-binding proteins (see Table 6.1). This definition leads to 1,217 (6%) proteins annotated as RNA-binding while 15,595 proteins are annotated with other functions and 3,458 are not annotated (unknown function). Table 1 lists the number of proteins found according to the keywords used. Although this definition of RNA binding proteins is subjected to annotation errors/omissions and choices of keywords, it provides a useful reference for analyzing our predicted RBPs.

Application of SPOT-Seq to human genome identified 2,937 proteins as RNA-binding after removing those proteins whose predicted structures have overlap with predicted trans-membrane regions by THUMBUP [161]. This filter is necessary because our method based on protein-RNA complex structures cannot predict the structures of trans-membrane proteins. Among 2,937 predicted RBPs, 519 proteins were annotated as RNA-binding and belong to one of the keyword classes shown in Table 6.1. In addition 1,848 proteins were annotated with functions other than RNA-binding and 570 proteins lack annotations. Fig. 6.1 shows a pie diagram for comparing fractions occupied by predicted RBPs in annotated RBPs, unknown proteins and proteins with other functions. The result reveals sensitivity (or coverage)

Table 6.1: The number of annotated RBPs according to keywords, compared to the number of proteins predicted as RBPs by SPOT-seq

Keywords	# of annoated	#of predicted	#Converage(%)
RNA binding	722	402	56
ribosomal	68	37	54
ribonucleoprotein	240	52	22
ribonuclease	67	12	18
RNA polymerase	120	16	13
Total	1,217	519	43

GO IDs related with RNA polymerase: GO:0000428: DNA-directed RNA polymerase complex; GO:0003899: DNA - directed RNA polymerase activity; GO:0003968:RNA -directed RNA polymerase activity; GO:0005665:DNA -directed RNA polymerase II; GO:0005666: DNA -directed RNA polymerase III; GO:0005736:DNA -directed RNA polymerase I complex; GO:0006368:RNA elongation from RNA polymerase II promoter; GO:0006369: termination of RNA polymerase II transcription; GO: 0016591:DNA -directed RNA polymerase II; 0030880 RNA polymerase complex;GO:0031379:RNA -directed RNA polymerase complex;GO:0031380:nuclear RNA -directed RNA polymerase complex;GO:0034062:RNA polymerase activity;GO:0042789:mRNA transcription from RNA polymerase II promoter;GO:0042795:snRNA transcription from RNA polymerase II promoter;GO:0042796:snRNA transcription from RNA polymerase III promoter; GO:0042797:tRNA transcription from RNA polymerase III promoter

of 42.6% (519/1,217). This sensitivity is consistent with 47% sensitivity from our benchmark study [36] despite that the latter was based on proteins with experimentally solved protein-RNA complex structures only. We noted that the sensitivity strongly depends on specific categories of RBPs. The sensitivity is the highest at 56% for the proteins annotated with the keyword of RNA binding and lowest at 13% with the keyword of RNA polymerase. Table 6.2lists top 10 templates employed for all predicted RBPs for human genome. The 60S ribosomal protein L3, RPL3 (chain C in pdb structure 3o58), is responsible for predicting 1181 proteins with 61 annotated as RNA binding. Four other 60S ribosomal proteins are also in the top 10 list. The surprising popular employment of RPL3 leads us to examine the accuracy associated with prediction based on 3o58. SPOT-seq was tested by 215 RNA-binding proteins and 5,765 non-RNA-binding proteins [36] . Among these proteins, 11 binding proteins and 15 non-binding targets employed protein chains contained in structure 3o58 as templates. Six are true positives and 0 are false positives based on the default thresholds. The Mathews correlation coefficient (MCC) for the use of 3o58 chains as templates

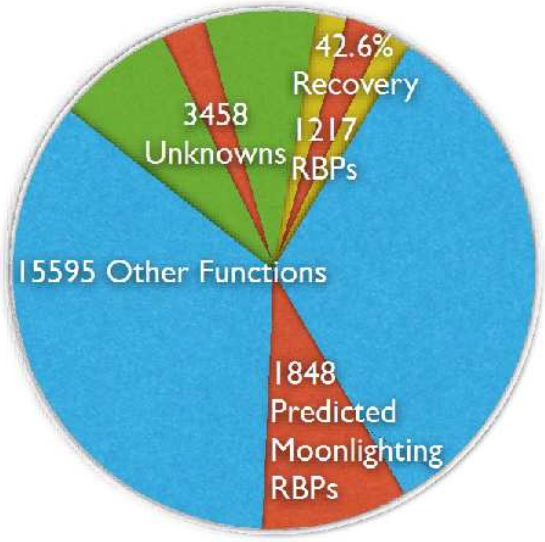

Fig. 6.1: A pie diagram for annotated RBPs (green), unknown proteins (yellow) and proteins with other functions (blue). All three regions contain predicted RBPs (in red) in significant fractions.

Table 6.2: Top 10 templates employed for all predicted human RBPs.

PDB ID	Gene Name	Protein Name	#Proteins (#A nnotated)	#Nonredudant
3o58C	RPL 3	60S ribosomal protein L3	1181(61)	835
1hvuA	gag-pol	Gag-Pol polyprotein	223(12)	177
3o58E	RPL5	60S ribosomal protein L5	180(10)	150
3ciyB	Tlr3	Toll-like receptor 3	149(2)	54
3o58F	RPL6A	60S ribosomal protein L6A	123(6)	114
3ivkB			112(0)	17
3a6pA	X PO5	Exportin-5	98(5)	91
3o58b	RPL32	60S ribosomal protein L32	90(5)	82
3o58T	RPL21A	60S ribosomal protein L21A	95(8)	60
1cvjA	PABPC1	Polyadenylate-binding protein 1	58(50)	41

is 0.64, similar to the overall MCC value of 0.62 when all templates are employed. Thus, the performance for prediction based on 3o58 chains is consistent with the overall performance.

6.3.2 Molecular functions related to 1848 moonlighting RNA-binding proteins

There are 1,848 predicted novel RBPs were annotated with functions other than RNA-binding. These proteins perform a moonlighting role of RNA-binding. We assess our predicted moonlighting RBPs by their shared molecular functions with known RBPs. In Table 6.3, we tabulate number of proteins and GO terms in molecular function that are unique or shared between predicted and annotated RBPs. More than 90% of predicted novel proteins [91%, 226/(226+21) for proteins with root annotations

Table 6.3: GO terms in molecular function that are unique in annotated or predicted RBPs and/or shared between them.

		# of Proteins[b]					#of GO IDs[c]			
			Root		Leaf		Root		Leaf	
Type[a]	Total	None	Unique	Shared	Unique	Shared	Unique	Shared	Unique	Shared
A	1217	118	92	477	47	483	95	189	192	96
A-A⋂P	698	102	56	221	29	290	84	178	143	83
A⋂P	519	16	36	256	18	193	11	11	39	13
P-A⋂P	2418	907	21	226	26	1238	148	189	250	96

[a] AA⋃P (annotated but not predicted RBPs), A⋃P (annotated and predicted RBPs), and PA⋂P (predicted but not annotated as RBPs). [b] The total number of proteins, the number of proteins without GO IDs, with unique GO IDs, and shared GO IDs between predicted and annotated proteins at root and leaf levels. [c] The number of GO IDs that are unique or shared between predicted and annotated proteins at root and leaf levels.

only or 98%, 1,238/(1,238+26) for proteins with leaf annotations] shared GO IDs with annotated RBPs. In other words, almost all functions of these predicted moonlighting RBPs are associated with known RBPs. We note that the entire human genomes have 1,411 leaf GO IDs and annotated RBPs have 288 leaf GO IDs. That is, 20% of all leaf GO IDs associated with RBPs indicate the extensive association of RBPs with other biological functions.

To illustrate shared functions between predicted and annotated RBPs, we showed four clusters of predicted and annotated RBPs with four GO IDs in Fig. 6.2. Each GO ID not only contains many predicted and annotated RBPs at the same time but also connects with each other through proteins having multiple GO IDs. Top 10 GO IDs (excluding RNA-binding functions) enriched with moonlighting RBPs are listed in Table 6.4. Many of these 10 GO IDs are associated with transcription regulatory activity, suggesting DNA-binding activity. Indeed, we found that 350 out of 1,217 annotated RBPs (29%) are also annotated as DNA binding proteins according to GO annotations. Similarly, 22% (114/519) of predicted and annotated RBPs and 39% (728/1848) of predicted novel moonlighting RBPs are DNA binding proteins. Thus, a significant fraction of proteins can interact with DNA and RNA at the same time. The full list of predicted RBPs with annotated DNA binding is available on http://sparks-lab.org

Table 6.4: Top 10 GO IDs enriched with annotated and predicted RBPs, ranked according to the number of annotated RBPs

GO-Id	Function	Proteins	A	A⋃P	P-A⋃P	(A/All)	(A+P-A⋃P/All)
GO:0008270	zinc ion binding	2307	148	84	604	6%	28%
GO:0030528	transcription regulator activity	1508	138	98	434	24%	35%
GO:0001883	purine nucleoside binding	1599	132	66	136	8%	13%
GO:0005524	ATP binding	1475	129	65	133	8%	13%
GO:0016563	Transcription activator activity	146	44	35	105	30%	79%
GO:0003702	RNA polymerase II transcription factor activity	245	37	28	67	15%	31%
GO:0000287	magnesium ion binding	454	34	24	32	7%	9%
GO:0003743	translation initiation factor activity	58	29	16	5	50%	31%
GO:0016564	Transcription repressor activity	317	27	19	81	9%	28%
GO:0005525	GTP binding	372	19	14	7	5%	3%

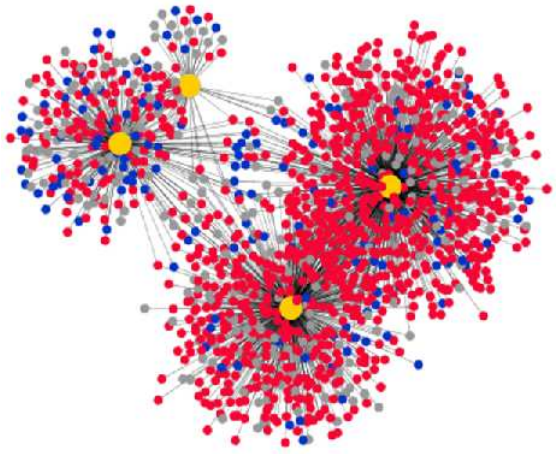

Fig. 6.2: The connection between proteins with four GO terms (GO:0030528, GO:0008270, GO:0001883 and GO:0000287) that are shared by annotated, not predicted (Grey); predicted and annotated (Blue), and predicted, novel (Red) RBPs. Each node represents a protein. One protein can connect to one or more GO terms in yellow

6.3.3 Validation of predicted novel RBPs by proteomic studies of human HeLa cells.

Sharing GO IDs between annotated and predicted RBPs support but do not validate predicted novel RBPs. Direct validation of our predicted RBPs is made possible by an recent proteomic experiment that obtained all mRNA-binding proteins of HeLa cells [17] . In this study, mRNA-binding proteins (mRBPs) in living HeLa cells were frozen by covalent UV crosslinking, captured by oligo(dT) magnetic beads after cell lysis, and identified by high resolution nano-LC-MS/MS. They found 860 mRBPs in which 375 are predicted RBPs. That is, the sensitivity for this dataset is 43.6% close to 42.6% sensitivity for all GO annotated RBPs. Similar sensitivity despite significantly different datasets confirms the overall accuracy of SPOT-Seq.

860 mRBPs discovered experimentally contain many novel RBPs. Using the same definition for RBPs as above, we obtained 746 proteins as novel RBPs in which 291 are predicted as RBPs. Thus, SPOT-Seq can detect novel RBPs in 39% sensitivity, close to the sensitivity for all RBPs (42.6%). In these 291 predicted and validated mRNA-binding proteins, the most frequently used templates belong to chains in PDB ID 3o58 (87 times). This validates the use of 3o58 as a template for predicting RBPs. Moreover, the majority of 291 predicted novel proteins (70%, 203/291) employed a template protein with mRNA binding function, indicating high accuracy in predicted binding RNA-type based on template RNA.

Castello et al. also defined a more stringent subset of previously unknown RBPs by excluding proteins that are previously experimentally validated, inferable by homology, and/or with a GO annotation containing RNA (not just RNA binding). This stringent set of previously unknown RBPs contains 315 proteins, 61 of which (19%) are predicted novel RBPs by SPOT-Seq. This large overlap demonstrates the ability of SPOT-Seq to go beyond homology-based inference of RNA-binding proteins and uncover truly novel RBPs.

Table 6.5: Number of proteins and RBPs involved in 11 different phenotypes

Disease	All	Annotated	A∪P	P-A∪P	Pathways
Cancer	372	10	0	41	14
Immune System	1579	53	8	115	30
Nervous System	3740	233	75	253	30
Cardiovascular	2668	157	71	166	44
Endocrine/M etabolic	1603	19	2	106	24
Digestive	2128	41	5	154	27
Urinary/reproductive	1497	14	5	109	20
Musculoskeletal/skin	3152	88	13	225	61
Respiratory	428	0	0	17	4
Congenital/metabolism	3299	103	17	192	101
Congenital/other	3543	198	86	245	83
Total	4602	337	151	284	176

6.3.4 Disease pathways associated with predicted RBPs

Validation of predicted novel RBPs provides incentive for analyzing their relevance to disease using known disease pathways of Kyoto Encyclopedia of Genes and Genomes (KEGG) database [162]. The KEGG database classified diseases into 11 types (Cancer, immune system diseases, nervous system diseases, cardiovascular diseases, digestive diseases, urinary and reproductive diseases, musculoskeletal and skin diseases, respiratory diseases, congenital disorder of metabolism, and other congenital disorders). These diseases correspond to 176 pathways and 4602 proteins. Among these proteins, 337 are annotated RBPs. 151 (44.8%) annotated RBPs are predicted by SPOT-seq. This is consistent with the overall sensitivity of 42.6%. In addition to recover known RBPs, SPOT-Seq also predicted 284 novel RBPs. The overall fraction of RBPs (both predicted and annotated) in all proteins involved in disease pathways is about 13%, slightly lower than 18% for all proteins in the human genome. Table 6.5 lists 11 diseases and the number of their related annotated RBPs and predicted RBPs. These newly predicted RBPs in disease pathways are expected to be useful for understanding disease mechanisms and generating new hypotheses for experimental testing. As an example, the Aminoacyl-tRNA biosynthesis pathway is shown in Fig. 6.3 to illustrate the extent of predicted and annotated RBPs involved. In this pathway, one node may contain more than one protein, and the number of

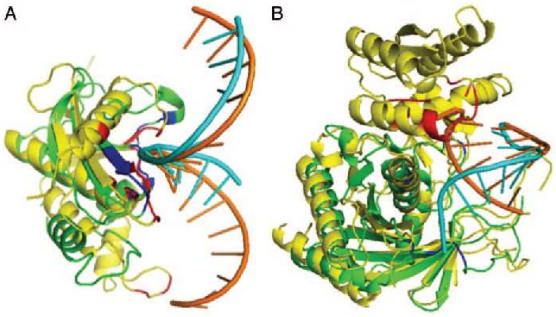

Fig. 6.3: Aminoacyl-tRNA biosynthesis pathway. Red, black and blue colors label nodes containing predicted novel RBPs, predicted and annotated RBPs and annotated RBPs, respectively. Each node contains more than one protein.

Table 6.6: Predicted novel RBPs in MutDB and their interactions with annotated RBPs

Predicted RBPs	Annotated RBPs	Refs.
FANCA (O15360)	BRCA1(P38398)	[164](Ganesan et al. 2002)
COL7A1(Q02388)	HSPA8(P11142)	[165–167]
KLF11(O14901)	ATXN1(P54253)	[168]
NKX2-1(P43699)	CALR(P27797)	[169]
COL17A1(Q9UMD9)	ACTN4(O43707)	[170](Jonson et al. 2007)
MSX1(P28360)	TBP(P20226), TAF1(P21675)	[171](Mittal and Hernandez 1997)
VCL(P18206)	RAVER1(Q8IY67)	[172]
GATA1(P15976)	SPI1(P17947)	[173]
MEN1(O00255)	POLR2B(P30876)	[174]

nodes is greater than number of proteins because each node can represent more than one gene product (proteins). For example, the node labeled as 6.1.1.17 is related with two proteins, EARS and EPRS. There are 11 annotated RBPs involved in this pathway, and 7 of them were predicted as RBPs by SPOT-seq. In addition, SPOT-Seq discovered 18 novel RBPs. One protein is BLM (P54132) that is known to interact with a RNA-binding protein FEN1(P39748) [163]. Moreover, most of the predicted novel RBPs (13/18=72%) employed templates that bind with tRNA. Interacting with known RBPs and predicted binding with tRNA provide additional supports for our predicted novel RBPs.

6.3.5 Disease-causing SNPs associated with predicted RBPs.

We searched the annotated and predicted RBPs for single-nucleotide polymorphism (SNP) and their associated phenotypes in the MutDB [175]. We found that 27 annotated/predicted RBPs and 135 predicted, novel RBPs are in the database. Among them, 6 annotated/predicted and 42 predicted, novel RBPs have SNPs in predicted RNA

Table 6.7: Predicted and annotated SNPs in RNA-binding region

Genename (Uniprot)[a]	Protein name	TPL[b]	Zscore	Energy	SNP region[c]	Phenotype
COL17A1 (Q9UMD9)**	Collagen- alpha-1(XVII)- chain	3o58C	19.91	-4.40	265-265	Epidermolysis- bullosa, junctional, non- Herlitz- type
COL3A1(P02461)*	Collagen- alpha-1(III)- chain	3o58C	18.64	-8.54	924-1188	Ehlers- Danlos- syndrome, type- III
COL9A2(Q14055)*	Collagen- alpha-2(IX)- chain	3o58C	18.56	-5.35	326-326	Epiphyseal- dysplasia, multiple, 2
COL1A2(P08123)*	Collagen- alpha-2(I)- chain	3o58C	18.17	-9.19	877-1148	Ehlers- Danlos- syndrome
COL10A1(Q03692)*	Collagen- alpha-1(X)- chain	3o58C	17.95	-3.65	617-618	Metaphyseal- chondrodysplasia, Schmid- type
RPL5(P46777)	60S- ribosomal- protein- L5	3o58E	17.95	-2.78	140-140	diamond- Blackfan- anemia- 6
COL2A1(P02458)*	Collagen- alpha-1(II)- chain	3o58C	17.94	-8.19	992-1197	Achondrogenesis, type- II- or- hypochondrogenesis
COL4A5(P29400)*	Collagen- alpha- 5(IV)- chain	3o58C	17.06	-6.35	289-609	Alport- syndrome
COL1A1(P02452)*	Collagen- alpha-1(I)- chain	3o58C	17.03	-6.16	947-195	Caffey- disease, Ehlers- Danlos- syndrome,
MAPT(P10636)*	Microtubule- associated- protein- tau	3o58C	15.99	-1.96	620-654	Dementia, frontotemporal, with- or- without- parkinsonism
EDA(Q92838)*	Ectodysplasin- A	3o58C	14.41	-3.06	61-302	Charcot- Marie- Tooth- disease, type-1D
COL6A2(P12110)*	Collagen- alpha-2(VI)- chain	3o58C	14.08	-5.54	498-498	Bethlem- myopathy
MECP2(P51608)	Methyl- CpG- binding- protein-2	3o58C	14.05	-3.51	167-305	Angelman- syndrome
GATA1(P15976)**	Erythroid- transcription - factor	3o58C	13.23	-1.33	216-218	X- linked, without- thrombocytopenia
EGR2(P11161)*	Early- growth- response- protein-2	3o58C	12.76	-2.01	355-383	Charcot- Marie- Tooth- disease, type-1D
KLF11(O14901)**	Krueppel- like- factor-11	3o58C	12.04	-1.73	347-347	Maturity- onset- diabetes- of- the- young, type- VII
COL11A2(P13942)*	Collagen- alpha-2(XI)- chain	3o58C	11.80	-8.25	808-808	Deafness, autosomal- dominant-13
COLQ(Q9Y215)*	Acetylcholinesterase- collagenic- tail-peptide	3o58C	11.66	-3.26	342-342	Endplate- acetylcholinesterase- deficiency
WAS(P42768)*	Wiskott- Aldrich- syndrome- protein	3o58C	11.60	-3.33	131-134	Neutropenia, severe- congenital, X- linked, Thrombocytopenia
COL7A1(Q02388)**	Collagen- alpha-1(VII)- chain	3o58C	11.43	-7.77	2348-2713	EBDr- inversa
FUS(P35637)	RNA- binding- protein- FUS	3o58C	11.30	-9.59	244-525	Amyotrophic- lateral- sclerosis- 6, autosomal- recessive, dementia
FOXL2(P58012)*	Forkhead- box- protein- L2	3o58C	11.22	-4.89	105-258	Blepharophimosis, epicanthus- inversus, and- ptosis, type-1
GLI2(P10070)*	Zinc- finger- protein- GLI2	3o58C	10.87	-3.65	932-932	Holoprosencephaly- 9
NKX2-1(P43699)**	Homeobox- protein- Nkx-2.1	3o58C	10.68	-4.61	213-213	Chorea, hypothyroidism
ALX3(O95076)*	Homeobox- protein- aristaless- like-3	3o58C	10.45	-3.44	203-203	Frontonasal- dysplasia-1
COL4A3(Q01955)*	Collagen- alpha-3(IV)- chain	3o58C	10.13	-6.64	1015-1015	Alport- syndrome, autosomal- recessive
CFP(P27918)*	Properdin	3o58C	10.00	-1.84	343-343	Cystic- fibrosis
VSX1(Q9NZR4)*	Visual- system- homeobox-1	3o58C	9.83	-1.77	159-244	Corneal- dystrophy, hereditary- polymorphous- posterior
TGIF1(Q15583)*	Homeobox- protein- TGIF1	3o58C	9.72	-2.54	280-280	Holoprosencephaly-4
NKX2-5(P52952)*	Homeobox- protein- Nkx-2.5	3o58C	9.66	-3.13	7-323	Atrial- septal- defect- with- atrioventricular- conduction- defects
ZFP57(Q9NU63)*	Zinc- finger- protein- 57- homolog	3o58C	9.62	-1.56	166-166	Diabetes- mellitus, transient- neonatal, 1
COL4A4(P53420)*	Collagen- alpha-4(IV)- chain	3o58C	9.46	-9.29	1201-1201	Alport- syndrome, autosomal- recessive
MED25(Q71SY5)*	Mediator- of- RNA- polymerase- II- transcription- subunit-25	3o58C	9.41	-5.56	335-335	Charcot- Marie- Tooth- disease, type-2B2
MSX1(P28360)**	Homeobox- protein- MSX-1	3o58C	8.82	-3.26	91-116	Orofacial- cleft- 5
WT1(P19544)	Wilms- tumor- protein	3o58C	8.79	-4.74	181-394	Denys- Drash- syndrome
VSX2(P58304)*	Visual- system- homeobox-2	3o58C	8.65	-2.60	200-227	Microphthalmia- with- coloboma-3
ZIC3(O60481)*	Zinc- finger- protein- ZIC-3	3o58C	8.65	-2.88	323-405	Heterotaxy, X- linked- visceral
TBX19(O60806)*	T- box- transcription- factor- TBX19	3o58C	8.61	-4.00	128-128	Adrenocorticotropic- hormone- deficiency
LAMB3(Q13751)*	Laminin- subunit- beta-3	3o58C	8.41	-2.97	199-199	Epidermolysis- bullosa, junctional, Herlitz- type
HOXD10(P28358)*	Homeobox- protein- Hox- D10	3o58E	8.13	-1.45	319-319	Charcot- Marie- Tooth- disease, foot- deformity- of
VCL(P18206)**	Vinculin	3a6pA	10.11	-0.95	975-975	Cardiomyopathy, dilated, 1W
FANCA(O15360)**	Fanconi- anemia- group- A- protein	3a6pA	9.24	-0.99	858-858	Fanconi- anemia, complementation- group- A
NIPBL(Q6KC79)*	Nipped- B- like- protein	3a6pA	8.29	-0.72	2430-2430	
RPS19(P39019)	40S- ribosomal- protein- S19	2xzmT	21.74	-2.52	15-120	Diamond- Blackfan- anemia-1
IGHMBP2(P38935)	DNA- binding- protein- SMUBP-2	2xzoA	19.44	-1.76	565-581	Neuronopathy, distal- hereditary- motor, type- VI
TRMU(O75648)	Mitochondrial- tRNA- specific-2- thiouridylase-1	2detA	27.11	-1.18	272-272	Liver- failure, acute- infantile
TUFM(P49411)*	Elongation- factor- Tu, - mitochondrial	1ob2A	25.27	-1.94	336-336	
MEN1(O00255)**	Menin	1i94L	8.47	-2.18	545-560	Multiple- endocrine- neoplasia-1

[a] Template PDB ID [b] Predicted SNP region. [c] P Predicted RBPs. T annotated RBPs

binding regions (Table 6.7). In 6 annotated/predicted RBPs, 80 in 170 SNPs (47%) are in the predicted RNA-binding regions. In 42 predicted novel RBPs, 844 in 1608 SNPs (52%) are in the predicted RNA-binding regions. Among 42 predicted novel RBPs, nine proteins interact with 10 annotated RBPs according to human protein reference database (HPRD) [176]. These 9 proteins and their interacting partners along with original citations are listed in Table 6.6.

Table 6.7 also lists the overlap between predicted RNA-binding residues with SNPs. For example, 40S ribosomal protein S19 is implicated in DiamondBlackfan anemia (DBA). Its known RNA-binding region [177, 178] agree with predicted RNA-binding amino-acid residues with a sensitivity of 42.1% (8/19). The predicted

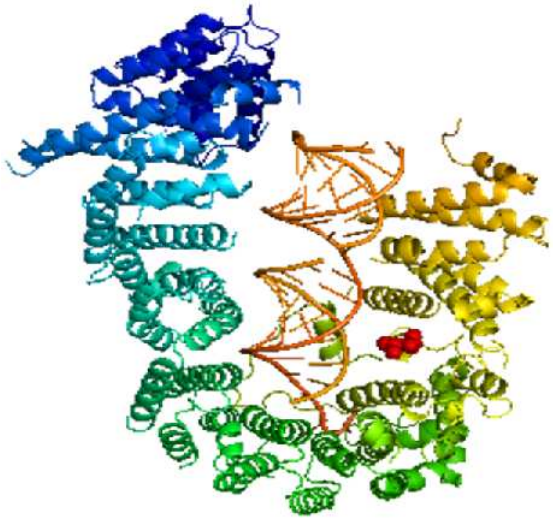

Fig. 6.4: Predicted complex structure for novel RBP: vinculin, related to cardiomyopathy dilated 1W. Locations of known SNP are shown as spheres.

RNA-binding residues in positions 15 (V→F), 47(P→L), 56 (R→Q), 55(T→M), 59(S→F), 62(R→Q, R→W), 101(R→H), and 120(G→R) are associated with known SNPs in the MutDB database [175]. As a second example of a known RBP, Wilms tumor protein (P19544) contains 3 disease-causing mutations (C330Y, R394P, R394W) in zinc finger domain [179–181]. These two mutated residues are predicted as RNA-binding residues by our method. This protein has three DNA-binding complex structures available within residue ranges of 318-428 (PDB ID#2PRT, #2JP9 and # 2JPA). Fig. 8.3 shows another example where the SNP is localized in the RNA-binding region in the predicted complex structure between tRNA and vinculin.

6.4 Discussions

In this study, a new method for RBP prediction based on known RBP complex structures was applied to human genome. The method uncovered 2,418 proteins that were not previously annotated as RBPs in the GO database. About half of these predicted novel RBPs were annotated as ORFs that lack GO annotations of molecular functions (908), or have only GO root ID (247), suggesting that they were poorly studied proteins. Some of these predicted novel RBPs (284) directly involve in disease pathways (Table 6.5), indicating their potential phenotypic roles. More importantly, 12% of these predicted novel proteins (291) are validated by a recent proteomic experiment that mapped all mRNA-binding proteins in living HeLa cells [17]. The consistent sensitivity (42.6% for

annotated RBPs in human genome and 43.6% for mRBPs in HeLa Cells) demonstrates the robustness of SPOTseq in making highly accurate prediction of RBPs.

Among all RBPs predicted, 80.5% are proteins with unknown functions or annotated with functions other than RNA-binding. This suggests that many more RBPs exist than those that are currently annotated. If we combine predicted RBPs with annotated RBPs and assume that majority of predicted and annotated RBPs are true, these RBPs would consist of 18% [=(1,848+570+1,217)/20,270] of all genes. Because the sensitivity of our technique is at about 43%, the actual number of RBPs is likely greater than 18% even if we take into account of errors in our prediction. The huge number of RBPs highlights the scope and significance of the protein-RNA interaction network.

Most of the RBPs predicted here have functions other than RNA-binding. This so-call moonlighting capability of RBPs is consistent with experimental screens of yeast and human proteins. It was found that novel RBPs uncovered in screens often have enzymatic activities [14, 15] as well as RNA-binding kinases and RNA-binding architectures [17] . Thus, moonlighting aspect of RBPs is likely more common than previously appreciated. In particular, 39% of predicted moonlighting proteins are related to DNA-binding. This is not caused by inability of SPOT-seq to distinguish RNA- from DNA-binding. In fact, the application of SPOT-seq to 250 DNA-binding proteins did not yield any false positive prediction of RBPs [36] . Thus, many proteins can interact with RNA and DNA at the same time.

A surprising result from our template-based technique is that many predicted RBPs employed the templates from 60 S ribosomal proteins (PDB ID 3o58). This is true for both predicted novel and annotated RBPs. We are confident about these predictions because our benchmark test indicates the accuracy of prediction based on 3o58 is the same as that based on other templates. Moreover, known and novel RBPs predicted from 3o58 have interspersed confidence levels as shown in Table 6.2. More importantly, 87 novel RBPs based on 3o58 templates are validated as mRNA-binding proteins [17] .

One caveat of the SPOT-seq method is its reliance on known protein-RNA complex structures as templates for predicting complex structures. This limitation contributed to the respectable but low sensitivity (43%) of the prediction. This sensitivity was also resulted from our emphasis on high precision (fraction of correct predictions in all predictions). As more protein-RNA complex structures are solved, our method will improve in recovering known RBPs and uncovering novel ones. Increasing the sensitivity of SPOT-seq by combing it with other sequence- and structure-based approaches [27–30, 34, 36, 116, 148, 149] is working in progress. Nevertheless, the ability to double the number of annotated RBPs with such sensitivity suggests that many more interesting novel RBPs remain to be uncovered.

Chapter 7 Prediction of RNA binding proteins comes of age from low resolution to high resolution

Abstract

Evidence is accumulating that the protein-RNA interaction network is substantially larger than protein-protein and protein-DNA interaction networks combined. Recent experimental studies begin to uncover more and more unconventional or moonlighting RNA-binding proteins (RBPs). At the same time, more and more protein-RNA complex structures are deposited into protein databank. These resources provide ample statistics for developing computational techniques dedicated to RBP prediction. This review compares traditional machine-learning based approaches with emerging template-based methods at several levels of resolution of prediction ranging from two-state binding/non-binding prediction, binding residue prediction, to protein-RNA complex structure prediction. The analysis indicates a promising future for highly accurate RBP prediction with a reasonable sensitivity using a template-based approach.

7.1 Introduction

RNA directly involves a wide variety of functions ranging from protein synthesis, post-transcriptional modification, to post-transcriptional regulation. Unlike DNA, located mostly in the cell nucleus, RNA is transcribed in nucleus and transported to cytoplasm as non-coding RNA or for translation. Diverse localizations and different functionality of RNA transcripts [154–156] along with only 3% human genome coded for proteins [182] suggest that the network of protein-RNA interactions is likely much larger and more complex than those of protein-DNA and protein-protein interactions combined [157] . These RNA-binding proteins (RBPs) are challenging to locate experimentally although some progress in high-throughput biochemical approaches are made [157–159,183] and hundreds of novel unconventional or moonlighting RBPs have

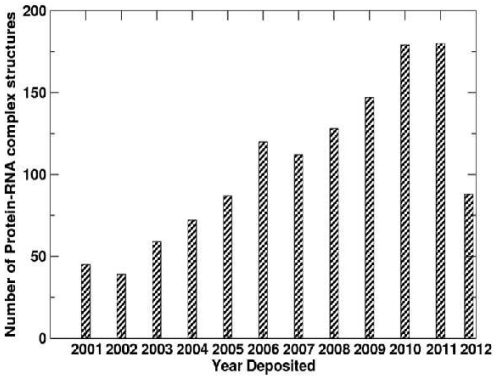

Fig. 7.1: Number of Protein-RNA complex structures deposited in protein data banks since 2001.

been discovered [14, 15, 17] . This, however, scratched only the surface of RBPs and their associated post-transcriptional network.

A complete understanding of the protein-RNA interaction between a specific protein and a RNA requires to determine their complex structure. Despite of difficulty in solving protein-RNA complex structures [184–186], the number of non-redundant complex structures deposited in protein databank has been quadrupled from 45 per year at 2001 to 180 at 2011 (at 90% sequence identity cutoff), as shown in Fig. 7.1. By comparison, the number of deposited structures is less than tripled from 2831 at 2001 and 8091 at 2011 (http://www.rcsb.org/pdb/statistics). The growing number of protein-RNA complex structures provides an increasingly larger dataset for analyzing the principles of protein-RNA recognition [137, 187–191]. However, not all members in the same structural folds have RNA-binding activities. For example, the Structural Classification Of Proteins (SCOP) [133] has 44 folds shared by both RNA and non-RNA binding proteins [36].

The challenge and expense of experimental determination of RBPs necessitates the development of accurate and efficient computational techniques. In this review article, we will classify different computational methods according to the resolution of prediction from low, medium, high to the highest. A low-resolution prediction is a simple two-state prediction of whether a protein is RNA binding or non-RNA binding. A medium-resolution prediction locates the amino-acid region of a RBP that binds to RNA (RNA binding site/motif prediction). A high-resolution prediction indicates the types of RNA binding to a RBP. The highest resolution prediction

will predict the three-dimensional structure of protein-RNA complexes with predicted RNA binding sequence. The highest resolution prediction can simultaneously make all lower resolution predictions including the RNA type, RNA-binding site, and the two-state RBP/non-RBP classification, but not vice versa. Most computational methods developed so far focused on low to medium resolution prediction [192, 193].

Here, we will provide a brief review based on the resolution as well as the information (i.e. sequence versus structure-based) employed in prediction.

7.2 Function Prediction in different resolutions

7.2.1 Low Resolution Function Prediction: Two-State RBP Prediction.

Structure-based Inference of RBPs. Negatively charged RNA preferentially binds to positively charged proteins. Electrostatic interactions are obviously an important feature for detecting RBPs. Shazman and Mandel-Gutfreund [117] employed Support Vector Machines (SVM) to combine electrostatic patches, solvent accessibility, cleft sizes and other global protein features for RBP prediction. This method trained on 76 RNA binding proteins and 246 non-nucleic acid binding proteins and achieved a Matthews correlation coefficient (MCC) of 0.72 based on the leave-one-out test. However, it is unable to distinguish RBPs from DNA-binding proteins. Ahmad and Sarai [194]employed neural networks that are based on charge, dipole moment, three eigenvalues of quadrupole moments generated from the structure. It was trained on 160 RBPs and 2441 non-RBPs and achieved 0.79 for an area under the ROC curve based on the leave-one-out test. Table 7.1 provides a list of features for the two methods described above. More recently, we have developed an alternative approach based on a template library of known protein-RNA complex structures [34, 42]. In this method, a target structure is aligned to the templates in the template library and a RBP is predicted if the structural similarity between the target and a template is higher than a certain threshold. Several structural alignment programs were tested. Among them, SPalign [42] was found to give the highest MCC value of 0.37 based on a dataset of 212 RNA binding domains and 6761 non-RNA binding domains with 250 RNA-binding domains

as templates. When the query structure is compared to template structures, the templates with sequence identity ¿30% to the query sequence are excluded in order to test the ability of the method to detect remote homologs. SPOT-struc (RNA) [34] improves over the method based on a structural similarity score only by using a relative structural similarity between RBPs and non-RBPs and by predicting the binding affinity between the query protein and the template RNA with a knowledge-based energy function based on distance-scaled finite ideal gas reference state (DFIRE) [33] It achieves a MCC value of 0.57 for the same dataset above. SPOT-struc (RNA) has the ability to separate RNA- from DNA-binding proteins because it yields zero false positives after excluding proteins known to bind both DNA and RNA when applied to a dataset of 331 DNA binding domains.

Sequence-based inference of RBPs. The main limitation of a structure-based technique is that the structures for most proteins are not yet known. One common technique is homology-based prediction assuming that proteins with similar sequences are likely to perform the same function. Enzymes [197, 198],for examples, tend to have a conserved function, if they share more than 40% to 50% sequence identity. However, such prediction will produce false negatives by failure to detect functionally identical remote homologs [199] and false positives by ignoring possible functional divergence for highly homologous sequences [197]. Thus, there is a need to go beyond simple homology-based search. Several SVM-based tools [27–30, 84, 116, 195]were developed. Different methods mainly differ in features employed. Commonly used features are the composition of amino-acid residues, hydrophobicity, amino acid composition, charge, hydrophobicity and accessible surface area. Early studies [27,116] did not remove homologous sequences in training and testing. Due to limitation of SVMs, most methods were trained with nearly equal number of RBPs and non-RBPs [28–30, 84, 195]. In a real-world situation, RBPs are only a fraction of all proteins. The reported MCC values are 0.53 for a dataset of 134 RBPs and 134 non-RBPs [30], 0.51 for a dataset of 69 RBPs and 100 non-RBPs by RNApred [84], 0.65 for a dataset of 687 RBPs and 687 non-RBPs [195]. Recently, we developed a template-based

Table 7.1: Structure and sequence-based features for RBP prediction

Method[Ref.]	Technique	Features
Structure-based		
[117]	SVM	Electrostatic surface patches, molecular weight, solvent accessibility, dipole, quadrupole, patch size, size of the largest clefts, number of atoms in positive and negative patches, patch surface overlap
[194]	NN	Charge dipole moment, quadrupole moment and functional property of protein chain
SPalign [42]	Template-based	Structural alignment
SPOT-Struc [34]	Template-based	Structural alignment plus binding affinity estimation
Sequence-based		
[28]	SVM	Hydrophobicity, secondary structures, solvent accessibility, van der Waals volume, polarity, polarizability and amino acid composition
[195]	Voting	Hydrophobicity, predicted secondary structure, predicted solvent accessibility, normalized Van Der Waals volume, polarity, and polarizability
RNApred [84]	SVM	Residue composition, predicted RNA binding residues, PSSM
[29]	SVM	Clustered amino acids according to dipoles and volumes of side chains.
[27]	SVM	Pseudo-amino acid composition, charge, hydrophobicity, accessible surface area
[196]	SVM	Amino acid composition, periodicities
SVMProt [116]	SVM	Amino acid composition, charge, polarity, and hydrophobicity
SPOT-Seq [36]	Template-based	Sequence-to-structure match and binding affinity estimation

approach called SPOT-seq [36] which is similar to SPOT-struc [34]except that the query structure is predicted by a fold-recognition technique called SPARKS-X [49]. More specifically, SPARKS-X attempts to match the query sequence to the templates of known protein-RNA complex structures. If a match is found (based on a Z-score), a binding affinity is predicted based on a knowledge-based energy function. The query sequence is an RBP if the binding affinity is higher than an optimized threshold. This coupled structure and binding prediction leads to a MCC value of 0.62 for independent test on 215 RBPs and 5765 non-RBPs with a template library of 1164 RNA-binding domains and RNA-binding chains. This MCC value is even higher than 0.56 given by SPOT-struc for the same dataset despite of using predicted structures in SPOT-seq, rather than actual structures in SPOT-struc, suggesting possible cancellation of errors of structure and binding prediction.

Method comparison. There is a lack of comparison between different methods for RBP prediction. Most methods described above do not have web-servers or their web-servers are no longer functional. We only found two available servers (RNApred [84], http://www.imtech.res.in/raghava/rnapred/ and SVMprot [116]). Both of them are sequence-based methods. They are compared to SPOT-seq along with our structure-based techniques SPOT-struc and SPalign with a dataset of 257 RBPs and 5765 non-RBPs in Table 7.2. This dataset is an independent test set for SPOT-seq at 25% sequence identity. RNApred predicted 203 out of 257 RBPs and 2415 out of 5765 non-RBPs as RBPs. RNApred achieved a MCC value of 0.15, sensitivity of 79%, and precision of 8%. SVMprot yields the MCC of 0.19, sensitivity of 50% and precision of 13%. By comparison, SPOT-seq has a MCC value of 0.60, sensitivity of 44%, precision of 84% for the same dataset. Thus SPOT-seq is significantly more powerful in separating RNA from non-RNA binding proteins. It is even more powerful than structure-based techniques that achieved MCC values of 0.46 (SPalign) and 0.50 (SPOT-struc). Fig. 7.2 displays the Receiver Operating Characteristic (ROC) curves for these sequence and structure-based methods. It is clear that SPOT-Seq, the template-based technique, is substantially more accurate than other sequence-based,

Table 7.2: Comparison of methods for low-resolution, two-state RBP prediction

	257 RBPs + 5765 non-RBPs							245 DBPs
Method	MCC	Sen.	Pre.	TP	FN	TN	FP	FP
Structure-based								
SPalign* [42]	0.46	33%	67%	85	172	5723	42	6
SPOT-Str* [36]	0.50	32%	84%	83	174	5749	16	3
Sequence-based								
RNApred [84]	0.15	79%	8%	203	54	3350	2415	168
SVMprot [116]	0.19	50%	13%	129	128	4898	867	55
SPOT-seq [34]	0.60	44%	84%	114	143	5743	22	0

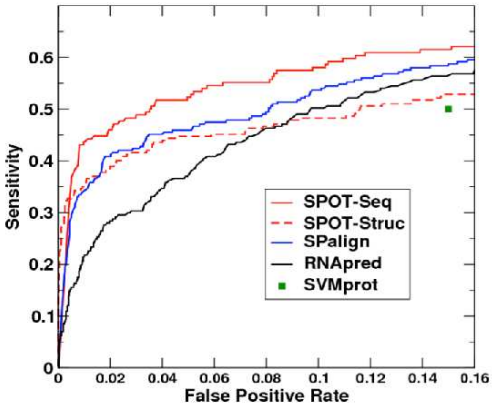

Fig. 7.2: The ROC curves for several RBP predictors. SPOT-seq, RNA-pred and SVMprot are sequence-based methods while SPalign and SPOT-struc are structure-based.

machine-learning techniques (RNApred and SVMprot) or structure-based techniques (SPalign and SPOT-struc). For structure-based technique, SPalign, although is less accurate than SPOT-Struc at low false positive rates, has higher sensitivity at high false positive rates. This suggests that replacing TMalign employed in SPOT-Struc by SPalign for pairwise structure alignment will likely further increase the power of SPOT-Struc.

Discriminating RBPs from DNA-binding proteins. DNA-binding proteins are important control for examining the accuracy of RBP prediction because DNA-binding interfaces are also positively charged as RNA-binding interfaces. Most methods are either unable to separate RNA from DNA binding proteins or not tested in this aspect. Table 7.2 confirms high false positives given by RNApred (69%) and SVM-prot (22%) when tested on 245 DNA-binding proteins, compared to zero-false positives given by SPOT-seq. These 245 DNA-binding proteins are a subset of DB250 which are modified by excluding 5 RNA-binding proteins [34].

7.2.2 Medium Resolution Function Prediction: Binding Residues Prediction

Locating functional residues is an important first step for understanding the mechanism of function. Thus, there are a significant number of studies in predicting RNA-binding residues. Most studies are machine-learning techniques trained from sequences or structures of known RBPs.

Structure-based prediction. How to capture key structural features is the challenging question for accurate prediction of RNA-binding residues from a given structure. Table 7.3lists structural features employed by several structure-based techniques [125, 126, 200–203].The methods range from docking, random forest classifier, neural network, SVM, nave Bayes classifier to linear regression. The notable features are sequence conservation, secondary structures, types of amino acid residues, solvent accessible surface area and interface propensity. There are some overlaps between the features employed for RBP prediction and binding residue prediction, except that one focuses on the whole protein level and the other is on the residue level. We developed a template-based approach called SPOT-struc (RNA) [34] that predicts binding sites based on structural alignment to known protein-RNA complex structures and prediction of protein-RNA binding affinity. SPOT-struc (RNA) is based on an alignment program called TM-align [139].Another method SPalign was developed to further improve the accuracy of alignment and identification of binding regions [42].

Sequence-based prediction. For sequence-based prediction, the prominent feature is sequence similarity and evolution information [30]. Additional features as shown in Table 7.3 include properties of amino acid residues, predicted secondary structures and solvent accessibility. Most methods are based on SVM. These features are typical features utilized in secondary structure prediction and ASA prediction as well (e.g. [50]). All above methods are machine-learning based tools. We developed a template-based technique called SPOT-seq [36] that infers RNA-binding residues according to predicted RNA-protein complexes between the model structure of the target protein and the structure of template RNA.

Table 7.3: Structure and sequence-based features for RNA-binding residue prediciton

Methods	Technique	Features
[200]	NN	Secondary structure, amino acid type
KYG [201]	Scoring	Residue doublet interface propensity, multiple sequence profiles
[125]	Scoring	Surface binding pocket, electropositive atoms, spatially evolution principle
[126]	SVM/Naive Bayes	residue contacts map, PSI-BLAST profile, Graph theory properties
Struct-NB [202]		surface roughness, interface residue propensity CX score
[127]	Linear Reg.	PSSM, secondary structure and solvent accessibility
OPRA [131]	Docking	pairwise residue-ribonucleotide interface propensities
[128]	Random forest classifier	interaction propensities, physicochemical characteristics, hydrophobicity, rASA, secondary structure, conservation score side-chain environment
SPalign [49]	Template-based	Structural alignment
SPOT-Struc [34]	Template-based	Structural alignment plus binding affinity estimation
BindN [94]	SVM	Side chain pKa value, hydrophobicity index, molecular mass
RNABindR [120]	SVM	smoothed PSSM
BindN+ [95]	SVM	side chain pKa value, hydrophobicity index, molecular mass, PSSM
NAPS [96]	bootstrap aggregation and cost sensitivity learning	PSSM
PBRpred [204]	SVM	PSSM, predicted secondary structure and solvent accessibility
PiRaNhA [205]	SVM	PSSM, residue interface propensity, predicted residue accessibility value
PRBR [206]	Random forest	secondary structure, evolution information, conservation information of physicochemical properties of amino acids, polarity-charge, hydrophobicity
SPOT-Seq [36]	Template-based	Sequence-to-structure match and binding affinity estimation

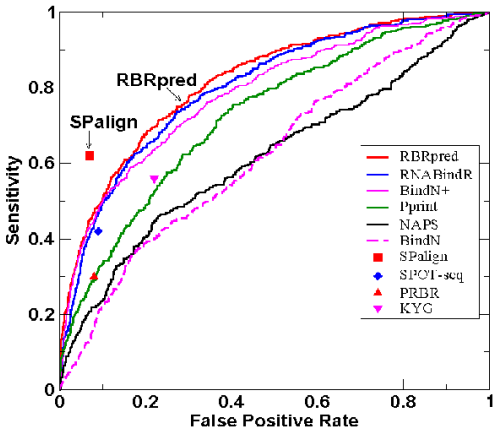

Fig. 7.3: Performance of RNA-binding prediciton by several sequence and structure-based techniques as labeled.

Method Comparison. One conclusion is that structure-based techniques do not have any advantage over sequence-based techniques. The second conclusion is that all methods have MCC below 0.6. However, different datasets make comparisons between different methods impossible. To compare different methods, we built a dataset of 106 RNA-binding domains (RB106) that were released in 2011 and 2012. RB106 is a non-redundant dataset with pairwise sequence identity lower than 35%. However, only 67 domains in 106 domains were predicted as RBPs by SPOT-seq because of lack of templates or low binding affinity. Thus, we also showed results for the RB67 set. In addition, we further remove the domains that have more than 45% sequence identity with RNA-binding domains released before 2011. This leads to a small dataset of 20 RBPs (RB20). We employed 45% sequence identity cutoff here because a lower cutoff will lead to fewer new RNA-binding complex structures.

Table 8.1 lists the performance of various structure and sequence-based techniques for the three datasets (RB106, RB67 and RB20). In structure-based techniques, SPalign has a consistent top performance among three structure-based techniques (SPalign, SPOT-struc and KYG). In both SPalign and SPOT-struc, all templates more than 35% sequence identity to the target are removed. In sequence-based methods, BINDN+ has the best performance in the MCC value for RB106 (MCC=0.59), followed by PBRpred (MCC=0.57). For RB20, PBRpred gives the highest MCC value (0.39), followed by BINDN+ (0.38) and RBABindR (0.37). SPOT-seq, on the other hand, yields the highest MCC value for those proteins predicted as RBPs (0.63 for RB67). SPOT-seq achieved an MCC value of 0.33 for RB20 by

Table 7.4: The performances of structure and sequence-based methods for predicting RNA-binding residues for three domain datsets(RB106, TP67, RB20)

Method	Sensitivity(%) RB106(TP67)	Precision(%) RB106(TP67)	MCC PB106(TP67)	MCC RB20
KYG [201]	62(61)	61(65)	0.43(0.44)	0.26
SPalign [49]	57(64)	61(67)	0.39(0.50)	0.43
SPOT-Struc [34]	55(61)	60(69)	0.36(0.49)	0.36
BindN [94]	57(56)	59(64)	0.39(0.40)	0.16
RNABindR [120]	69(77)	67(65)	0.52(0.53)	0.37
BindN+ [95]	70	74(77)	0.59(0.62)	0.38
NAPS [96]	42(45)	55(56)	0.28(0.28)	0.18
PBRpred [204]	74(78)	69(70)	0.57(0.59)	0.39
PRBR [206]	55(56)	69(72)	0.46(0.47)	0.22
SPOT-Seq [36]	81(68)	50(82)	0.39(0.63)	0.33

using the templates that have no sequence identity higher than 45% to target (45% is employed here to be consistent with the cutoff for building this small novel RBP structure database). It is clear that sequence-based techniques are as accurate as or more accurate than template-based techniques in predicting RNA binding residues. All methods, however, have dramatic reduction of accuracy if sequence identities to known RBPs are lower than 45%. The performance of various methods is also compared by the ROC curves in Fig. 7.3 Regardless of datasets, two best performing methods are RBPpred and BindN+.

7.2.3 High-Resolution Function Prediction: Binding RNA Type Prediction

Predicting the type of RNA binding with a given RBP provides a more detailed information on the function of RNA-binding proteins. Yue et al. [28] developed a sequence-based predictor for separating rRNA-binding from RNA-binding proteins. They found that rRNA-binding proteins can be more accurately predicted than RNA-binding proteins. Shazman and Mandel-Gutfreund [117] employed a multi-class SVM to classify rRNA, tRNA, and mRNA-binding proteins based on electrostatic properties derived from protein structures. It has the highest success rate for tRNA-binding proteins (13/13) but a lower success rate for rRNA (32/46) and mRNA (17/23) binding proteins. This method, however, cannot separate RNA from DNA

binding proteins. We developed the sequence-based technique SPOT-seq that can predict the RNA types by assuming that the query protein and its matching template RBP bind to the same type of RNA [36]. SPOT-seq achieved success rate of 69% (33/48) for tRNA, 56% (15/27) for rRNA and 96% (54/56) for mRNA for an independent test set of 215 RNA-binding proteins, compared to 62%, 73% and 91% for the training set of 216 RBPs. It should be noted, however, that the RNA structural motif, rather than the RNA functional type, is the key for the RBP function as many proteins can bind with different types of RNAs.

7.2.4 Highest Resolution Function Prediction: Protein-RNA Complex Structure Prediction

To understand the mechanism of protein-RNA binding, atomic resolution of protein-RNA complex structures is required. One method to predict protein-RNA complex structures is protein-RNA docking that relies on known protein and RNA structures. Such docking techniques for protein-RNA interactions can be modified from many docking software tools for protein-protein and protein-ligand docking after equipping with a scoring/energy function for protein-RNA interaction. For example, Zheng et al utilized the RosettaDocking [207] program to generate protein-RNA complex decoys and evaluate the ability of a knowledge-based energy function based on a conditional-probability function to discriminate docking decoys [130].Perez-Cano et al. employed the FTDOCK [208] program plus propensity-based statistical potentials [131] . Tuszynska and Bujnicki employed the GRAMM [209]docking program and two separate statistical potentials (QUASI-RNP based a quasi-chemical reference state and DARS-RNA based on the reference state from decoys) for scoring [210].Setny and Zacharias employed the protein-docking program ATTRACT [211] and a knowledge-based energy function employing a quasi-chemical approximation [212]. These studies demonstrated the usefulness of knowledge-based energy functions for decoy discrimination and selection of near-native docking decoys. We also developed a DFIRE-based statistical potential that increases true positive rates and decreases false

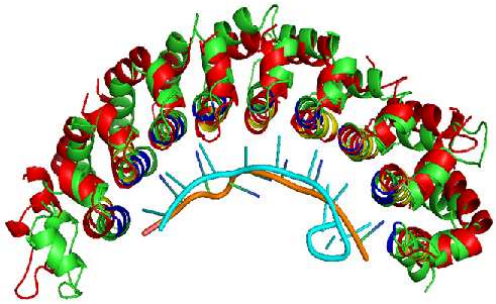

Fig. 7.4: Comparison between the predicted (red) and actual (green) structure and predicted (yellow) and actual (blue) binding residues. The RNA structure of actual is cyan and that of the predicted is orange. The target is 1m8yB and the template is 3k5qA.

positive rates in predicting RNA-binding proteins [34]. Protein-RNA docking, however, is more challenging than protein-protein docking because RNA structures are more flexible than protein structures. This is demonstrated by critical assessment of predicted interaction (CAPRI, 2009). CAPRI, which typically assessed protein-protein docking models, included a protein-RNA complex structure in a recent round [213].All docking predictions failed for this protein-RNA complex target because of inaccurate model RNA structure.

Another approach to predict protein-RNA complex structures is to use known protein-RNA complex structures as templates. SPOT-seq [36] and SPOT-struc [34] are sequence and structure-based techniques for predicting protein-RNA binding complex structures based on template-based structure prediction program SPARKS X and structural alignment program TM-align [139], respectively. Both methods can provide quite accurate prediction of binding residues and complex structures if a significantly matching template is found. For example, SPOT-seq can locate matching templates with strong predicted binding affinity for 114 out of 257 RBPs targets. One example is shown in Fig. 8.3. In this figure the target protein is 1m8yB (human Puf protein, Pumilio1), the SPOT-seq selected template is 3k5qA (Caenorhabditis elegans fem-3 binding factor 2). The sequence identity between these two proteins is 24.9%. The advantage of SPOT-seq or SPOT-struc is their computational efficiency that allows large genome-scale prediction.

7.3 Summary and Outlook

Constantly increasing number of protein-RNA complex structures makes it possible for the development of various techniques for predicting RNA-binding proteins at different levels of functional details. Sequence-based techniques using machine-learning methods are ineffective in separating RNA-binding from non-RNA binding proteins, DNA-binding proteins, in particular. Our result shows that a template-based technique is the only viable approach for RNA-binding discrimination. On the other hand, for a known RNA-binding protein, the best machine-learning techniques are often more accurate in locating RNA-binding residues than a template-based approach. This is true particularly for those proteins that are not predicted as RBPs by the template-based approach. Only a few techniques have been developed to predict the types of RNA interacting with a RBP. A template-based approach can make a reasonable prediction based on the type of RNA in the matching template-RNA complex structure. Similarly, a template-based approach is the only reliable tool available for predicting protein-RNA complex structure. As more and more protein-RNA complex structures deposited into protein databank, one can expect that a template-based approach will be increasingly useful. An application of such an approach to human genome has yielded more than 2000 novel RBPs and a recovery of 42.1% in known RBPs and a recovery of 41.5% newly discovered 860 mRNA-binding proteins [17] [Zhao et al. submitted]. The consistency of the recovery (or sensitivity) in two separate datasets highlights the robustness of template-based tools for predicting truly novel RNA-binding proteins. Further, the machine-learning based and template-based approaches are likely complementary each other. Combining these two approaches will likely further improve the accuracy of RNA-binding function prediction.

Chapter 8 Structure-based prediction of carbohydrate-binding proteins, binding residues and complex structures by a template-based approach

8.1 Introduction

Carbohydrates perform essential roles in cell processes in living organisms by interacting with proteins through both non-covalent (carbohydrate-protein binding) and covalent (glycosylation) interactions. Glycosylation of proteins and lipids coats the surfaces of all living cells and tissues with carbohydrates. The spatial patterns of such carbohydrate coating change during cell development1 and tumor progression and metastasis [214, 215]. Thus, recognition of cell-surface carbohydrates, one of the key functions of carbohydrate-binding proteins (CBPs), is subject of intensive studies for biomarker discovery and inhibitor design [214,216]. Abundant carbohydrates in human cell surfaces are also exploited by carbohydrate-binding proteins in pathogens for cell invasion and detection avoidance. As a result, CBPs in pathogens have been employed as potential drug targets [217]. Thus, it is critically important to locate all CBPs and elucidate their binding mechanisms.

Experimentally, glycan arrays have been developed for high-throughput searching of novel CBPs and investigation of their binding specificity [218–220]. However, it is challenging to construct a sizeable, diverse glycan array because of difficulty in synthesis and isolation of carbohydrates. Here, we focus on an alternative approach: prediction of CBPs and their binding residues by computational techniques.

Currently, predicting CBPs and their binding residues are treated as two separate problems [221–225]. Someya et al [221] predicted carbohydrate-binding proteins by combining protein sequences information with support vector machines (SVM). This approach employed triple sequence patterns and frequencies of grouped amino acids as features and has achieved 0.67 for Mathews correlation coefficient (leave-one-out cross validation) based on a dataset of 345 CBPs and non-CBPs. This method is limited to

CBP prediction. Most of the methods developed for predicting carbohydrate-binding residues, on the other hand, assume that their structures are known. For example, Shionyu-Mitsuyama et al. predicted binding residues by building empirical interactions rules [222]. Tsai et al. utilized 3D probability density maps [224]. Others employed machine-learning techniques based on binding propensity and solvent accessibility [226] or selected geometric and chemical features [227]. These methods, however, cannot distinguish CBPs from non-CBPs.

Here, we will introduce a single template-based method for prediction of CBPs and carbohydrate-binding residues. This work is inspired by our highly effective template-based technique named SPOT-Struc for structure-based prediction of DNA-/RNA- binding proteins and their binding sites [32, 34]. In this approach, the target structure is first structurally aligned to the proteins with known protein-RNA/DNA complex structures. Significantly aligned structures are then employed for building model complex structures between target structure and template RNA/DNA and for predicting binding affinities.

In this work, we will extend SPOT-Struc to CBPs. Such an extension is possible because of the existence of a reasonable size of complex structures of protein and carbohydrates in protein databank18 despite their low binding affinity and highly flexible structures of carbohydrates. This complex structure dataset allows us to develop the first distance-dependent knowledge-based energy function for protein-carbohydrate interaction that is essential for the accuracy of SPOT-Struc for CBPs. A distance-scaled, finite, ideal gas reference (DFIRE) state will be used as for proteins [33] and protein-DNA/RNA interactions [32, 34]. This knowledge-based energy function is then combined with a recently developed structure alignment method SPalign [42] for predicting CBPs and binding residues. This method is tested on 122 non-redundant RBPs and 2880 non-RBPs and achieved the Mathews correlation coefficients of 0.61 and 0.58 for prediction of CBPs and carbohydrate-binding residues, respectively. The sensitivity and precision of CBP prediction are 45% and 85% respectively. A

similar-level sensitivity is achieved for APO and HOLO structures. Application of this method to structural genomics targets revealed several novel CBPs.

8.2 Methods and Materials

8.2.1 Datasets

Template library of carbohydrate-binding proteins (T562). A template library was built based on the PROCARB database that contains 604 protein-carbohydrate complex structures [228]. We then selected only those proteins with more than 5 residues binding with carbohydrates. Here, a residue is defined as a carbohydrate-binding residue if it has one or more heavy atoms that are within 6.5 distance from any heavy atoms of carbohydrates. We further divided selected proteins into domains according to DDomain classifications. Both domains and their corresponding chains are included in the final template library that has 562 CBPs. We have included both domains and chains in the template library so as to improve the possibility of locating a suitable template.

Positive Binding-domain Dataset (BD122). We built a positive database of carbohydrate-binding domains for training and cross validation by firstly excluding the chains in T562. We further remove the redundant proteins by using BLASTClust24 with a sequence identity cutoff of 30%. The final dataset contains 122 CBPs.

Negative (non-binding) dataset (NB3442). We built the negative dataset by querying the PDB database and removing all PDB files containing carbohydrates. The protein chains are splitted into domains by DDomain. All redundant domains are removed by BLASTClust [134] with a sequence identity cutoff of 30%. One representative protein was randomly selected from each cluster. The final dataset contains 3442 protein domains.

APO45/HOLO45 dataset. To examine the effect of binding-induced change of protein conformations on accuracy and sensitivity of CBP detection, we built a dataset with both bound (HOLO) and unbound (APO) structures of CBPs. We located the APO structures by selecting homologous sequences of proteins in BD122. All APO chains are

divided into domains or by DDomain. Only HOLO and APO domains with sequence identity¿50% were selected. Here, the pair-wised sequence identity was calculated by ALIGN0 program from FASTA2 package [136]. We found 45 APO-HOLO domain pairs. The majority of the pairs (31 out of 45) have sequence identity more than 80%.

Structural genomics targets (SG2076). Our method is applied to 2076 structural genomics targets that was obtained by us from previous study on structure-based prediction of DNA-binding proteins16. This dataset was obtained by querying structural genomics targets in the protein databank. All structures were divided into domains by the automatic domain parser DDOMAIN25. Redundancy was removed by using BLASTClust [134] with a sequence identity cutoff of 30%.

8.2.2 DFIRE-based energy function for protein-carbohydrate interactions

We employed the same equation as the DFIRE-based interaction for protein-RNA interactions [34] as below

$$\overline{u}_{i,j}^{\mathrm{DRNA}}(r) = \begin{cases} -\eta \ln \dfrac{N_{obs}(i,j,r)}{\left(\frac{f_i^v(r) f_j^v(r)}{f_i^v(r_{cut}) f_j^v(r_{cut})}\right)^{\beta} \frac{r^{\alpha}\Delta r}{r_{cut}^{\alpha}\Delta r_{cut}} N_{obs}^{lc}(i,j,r_{cut})}, & r < r_{cut}, \\ 0, & r \geq r_{cut}, \end{cases} \tag{8.1}$$

where the volume-fraction factor $f_i^v(r) = \frac{\sum_j N_{obs}^{\mathrm{Protein-RNA}}(i,j,r)}{\sum_j N_{obs}^{\mathrm{All}}(i,j,r)}$, $N_{obs}(i,j,r)$ is the number of pairs of atoms i and j within the spherical shell at distance r_cut observed in a given structure database,Δr_{cut} is the bin width at r_cut, the value of α (1.61) was determined by the best fit of r^{α} to the actual distance-dependent number of ideal-gas points in finite protein-size spheres19 and β is set to 0.33. We divided the atom types into 174, which includes 167 protein and 7 carbohydrate atom types.

8.2.3 Prediction protocol

The protocol for CBP prediction is as follows. First, the target structure is aligned against those templates with sequence identity ¡ 30% from the template library T562 by

structure alignment tool SPalign [42]. SP-score is employed to measure the structural similarity between template and query structures. If the structure similarity is higher than a threshold, the model for the complex structure between the query protein and the template carbohydrate is constructed by replacing the template protein structure with the query structure in the template complex structure. The model complex structure will be utilized to calculate the binding affinity by the DFIRE energy function. The binding affinity is obtained by simplifying the predicted protein model with carbon α and carbon β. If the binding affinity is lower than a threshold, the query is predicted as CBPs. If binding affinity does not pass the threshold (or structural similarity SP-Score is lower than a threshold), the query is predicted as non-carbohydrate binding proteins. These two thresholds are optimized by maximizing the Matthews correlation coefficient (MCC) (see below).

8.3 Results

8.3.1 SPalign for CBP prediction

We first examine the ability of using SP-score from SPalign for CBP prediction. SP-score is a structural-alignment score that is independent of the sizes of proteins in comparison. SP-score ranges from 0 to 2. A higher SP-score indicates higher structural similarity. A SP-score at about 0.5 indicates the same structural folds likely shared by the two structures in comparison 21. Fig. 8.1 compares the distributions of SP-scores obtained by comparing template structures to the structures in BD122 (filled bars) to those in NB2897 (open bars). The comparison is made after removing any templates with sequence identify more than 30% to the positive query structure. The result shows that only 6% non-binding targets from NB3442 have a SP-score of more than 0.6 with a template structure. By comparison, 25% of binding targets can find a template with SP-score ¿0.6. It is clear that a structure-alignment program alone can provide a reasonable prediction of CBPs. We found that SP-align can achieve the highest MCC 0.56 with sensitivity of 42% and precision of 78% for the SP-score threshold of 0.784.

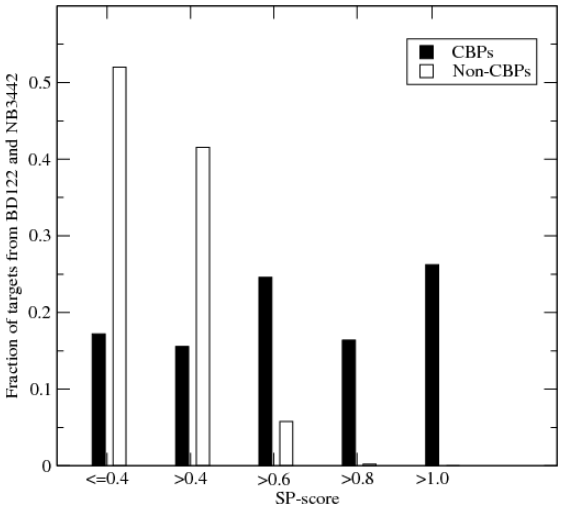

Fig. 8.1: Distributions of top SP-score ranked templates by comparing proteins in the positive BD122 (filled bars) and negative NB2987 (open bars) datasets to the template structures (T562) after excluding templates with more than 30% sequence identity to the query sequence from BD122.

Table 8.1: Performance of PSI-BLAST, SPalign, and SPOT-Stuc for DB122 and NB2987 based on leave-homolog-out cross validation

Method	Precision	Sensitivity	MCC
PSI-BLAST	90%	30%	0.51
SP-align	80%	42%	0.57
SP-align+Energy (SPOT-Struc)	88%	45%	0.62

8.3.2 Combining SP-align with DFIRE-based energy function

To further improve the prediction ability of SP-align, we combined SP-align with binding affinity based on the extended DFIRE energy function, DCBP [Equation(1)]. Two thresholds, SP-score and binding affinity, were optimized by using the leave-one-out scheme on BD122/NB3442. The grid for SP-score is 0.01. For a given SP-score, we locate the binding affinity that yields the highest MCC value. The final MCC value is 0.61 with 0.72 and -0.30 as the thresholds for SP-score and energy thresholds, respectively. The corresponding sensitivity and precision are 45% and 84%, respectively. This result indicates that combining SP-align and binding affinity can significantly improve over SP-align (9% for the MCC value, 7% for sensitivity, and 6% for precision) as shown in Table 8.1.

For a baseline comparison, we also predict CBPs by using PSIBLAST24 a commonly used tool for sequence-to-profile homolog search. We made four iterations of search by PSIBLAST utilizing the NCBI non-redundant protein sequence library. It predicts a target as CBP if the most significant template from T546 has an E-value

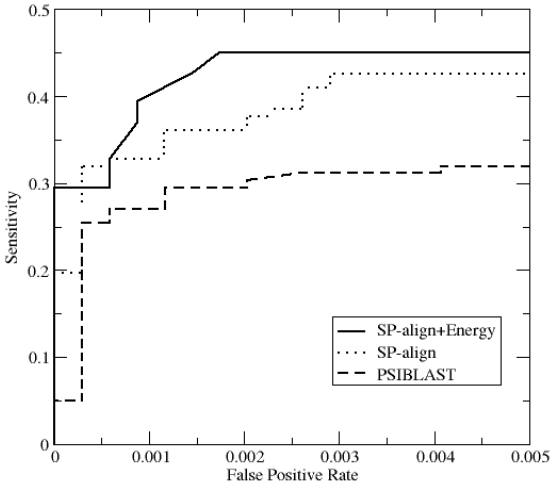

Fig. 8.2: Sensitivity versus false positive rate, given by PSI-BLAST, SPalign and SPOT-Struc (SPalign + Energy).

smaller than a threshold. As with SPalign-based techniques, the templates are removed if their sequence identities with a target are higher than 30%. The highest MCC value of PSIBLAST is 0.51 with precision of 92%, sensitivity of 30%. As shown in Table 8.1, the MCC value is 10% lower than SP-align and 20% lower than SP-align combining with energy. The combination of SP-align with energy is the most effective method in detecting CBPs. The Receiver operating characteristic (ROC) curves for PSI-BLAST, SPalign and SPalign+ Energy (SPOT-Struc) are shown in Fig. 8.2.

8.3.3 The effect of bound/unbound structures on CBP prediction (APO/HOLO dataset)

We examine the effect of bound/unbound structures on CBP prediction based on the leave-homolog-out cross validation. For a target protein, if its SP-score and binding energy value satisfies the above-optimized thresholds, it will be predicted as a CBP. The numbers of positive predictions for HOLO and APO sets are 21 and 19, respectively, and the corresponding sensitivities are 42% (19/45) and 36% (16/45), respectively.

Not all correctly predicted targets in the APO set overlap with those in the HOLO set. For 13 overlapped targets, the conformational change due to binding is small (SPscore ¿0.74). Six correctly predicted targets in HOLO are missed in APO. Two of the six targets are not predicted as CBPs because their suitable templates were excluded due to template-target sequence identities are greater than 30%. The remained four targets have significant structural changes (SP-scores¡0.2) from the corresponding

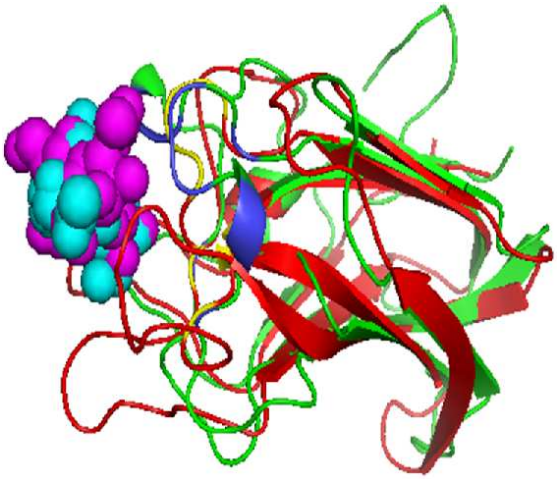

Fig. 8.3: Comparison of predicted and native binding residues for target 2j1uA. The red and green colors represent predicted and native structures, respectively. The magnate and cyan denote the template and native carbohydrate structures, respectively. The predicted and native binding residues are colored in yellow and blue, respectively.

HOLO structures. Interestingly, 3 APO targets are correctly predicted as CBPs but not the corresponding HOLO targets. These 3 APO targets have significant changes in their structures from their HOLO structures (SPscores ¡0.36). These large structural changes made them close to some of the templates that do not match to the HOLO structures. These results suggest that using APO structures does not lead to a large reduction of the sensitivity of our method.

8.3.4 Binding sites prediction

Predicted structures from SPOT-Struc can be employed to predict binding residues. A residue is defined as binding site if any heavy atom for that residue is ¡6.5 away from any heavy atom of carbohydrate. All other residues are defined as non-binding residues, regardless if they are on the surface or in the protein core. The predicted binding sites are evaluated against actual binding sites by using the MCC value, sensitivity and precision. For 54 correctly predicted CBPs from DB122, an average MCC value of 0.58 with standard deviation 0.29 was achieved with a sensitivity of 66% and a precision of 62%.

As an example, Figure 8.3 compares predicted CBP binding sites with native binding sites for target 2j1uA. This is a Fucolectin-related protein in Streptococcus pneumoniae serotype 4. For this target, the prediction achieved an MCC of 0.90 although the sequence identity between this target and template 2j7mA is only 17.3%.

Table 8.2: Structural genome targets predicted as CBPs

Target	Template	SP-score	Energy	Function
1t9fA	1v6vA2	0.788	-2.2	CBP[a]
2jz4A1	1vbpA	0.744	-1.8	CBP
1vdwA	2qvrA	0.883	-1.7	CBP
1y89B	2ri1A	0.952	-1.7	CBP
3hnmA	2j1tA	0.758	-5.7	CBP
1mtpA	8apiA	1.207	-2.0	NB[b]
3e5zA	1ms1A1	0.734	-2.4	CBP
3ejnA	3ck7B	0.822	-3.2	NB
1ny1A	1w1a1	1.368	-3.0	CBP
3eypA2	2j1uA	0.828	-2.2	CBP
3ebvA	2dt1A	0.842	-5.2	CBP
3cbwA	2cipA	0.900	-9.1	CBP
1oq1A	2d6oX	0.726	-2.1	UK[c]
3gglA	2bzdA3	0.942	-1.7	CBP
3ibsA	2vdkB	0.788	-0.5	NB
1xpwA	2v72A	0.881	0.9	CBP
1p1mA	2vhlA	0.759	-0.3	NB
1ni9A	2r8tA	1.219	1.0	CBP
1ujtA	2q7nA5	0.724	-1.9	CBP
1zoxA	1mfbH1	0.818	-1.3	NB
2p4oA	1ms1A1	0.733	-1.0	CBP

[a] Having putative function related to carbohydrate-binding. [b] Function unknown. [c] Annotated with other functions.

8.3.5 Application to structural genomics targets

This method was further applied to 2076 structural genomics domains. The trained thresholds (0.72 for SP-score and -0.30 for the binding energy) were employed. Twenty one targets from 2076 domains were predicted as CBPs. Among them, 15 out of 21 (71%) are annotated as putative CBPs by NCBI annotations [The NCBI BioSystems database]. One target is with unknown functions (1oq1A). The remained five targets (1mtpA, 3ejnA, 3ibsA, 1p1mA and 1zoxA) are annotated with other functions . Among these proteins, 2 proteins have the molecular function related with binding with others as recorded by Uniprot database. Protein 1mtpA (Tfu_1933) is a protein binding with peptide and annotated as serine-type endopeptidase inhibitor . Protein 1p1mA (MTA/SAH deaminase) is annotated as metal-binding protein. Table 8.2 lists 21 predicted CBPs.

Chapter 9 Discriminating between disease-causing and neutral non-frameshifting micro-INDELs by SVM and integration of sequence- and structure-based features

Abstract

Micro-INDELs (insertions or deletions of $\leq$ 20 bp) constitute the second most frequent class of human gene mutation after single nucleotide variants. Despite the relative abundance of non-frameshifting (NFS) INDELs, their damaging effect on protein structure and function has gone largely unstudied. We have developed such a technique (DDIG-in; Detecting DIsease-causing Genetic variations due to INDELs) by comparing the properties of disease-causing NFS-INDELs from the Human Gene Mutation Database with putatively neutral NFS-INDELs from the 1,000 Genomes Project. The final SVM model yielded a Mathews correlation coefficient of 0.68 for INDEL discrimination and is robust against annotation errors.

9.1 Introduction

The largest class of human gene mutation is the single nucleotide variant (SNV) which comprises 67% of known pathological mutations [229] .This is followed by microinsertions and microdeletions (micro-INDELs of $\leq$ 20 bp) which comprise 22% of known pathological mutations [230] . In addition, with the broad implementation of next generation sequencing (NGS) technology in genetic studies, several million polymorphic micro-INDELs have been identified and analyzed in the human genome [231–234].Many more genetic variants, including micro-INDELs, are currently being discovered at an unprecedented rate. Obviously, it is impractical to examine the impact of each variant on biological function individually. Hence, there is a critical need for effective bioinformatics tools that are capable of distinguishing potentially disease-causing variants from those that are functionally neutral.

Most available tools for prioritizing genetic variants are however limited to non-synonymous SNVs. Examples are SIFT [235] , POLYPHEN [236] , and MutPred [237] (for recent reviews, see [238–241]). These tools are not applicable to INDELs because INDELs change the number of nucleotides in the gene and hence are expected to have a much greater impact on molecular function than single nucleotide substitutions. There are two main types of INDEL within exons: frameshifting (FS) and non-frameshifting (NFS). NFS-INDELs insert/delete multiples of three nucleotides leading to the addition or removal of specific amino-acid residues at the INDEL site. FS-INDELs, on the other hand, insert/delete a discrete number of nucleotides that are indivisible by three and therefore alter the entire reading frame resulting in either a completely different amino-acid sequence C-terminal to the INDEL site, or premature termination of translation. Two bioinformatics methods were recently designed to discriminate between functional and non-functional FS-INDELs [242, 243] and nonsense mutations (premature stop codons) [242]. However, to our knowledge, there is no technique available that is capable of analyzing NFS-INDELs. Methods for interrogating FS-INDELs would not be applicable to NFS-INDELs because FS-INDELs modify the entire amino-acid sequence C-terminal to the INDEL site (unless a second INDEL were to exist), whereas NFS-INDELs simply alter the amino-acid sequence at the INDEL site. Such a technique for NFS-INDEL prioritization is urgently required because NFS-INDELs constitute a significant fraction of all exonic INDELs (theoretically, it is about one third). In practice, we found that only 26% of 9,327 exonic micro-INDELs are NFS INDELs in the 1,000 Genomes Project data [244].

In this paper, we have developed a method that we have termed DDIG-in (Detecting DIsease-causing Genetic variants due to microinsertions/microdeletions) to prioritize NFS-INDELs by comparing disease-causing INDELs from the Human Gene Mutation Database (HGMD) [229] with putatively neutral NFS-INDELs from the 1,000 Genomes Project [244] , respectively. We developed and examined a total of 58 sequence- and structure-based features of INDEL sites and found that

the feature based on predicted unstructured regions by disorder predictor SPINE-D [245] was the most discriminating one. This feature can, on its own, achieve a value of 0.56 for the Mathews Correlation Coefficient (MCC), and 0.82 for the Area Under the Receiver-Operating Characteristic (ROC) Curve (AUC). We developed two separate Support Vector Machines (SVM) methods for NFS-microdeletions and NFS-microinsertions that were 10-fold cross-validated and independently tested on microinsertions and microdeletions, respectively. A similar level of accuracy between independent testing and ten-fold cross-validation indicates not only the robustness of our training procedure but also a similar deleterious impact of NFS microdeletions and microinsertions. Of the 58 features tested, nine features were selected by maximizing the discriminatory roles for detecting disease-causing NFS microinsertions and microdeletions in a non-redundant dataset of micro-INDELs. Our DDIG-in method received further confirmation from the observation that NFS-INDEL variants with higher predicted disease-causing probabilities were characterized by lower average minor allele frequencies in the general population (based on data from the 1,000 Genomes Project). DDIG-in, is available at http://sparks-lab.org/ddig.

9.2 Methods

We tested many features for their potential roles in INDEL discrimination. These features are summarized in Table 9.1 and are described in detail below.

Nucleotide sequence-level features. We examined the following nucleotide sequence-level features as potential discriminators between disease-causing and neutral NFS-INDELs: the distances from the INDEL site to the nearest upstream and downstream splice sites and the DNA conservation score derived from phyloP(phylogenetic p-values) [246]. We examined the distances from nearest splice sites because mutations near splice sites have the potential to give rise to alternative splicing patterns [247].All DNA conservation scores were downloaded from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/phyloP46way/, based on multiple alignments of 45 vertebrate genomes to the human genome. To calculate a DNA

Table 9.1: List of all features considered.

Features	Description
Nucleotide Level	
Microdeletion/ microinsertion positions (2)	Distances to nearest 5' and 3' splicing positions
DNA conservation scores (3)	Maximum, minimum, average
Protein Level	
Evolution feature (30)	Maximum, minimum, average values (7 transition probabilities between match(M), microdeletion(D) and microinsertion(I) (MM, MI, MD, IM, II, DM, DD), 3 effective numbers of match/microinsertion/microdeletion)
Length (4)	Protein length, Microdeletion/microinsertion length, Distances to terminals
δS (1)	the INDEL-induced change to the HMM match score
Disorder score (3)	Maximum, minimum, Average
Secondary structure (12)	Maximum, minimum, Average probability (C, H, E), Predicted Secondary structure (C, H, E)
Accessible surface area (3)	Maximum, minimum, average

conservation score for a microdeletion, we considered all the deleted bases (n_{del}) plus a fixed number of bases before and after the deleted bases (the half-window size, n_{window}). We obtained the average, minimum and maximum DNA conservation scores based on phylogenetic p-values over the specified bases around the deleted bases (i.e., n_{del}+2n_{window}). For microinsertions, we considered the two bases flanking the microinsertion plus a fixed number of additional neighboring upstream and downstream bases (i.e., 2+2n_{window}). The maximum, minimum and average conservation scores for 2+2n_{window} bases were also obtained. These five nucleotide sequence-level features (2 distances+31 DNA conservation scores) were studied here to assess their utility in INDEL classification.

Protein sequence-level features. We obtained features at the amino-acid sequence level using a program called HHBlits that derives multiple protein sequence alignments based on profiles generated from hidden Markov chain models (HMM) [108] (downloaded from http://toolkit.tuebingen.mpg.de/hhblits/).This program compares two sequences at the HMM profile level and searches for homologous sequences

from the UniProt sequence database. It is a more sensitive technique than the sequence-to-profile homolog search tool PSI-Blast [134] commonly used in classifications of non-synonymous SNVs (e.g. SIFT [235]) because HHBlits employs a position-dependent gap penalty and calculates transition probabilities not only between matches of two residues (i.e. two residues from two sequences are aligned) but also between other states (match to microdeletion, match to microinsertion, microdeletion to match, microinsertion to match, microinsertion to microinsertion and microdeletion to microdeletion). That is, there are a total of seven position-dependent transition probabilities. In addition, for each position, we can obtain three effective numbers of homologous sequences (neff) aligned to microinsertion, to microdeletion and to amino-acid residues, irrespective of residue type. The maximum, minimum and average of all these amino-acid residue level properties [3$\times$(7+3)=30 features] were obtained for a specified region. For the microdeletions, this region included deleted residues plus several residues before and after the deleted residues (n_{del}+2n_{window}). For microinsertions, this region comprised the two nearest neighboring residues flanking the inserted residues plus a fixed number of residues before and after these two residues (2+2n_{window}). In addition, we calculated a protein-level feature: the change to the HMM-HMM alignment score by the whole protein sequence before and after the microdeletion or microinsertion. We also examined four features of microinsertion/microdeletion length, protein length and distances to the protein amino and carboxy terminal ends. A total of 35 features (30+1+4) were generated from protein sequences.

Protein structure-level features. The first protein structure-level feature was based on amino acid sequence-based prediction of structured and unstructured regions by a neural-network-based disorder predictor, SPINE-D [245]. We employed SPINE-D because it is among the most accurate methods based on benchmarks [245] according to the 9th Meeting for Critical Assessment of Structure Prediction Techniques (CASP 9, 2010) [245, 248]. We examined the maximum, minimum and average values of disorder probabilities over the specified region described above (n_{del}+2n_{window} for

microdeletion, 2+2n_{window} for microinsertion). In addition, we obtained predicted secondary structures, secondary structure probability, and solvent accessible surface area for the same specified region from SPINE-X [249] . SPINE-X has achieved 82% accuracy in secondary structure prediction [249] and 0.74 for the correlation coefficient between predicted and measured solvent accessible surface area (ASA) [50] based on large-scale benchmark tests. As with the disorder feature, we obtained the maximum, minimum and average values of predicted secondary probabilities in three states and predicted real-value solvent accessibility over the specified region for microdeletions or microinsertions. We also studied the fractions of three secondary structure types over the same specified region. A total of 18 structure-based features (31 disorder, 3 fractions of secondary structure types, 33 secondary structure probability and 31 ASA) were generated for studies. Dataset of Positive INDELs. The positive (disease-causing) dataset was obtained from the HGMD (HGMD Professional v. 2012.2) [229]. Initially, a total of 25,384 INDELs were identified after mapped to CCDS (20110907 version). After excluding frameshift (FS) INDELs and those INDELs that were located in an intron or at a stop codon, we obtained a dataset of 2,479 exonic disease-causing NFS-INDELs in 743 protein-coding genes. Of these, 1,998 and 481 were microdeletions and microinsertions, respectively. To examine the possible effect of homologous sequences on training our bioinformatics method, we also constructed a non-redundant dataset lacking homologous sequences that had ¿35% sequence identity between any pair of sequences. This was accomplished by pairwise sequence alignment and clustering by BlastClust [134] and only one representative sequence was chosen from each cluster. A 35% protein sequence identity cutoff was employed because this cutoff lies at the boundary that distinguishes close homologs from remote homologs [250, 251] . This removal of homologous sequences yielded 1,762 microdeletions and 445 microinsertions from 680 protein-coding genes. We also examined the overlap between microinsertion and microdeletion datasets. We considered that a microinsertion and a microdeletion were located at the same site if at least one of the two nearest neighboring residues flanking the

inserted residues in the microinsertion contributed to the deleted residues in the microdeletion. This definition yielded 21 of 743 proteins; they were CCDS13330.1, CCDS8539.1, CCDS13989.1, CCDS5313.1, CCDS2145.1, CCDS30981.1, CCDS747.1, CCDS4306.1, CCDS13858.1, CCDS5773.1, CCDS6392.1, CCDS1390.1, CCDS11892.1, CCDS14083.1, CCDS10458.1, CCDS12198.1, CCDS2463.1, CCDS11453.1, CCDS11127.1, CCDS1071.1, and CCDS45080.1. The minimal overlap suggested that the microinsertion and microdeletion sets could to all intents and purposes be treated as independent test datasets against each other.

Dataset of Putatively Neutral INDELs. The putatively neutral dataset was retrieved from the micro-INDEL variants identified during the 1000 Genomes Project (http://www.1000genomes.org/, 20101123 release), in which apparently healthy individuals from five major populations were sequenced [252] . As with the HGMD data, the INDELs were located using hg19 as the reference genome. From 9,327 exonic INDELS (excluding more than 3 million intronic INDELs), we identified a total of 2,413 NFS-INDELs of which 1,944 were microdeletions and 469 were microinsertions. These 2,413 NFS-INDELs were derived from 1,929 protein-coding genes after excluding FS-INDELs and those INDELs that were located in an intron or at a stop codon. Removal of homologous sequences (based on a protein sequence identity cut-off of 35%), yielded 1,795 microdeletions and 446 microinsertions (a total of 2241 neutral micro-INDELs) from 640 protein-coding genes. Unlike the disease-causing NFS-INDEL dataset, there was no overlap between the positions of the microdeletions and those of the microinsertions in this dataset. Minor allele frequencies were retrieved for all 2,241 NFS-INDELs from the 1000 Genomes Project. Both datasets (with and without homologous sequences) were employed to train and test our models to examine the effect of homologous sequences. It should be noted however that we cannot wholly exclude the possibility that a small subset of this putatively neutral dataset could still be of functional importance (more in the Discussion section).

9.2.1 Structural and Sequence Features

We tested many features for their potential roles in INDEL discrimination. These features are summarized in Table9.1 and are described in detail below.

Nucleotide sequence-level features. We examined the following nucleotide sequence-level features as potential discriminators between disease-causing and neutral NFS-INDELs: the distances from the INDEL site to the nearest upstream and downstream splice sites and the DNA conservation score derived from phyloP(phylogenetic p-values). We examined the distances from nearest splice sites because mutations near splice sites have the potential to give rise to alternative splicing patterns [247] . All DNA conservation scores were downloaded from http://hgdownload.cse.ucsc.edu/goldenPath/hg19/phyloP46way/, based on multiple alignments of 45 vertebrate genomes to the human genome. To calculate a DNA conservation score for a microdeletion, we considered all the deleted bases (n_{del}) plus a fixed number of bases before and after the deleted bases (the half-window size, n_{window}). We obtained the average, minimum and maximum DNA conservation scores based on phylogenetic p-values over the specified bases around the deleted bases (i.e., n_{del}+2n_{window}). For microinsertions, we considered the two bases flanking the microinsertion plus a fixed number of additional neighboring upstream and downstream bases (i.e., 2+2n_{window}). The maximum, minimum and average conservation scores for 2+2n_{window} bases were also obtained. These five nucleotide sequence-level features (2 distances+31 DNA conservation scores) were studied here to assess their utility in INDEL classification.

Protein sequence-level features. We obtained features at the amino-acid sequence level using a program called HHBlits that derives multiple protein sequence alignments based on profiles generated from hidden Markov chain models (HMM) [108] (downloaded from http://toolkit.tuebingen.mpg.de/hhblits/).This program compares two sequences at the HMM profile level and searches for homologous sequences from the UniProt sequence database. It is a more sensitive technique than the sequence-to-profile homolog search tool PSI-Blast [134] commonly used in

classifications of non-synonymous SNVs (e.g. SIFT [235]) because HHBlits employs a position-dependent gap penalty and calculates transition probabilities not only between matches of two residues (i.e. two residues from two sequences are aligned) but also between other states (match to microdeletion, match to microinsertion, microdeletion to match, microinsertion to match, microinsertion to microinsertion and microdeletion to microdeletion). That is, there are a total of seven position-dependent transition probabilities. In addition, for each position, we can obtain three effective numbers of homologous sequences (neff) aligned to microinsertion, to microdeletion and to amino-acid residues, irrespective of residue type. The maximum, minimum and average of all these amino-acid residue level properties [3$\times$(7+3)=30 features] were obtained for a specified region. For the microdeletions, this region included deleted residues plus several residues before and after the deleted residues (n_{del}+2n_{window}). For microinsertions, this region comprised the two nearest neighboring residues flanking the inserted residues plus a fixed number of residues before and after these two residues (2+2n_{window}). In addition, we calculated a protein-level feature: the change to the HMM-HMM alignment score by the whole protein sequence before and after the microdeletion or microinsertion. We also examined four features of microinsertion/microdeletion length, protein length and distances to the protein amino and carboxy terminal ends. A total of 35 features (30+1+4) were generated from protein sequences.

Protein structure-level features. The first protein structure-level feature was based on amino acid sequence-based prediction of structured and unstructured regions by a neural-network-based disorder predictor, SPINE-D [245] . We employed SPINE-D because it is among the most accurate methods based on benchmarks [245] according to the 9th Meeting for Critical Assessment of Structure Prediction Techniques (CASP 9, 2010) [245, 248] . We examined the maximum, minimum and average values of disorder probabilities over the specified region described above (n_{del}+2n_{window} for microdeletion, 2+2n_{window} for microinsertion). In addition, we obtained predicted secondary structures, secondary structure probability, and solvent accessible surface

area for the same specified region from SPINE-X [249] . SPINE-X has achieved 82% accuracy in secondary structure prediction [249] and 0.74 for the correlation coefficient between predicted and measured solvent accessible surface area (ASA) [50] based on large-scale benchmark tests. As with the disorder feature, we obtained the maximum, minimum and average values of predicted secondary probabilities in three states and predicted real-value solvent accessibility over the specified region for microdeletions or microinsertions. We also studied the fractions of three secondary structure types over the same specified region. A total of 18 structure-based features (31 disorder, 3 fractions of secondary structure types, 33 secondary structure probability and 31 ASA) were generated for studies.

Parameter Optimization for SVM. We employed LIBSVM [LIBSVM: a library for support vector machines (SVM) [http://www.csie.ntu.edu.tw/ cjlin/libsvm/] to combine the features listed above for NFS-INDEL classification. There are two parameters for SVM: a nonlinear kernel of radial basis function with a gamma parameter and the cost parameter (C) that allows a soft region for misclassification. In addition, we employed a half-window size (n_{window}) to include several amino-acid residues before and after the microdeletion/microinsertion site as defined above. For example, a half-window size of 0 would contain all residues deleted in a microdeletion and two residues flanking the inserted residues for a microinsertion. To reduce the number of parameters, a uniform widow size was applied to all features requiring a window size. A simple grid search was done with a grid of 2 ranging from -5 to 15 for logC and ranging -15 to 3 for log(gamma) and a window size ranging from 0 to 7. That is, we searched for the parameters that yielded the highest Mathews correlation coefficient (MCC) for 10-fold cross-validations (9 fold for training and 1 fold for testing) while employing all features. We also examined the dependence of MCC values on C, gamma, and n_{window} and found that MCC values change little across a wide range of C, gamma and n_{window} values (See Discussion). This served to confirm the robustness of the parameters we found.

9.2.2 Training and Cross-validation

The training set (positive and putatively neutral datasets) was randomly divided into 10 parts, nine of which were used for training, the rest for testing. This process was repeated 10 times (ten-fold cross-validation). We performed 10-fold cross-validation on SVM models for microdeletions or microinsertions only, as well as for the combined set of microdeletions and microinsertions. Microinsertion and microdeletion datasets were also used as independent test sets against each other in order to evaluate the overall robustness of the classification technique employed. In other words, the methods trained with the microinsertion set never saw the microdeletion dataset and vice versa.

9.2.3 Feature Selections

To identify the most informative subset of features, a previously described greedy feature selection algorithm for SNV classification [253] was employed. This iterative greedy algorithm starts with the feature shown to have the highest discriminatory power (disease versus neutral) based on the MCC value. The second feature was then selected on the basis that the combination of the first and the second features yielded the highest MCC value among all combinations between the first and other features. Similarly, the third feature was added to the first two if the addition of the third feature further improved MCC and the improvement was the largest obtained by comparison with the other remaining features. The iteration of adding an additional feature from the remaining features was halted if the MCC value failed to increase. Here, the MCC value was derived from the 10-fold cross validation.

9.3 Results

9.3.1 Single feature performance

We first examined the ability of a single feature to discriminate between disease-causing and neutral NFS-INDELs. Table 9.2 compares the top five performing features for microdeletions and microinsertions, separately, based on a half-window size

Table 9.2: Top five performing features for microdeletion and microinsertion discrimination.

Features	MCC[a]	AUC[b]	Precision	Recall
Deletion				
Disorder (Min, Ave, Max)	0.558,0.557,0.551	0.824,0.825,0.818	74%,74%,73%	85%, 85%, 84%
ASA[c] (Min, Ave, Max)	0.542,0.47,0.302	0.81,0.781,0.659	73%, 71%, 68%	88%, 81%, 57%
DNA conservation (Max, Ave, Min)	0.468, 0.367, 0.144	0.781, 0.742, 0,561	68%, 72%, 66%	79%, 71%, 23%
Neff[d] (Min, Ave, Max)	0.449,0.439,0.43	0.735,0.749,0.729	68%, 66%, 67%	85%,87%, 85%
Probability of sheet (Max, Min Ave)	0.32, 0.305, 0.284	0.678,0.658, 0.632	69%, 69%,64%	60%, 53%,51%
Insertion				
Disorder (Min, Max, Ave)	0.556,0.546,0.545	0.813,0.816,0.80	78%, 80%, 79%	75%,74%, 75%
ASA (Min, Ave, Max)	0.501,0.454, 0.317	0.80,0.78,0.670	71%, 78%,71%	85%, 65%,52%
Neff (Min, Ave, Max)	0.467,0.455,0.438	0.751,0.747,0.742	68%, 68%, 67%	86%, 85%, 84%
DNA conservation (Max, Ave, Min)	0.453,0.422, 0.234	0.758,0.752,0.597	72%, 74%, 76%	75%, 65%,27%
Transition probability of microinsertion to match (Min)	0.372	0.708	72%	62%

Note: Max, min, and ave are arranged in the order of MCC values. [a]MCC: Mathews correlation coefficient. [b]AUC: area under the curve. [c]ASA, solvent accessible surface area. [d]Neff: the number of effective homologous sequences aligned to residues, irrespective of residue type.

of 2 (n_{window}=2). Similar results were obtained with different window sizes (see Discussion). The results indicated that the top two performing features for microinsertions and microdeletions were both the same (disorder and solvent accessible surface area). This was followed by DNA conservation or effective number of homologous sequences aligned to residues instead of gaps. Both features represent evolutionary conservation scores but at the nucleotide and amino-acid residue levels, respectively. The effective number of homologous sequences aligned to amino-acid residues can be regarded as the conservation of amino-acid sequence position (not aligned to microdeletion or microinsertion regions). The 5th most discriminative feature was the length of microdeletion for microdeletions and transition probability for microinsertions. Inspection of Table 9.2 reveals that a single disorder feature alone can achieve an MCC value of 0.56 and an AUC of 0.82. At this MCC value, it has 74% precision and 85% recall (or sensitivity). Fig. 9.1 depicts the distributions of DNA conservation score, disorder probability, and ASA for the disease-causing and putatively neutral microdeletions (Fig. 9.1A) and microinsertions (Fig.9.1B), respectively. It is clear that the disease-causing NFS-INDELs occur more frequently within regions characterized by a greater degree of evolutionary conservation at the nucleotide level, lower disorder probability (structural regions), and lower ASA (buried core regions). The results summarized in Table 9.2 and Fig. 9.1 support the view that disruption of protein structure (and hence protein function) is the single most important reason why the NFS-INDELs are deleterious from the various features examined. Similar top-ranked features for microdeletions and microinsertions suggest that a single predictive method may be developed for microinsertions and microdeletions combined.

9.3.2 SVM for Microdeletions only

To combine different features to improve INDEL discrimination, we first employed support vector machines for the microdeletions. The microdeletion database included 1,998 disease-causing and 1,944 neutral NFS-INDELs. When all 58 features (listed

Table 9.3: List of selected features for different training sets

Deletions	Insertions	INDELs	Non-redundant INDELs
Disorder (min)	Disorder (min)	Disorder (min)	Disorder (min)
DNA conservation (max)	DNA conservation (max)	DNA conservation (max)	DNA conservation (max)
Deletion length	P(m-i)e (min)	δS[d]	δS[d]
ASA[a] (min)	δSi[d]	Neff[c] (ave)	Neff[c] (min)
P(m-d)[b] (ave)	P(m-i)[e] (ave)	Distance to protein downstream	ASA[a] (ave)
Neffc (min)	Disorder (ave)	Distance to the nearest splicing site (upstream)	INDEL length
Distance to the nearest splicing site (downstream)	Helical probability (max)	ASA[a] (max)	ASA[a] (max)
ASA[a] (max)	P(m-m)[f] (ave)	Neff[g] (min)	P(m-m)[f] (max)
δS[d]			DNA conservation (ave)
ASA (ave)			

[a]ASA, solvent accessible surface area. [b]P(m-d), match-to-deletion transition probability. [c]Neff: the number of effective homologous sequences aligned to residues. [d]δS, INDEL-induced change to alignment score. [e]P(m-i), match-to-insertion transition probability. [f]P(m-m), match-to-match transition probability. [g]Neff-del: the number of effective homologous sequences aligned to deletion.

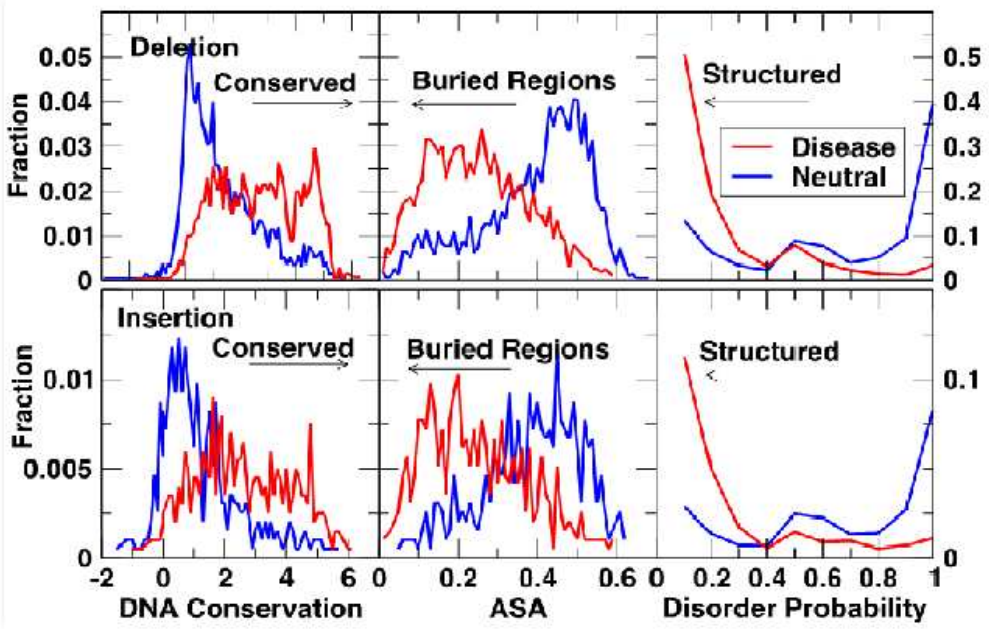

Fig. 9.1: Distributions of the average DNA conservation score from phyloP (phylogenetic p-values) (Left), the average solvent Accessible Surface Area (ASA, Middle), and the average disorder probability (Right) of disease-causing (Red) and neutral (Blue) INDELs [microdeletions (top panel) and microinsertions (bottom panel)].

in Methods) were employed, LIBSVM achieved an MCC value of 0.682, an accuracy of 84% and an AUC of 0.90 by ten-fold cross-validation. To avoid overtraining, and in order to remove redundant features, we utilized a greedy feature selection method (see Methods) and selected 10 features as shown in Table 9.3 . They were minimum disorder, maximum DNA conservation, microdeletion length, minimum ASA, average HHBlits match-to-microdeletion transition probability, the minimum effective number of aligned sequence to amino-acids, the distance to the nearest downstream splice site, maximum ASA, INDEL-induced change to matching score, and average ASA. The MCC and AUC values for this reduced feature set were 0.675 and 0.90, respectively. The precision and recall rates were 81% and 89%, respectively. The ROC curve from the ten-fold cross-validated result of the 10-feature model was compared to the results obtained from single features in Figure 9.2 (top panel). We tested the above SVM models on the microinsertion dataset. We were able to treat the microinsertion dataset as a quasi-independent test set because only 21 proteins (from 743 proteins) harbored

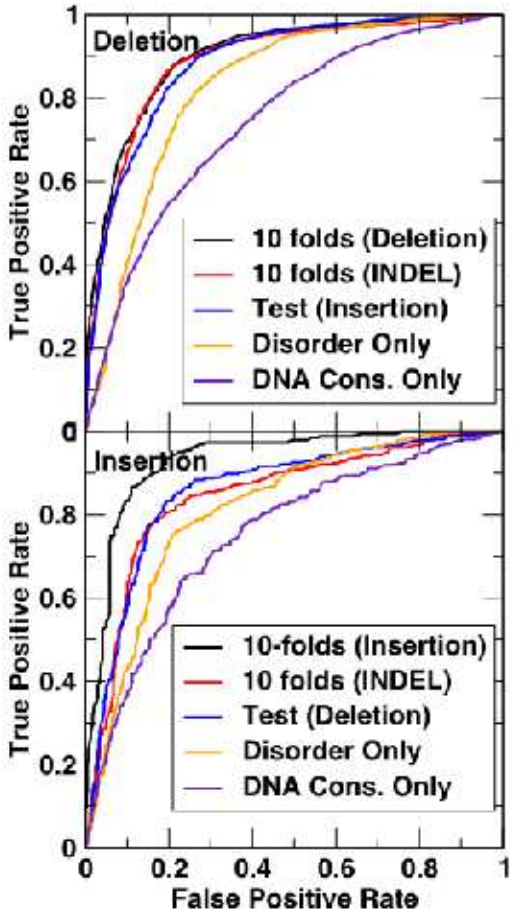

Fig. 9.2: The ROC curves for the microdeletion (Top) and microinsertion (Bottom) sets, respectively, by ten-fold cross-validation on the set (black), ten-fold cross-validation on both insertions and deletions (Red), independent test by training on the microinsertions (top) or microdeletions (bottom) (Blue), by disorder feature only (Orange) and by DNA conservation score only (Purple) as labeled.

microinsertions and microdeletions at the same location. The full 58-feature model yielded an MCC value of 0.59, an accuracy of 74%, a precision of 82%, a recall of 76%, and an AUC of 0.84. By comparison, the above 10-feature model yielded an MCC value of 0.654, an accuracy of 83%, a precision of 82%, a recall of 85%, and an AUC of 0.86. This result is indicative of the same highly discriminating power of the microdeletion-trained model for microinsertions and highlights the importance of feature selection to avoid overtraining.

9.3.3 SVM for Microinsertions only

In a similar vein, we applied SVM to perform ten-fold cross-validation on the microinsertion set and employed the greedy feature selection to remove redundant features and avoid overtraining. This yielded a total of 8 best performing features listed in Table 9.3. Three features (the minimum disorder probability, the DNA conservation, and INDEL-induced change to HMM match score) were the same as those in the

10-feature model for microdeletions. This 8-feature model achieved an MCC of 0.71, an accuracy of 86%, a precision of 85%, a recall of 86% and an AUC of 0.88. This may be compared to 0.654 for MCC, 83% for accuracy, 82% for precision, 85% for recall and 0.86 for AUC, the independent test result for the 10-feature model trained on the microdeletion dataset. The ten-fold cross-validation is more accurate than the independent test, in all probability due to the smaller size of the microinsertion dataset (only 481 and 446 disease-causing and putatively neutral microinsertions available for this analysis). Application of this 8-feature model to the microdeletion dataset as an independent test set yielded an MCC of 0.64, an accuracy of 82%, a precision of 78%, a recall of 89%, and an AUC of 0.89. This result was comparable to 0.675 for MCC, 84% for accuracy, 81% for precision, 89% for recall and 0.90 for AUC based on the10-fold cross-validation with 90% microdeletions as the training set for the 10-feature model. The ROC curve for microinsertions given by the 8-feature model (ten-fold cross-validation) is compared to the ROC curves from single features of disorder and DNA conservation and the independent test result from the 10-feature model trained on microdeletions in Fig. 9.2 (bottom panel).

9.3.4 SVM for both Microinsertions and Microdeletions

The high discriminatory power of the microdeletion-trained model for microinsertions (and vice versa) suggested that it should be possible to treat microinsertions and microdeletions as a single dataset. The same feature selection procedure yielded a total of 8 best-performing features for combined microinsertions and microdeletions as shown in Table 9.3. This set of features yielded 0.670 for MCC, 83% for accuracy and 0.89 for AUC. When we examined microdeletions and microinsertions separately, the results were 0.671 for the MCC, 84% for accuracy, and 0.89 for AUC in the case of microdeletions, 0.663 for the MCC, 83% for accuracy, and 0.88 for AUC in the case of microinsertions. The ROC curves given by the SVM model trained by both microinsertions and microdeletions yielded similarly accurate ROC curves given by

independent tests for microdeletions or microinsertions, as shown in Fig. 9.2. This further confirms the robustness of the SVM model.

9.3.5 Effect of Homologous Sequences

The above results are based on datasets which had not had any homologous sequences removed. If a method is trained on one sequence and tested on a highly homologous sequence, the resulting accuracy estimate of the method may be inflated because of the similarity of the two sequences. The presence of homologous sequences may also bias training toward a particular type of protein. To explore such a possible effect, we reconstructed the SVM model based on the non-redundant set of NFS-INDELs (2,207 disease-causing and 2,241 neutral) in which all protein sequences exhibited $\leq$ 35% sequence identity between each other (see Methods). For this non-redundant set, the greedy-feature selection yielded 9 best-performing features as shown in Table 9.3 and the final model with a ten-fold cross-validated MCC value of 0.684, accuracy of 84% precision of 81%, recall of 89% and an AUC of 0.886. Application of this model back to the set without removing homologous sequences yielded an MCC of 0.71, an accuracy of 85%, precision of 81%, recall of 92% and an AUC of 0.91. This result represented a marked improvement over 0.67 for MCC, 83% for accuracy and 0.89 for AUC by training and cross-validating the same set. This confirms the importance of removing homologous proteins prior to training our SVM model.

9.3.6 Minor allele frequency

We obtained allele frequencies for all putatively neutral NFS-microdeletions and microinsertions derived from the 1000 Genomes Project data. The allele frequency in the population should in general reflect the fitness of that allele with respect to its intended biological function [246, 254–257] . Fig. 9.3 compares average predicted disease probabilities with average allele frequencies grouped into 20 bins (bin size, 0.05). The predicted disease probabilities are based on the 10-fold cross-validation by the 9-feature model trained on both microinsertions and microdeletions after

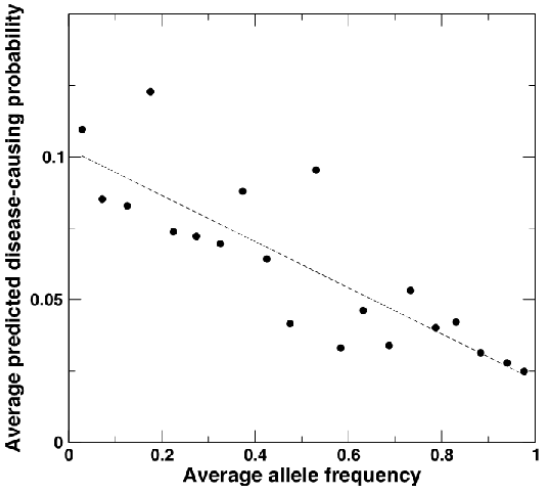

Fig. 9.3: The average predicted disease-causing probabilities as a function of the average allele frequency in the neutral INDEL dataset derived from 1000 Genomes Project data. This was done by dividing allele frequencies into 20 bins. The dashed line is from a linear regression fit. The correlation coefficient is -0.84.

removing homologous sequences. As expected, there was a strong negative correlation (correlation coefficient,-0.84), indicating that NFS-INDELs with higher predicted disease-causing probabilities tend to occur with lower allele frequencies in the general population.

9.4 Discussions

We have developed a method, termed DDIG-in, for prioritizing NFS-INDELs by predicting the disease-causing probability for a given micro-INDEL. The method is based on nucleotide and amino-acid sequences and predicted structural features of proteins. The result suggests that highly accurate and robust prediction for both microinsertions and microdeletions can be made with only 9 features. They are minimum disorder score, maximum DNA conservation score, the INDEL-induced change to the HMM alignment score, minimum effective number of aligned sequence to amino acids, average ASA, microinsertion/microdeletion length, maximum ASA, maximum HHBlits match-to-match transition probability, and average DNA conservation score. Interestingly, predicted ASA and DNA conservation are employed twice, once as the average value and a second time as the maximum value for the entire NFS-INDEL region. The difference between these two ASA or DNA conservation features measures the fluctuation of ASA or conservation for the INDEL region. The method was examined by ten-fold cross-validation as well as by an independent test.

The consistency between 10-fold cross-validations and independent tests (84-85% for accuracy, 0.88-0.90 for AUC) supports the robustness of the final method developed.

One point to consider is that the most discriminating feature was predicted disordered (or structured) regions by SPINE-D. As Table 9.2 shows, the disorder feature alone can achieve an MCC value of 0.56 for both microinsertions and microdeletions. Although predicted disorder probabilities have previously been found to be useful in SNP discrimination [237,258] , with disease-causing missense mutations being shown to be less likely to occur within disordered regions [259] , its importance has never before been shown to be so prominent. This is probably due, at least in part, to the improvement of SPINE-D over previous algorithms [245] . It may also suggest the uniqueness of NFS-INDEL classification. This result is not unexpected because fully disordered regions (Disorder probability 1) are structurally flexible and hence more permissive of modification by microinsertion or microdeletion as long as functional residues within the disordered regions remain intact. Indeed, we found that binding sites at intrinsically disordered regions of proteins are often located in semi-disordered regions (regions with a disorder probability of 0.5), consistent with near equal probability of disease-causing or neutral NFS-INDELs at disorder probability 0.5 in Fig. 9.1.

Here, we assumed from the outset that the microdeletion and microinsertion variants identified during the course of the 1,000 Genomes Project are neutral. Although this assumption is not unreasonable, it should be appreciated that the training set may contain false negatives, especially for some late-onset disorders. To examine the effect of this, we removed those neutral variants with a minor allele frequency (MAF) of $< 2\%$ and examined the effect of the removal of those variants on the accuracy and training of our NFS-INDEL discriminatory tool. This yielded 1,609 neutral cases plus 2,207 positive cases from the non-redundant set. The 10-fold cross-validation with the same 9 features, but retrained without INDELs with a MAF of $<$2%, yielded an MCC of 0.70, an accuracy of 85% and an AUC of 0.883. By comparison, application of the original 9-feature model (trained with neutral INDELs with a MAF of $<$2%)

to the set of neutral INDELs without a MAF of <2% yielded an MCC of 0.74, an accuracy of 87% and an AUC of 0.92. The fact that the 9-feature model trained without MAF <2% INDELs was less accurate than the 9-feature model trained with MAF <2% INDELs suggests that including MAF <2% INDELs (which potentially contained false negatives) facilitated machine learning. In other words, potential false negatives within the small frequency putatively neutral NFS-INDELs did not adversely affect SVM training. This is supported by strong negative correlations between the MAF and the disease-causing probability (Fig. 9.3).

To further examine the effect of potential annotation errors in our datasets, we randomly introduced 5% or 10% errors to 9 folds by assigning neutral to disease-causing and disease-causing to neutral INDELs and testing the method for the remaining 1-fold. This was repeated for 10 times. We also randomly introduced 5% or 10% errors 10 separate times to obtain an average effect. As described above, the 10-fold cross-validation with the same 9 features (Table 9.3) but retrained without INDELs with a MAF of <2% yielded an MCC of 0.696. Adding 5% and 10% errors to 9 training folds yielded the average MCC values for the test set of 0.684 and 0.674, respectively. This small change in MCC values due to 5%-10% errors confirms that our method is robust against potential assignment errors in the training set.

Another way to examine the robustness of a method is to test its dependence on various parameters. Figure 9.4 shows the Mathews correlation coefficient as a function of SVM gamma and cost parameters and the half-window size for the NFS-INDEL dataset for the case when all features were employed. It shows that MCC values change a little for the entire range of n_{window} from 0 to 7 and for a large range of gamma and cost parameters. Recently, Kumar et al. [260] found that most commonly used tools for non-synonymous SNV classification yield high false positive rates for ultra conserved sites. To examine the dependence of the accuracy of our method on conservations of INDEL sites, we calculated conservation scores according to relative entropy (RE) [$= 100 \sum_{i=1}^{20} p_i log(p_i/q_i)$] where p_i is the probability of amino acid types at a sequence position obtained from PSI-BLAST [134], and q_i is the background probability from

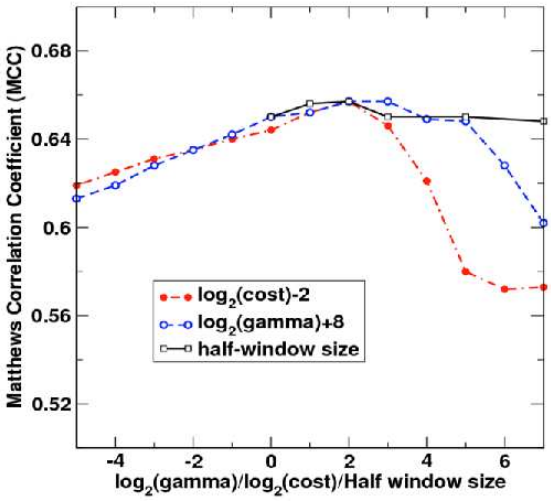

Fig. 9.4: Ten-fold cross-validated Matthew correlation coefficient for the NFS-INDEL set as a function of SVM gamma and cost parameters and half window size when trained on all features. Note that a logarithmic scale is used for gamma and cost parameters and log2(gamma) and log2(cost) are shifted to facilitate comparison.

the blosum62 matrix.We divided our dataset into three portions (high, median, low) according to the average relative entropy of deleted residues or two residues around the insertion position (RE$\geq$150, 70$\leq$RE$<$150, RE$<$70). As in Kumar et al [260]., we also observed an elevated false positive rate at highly conserved sites (33%), relative to poorly conserved sites (14%). Interestingly, the true positive rate at highly conserved sites is also higher (95% at high RE sites versus 72% at low RE sites). Thus, the overall performance of our method is not strongly dependent on conservation of INDEL sites. The MCC values are 0.67, 0.63 and 0.58 for high, median and low RE INDELs, respectively. The relative independence of our method on the conservation of INDEL sites reflects the fact that sequence conservation is not the dominant feature in our INDEL discrimination technique.

It is worthy of note that the INDEL length is one of the top features selected by SVM. This is reasonable because longer INDELs will likely have greater impact upon protein structure and function. However, it could also be due to bias in our datasets because, empirically, the majority of INDELs involve short lengths of 1 or 2 residues in both our datasets, a reflection of the inherent bias of the underlying mutational mechanism in vivo. Such an unbalanced dataset renders size-controlled or stratified sampling impossible. Thus, to determine whether the length dependence is a result of dataset bias or is instead of true functional origin would require further studies employing much larger datasets for both disease-causing and neutral INDELs.

Nevertheless, the effect of this feature on the overall accuracy is small. Removing this feature only decreases the MCC value from 0.684 to 0.664 for our non-redundant INDEL sets.

In addition to the features listed in Table 9.1, we also performed a test for the usefulness of biochemical properties of amino acid residues such as residue size and hydrophobicity scale for INDEL discrimination. This is in part because such features have been found to be effective in protein secondary structure prediction [249,261]. We examined seven representative physical parameters including a steric parameter (graph shape index), hydrophobicity, volume, polarizability, isoelectric point, helix probability, and sheet probability [249, 261]. None of these features were found to further improve the MCC value for INDEL discrimination.

This work is consistent with various studies that have examined the sequence context of microdeletions and microinsertions. These studies found that INDELs occurred non-randomly and were highly influenced by the local DNA sequence context [230, 262, 263]. This probably accounts for the success of our algorithm in NFS-INDEL classification based upon local sequence and structural information. Furthermore, microinsertions and microdeletions exhibit strong similarities in terms of the characteristics of their flanking DNA sequences, implying that they are generated by very similar underlying mechanisms [230] . Again, this accords with our ability to design a single tool capable of discriminating between microdeletions and microinsertions of pathological importance and neutral microdeletions/microinsertions.

This study focused on NFS-INDELs only because FS-INDELs would require a quite separate algorithm to effect their classification. Such an algorithm would require features based on the entire region after the INDEL site, rather than simply the local region around the INDEL site. This is because the frame-shift in FS-INDELs results either in a completely different amino-acid sequence C-terminal to the INDEL site or premature termination of translation. Expansion of DDIG-in so as to include FS-INDELs is however in progress. In the meantime, our sequence- and structure-based tool will complement two recently developed methods [242, 243] that are based on

information derived only from nucleotide and amino-acid sequences. In addition to extension to cover FS-INDELs, we intend to incorporate new features other than sequence- and structure-based features. Other such features (e.g. predicted functional regions) may well be useful in further improving the micro-INDEL classification as was previously achieved for SNP classification [238–240, 264].

Chapter 10 Conclusion

This dissertation reported a template-based method for prediction of protein functions. The idea behind this work is that combining protein structure information with binding affinity can predict protein interactions more accurately than traditional sequence/structure homology searching methods. This was made through removing false positives generated by homology searching through further filtering with predicted binding affinity. This approach was applied to prediction of RBPs, DBPs and CBPs [32, 34, 36]. For all datasets we studied, the template-based method made significant improvements over methods based on structure homology or sequence homology only. Our highly accurate function prediction methods are contributed by accurate and effective structure alignment method [42], structure prediction method [49] and knowledge-based statistical energy function [33].

The structure alignment method used in this work is SP-align [42], where a new SP-score was defined to measure structure similarity. SP-score was designed by adding a new scaling parameter to remove protein size dependency . The performance of SP-align was found better than the commonly used structure alignment method TM-align [139] on prediction of RBPs and DBPs. TM-align evaluates structure similarity by TM-score which was found dependent on protein size [49]. Two protein structure prediction tools SPARKS-X [49] and HHpred [108] were employed for the prediction of protein functions from sequence. The DFIRE-based, all atom energy functions were utilized for the prediction of binding affinity. They were shown to be more accurate than other residue-contact based energy function [32].

By integrating sequence, structure and binding affinity information, we developed a series of template-based methods for protein function prediction . They were employed to scan proteins from structure genomics and the human genomics. Proteins predicted with novel functions provide resources for hypothesis generation for biologists. Moreover, uncovered novel functions of proteins in disease pathway can help us to better understand human disease mechanisms.

By integrating protein structure and other features, we developed the first approach for discriminating the disease-causing non frame-shifting insertions or deletions of nucleotides [265]. This method was trained by SVM model based on disease-causing and neutral mutations from HGMD [229] and 1000 genomes project, respectively. The structural features, especially disorder probability, are more discriminative than transitional sequence-based features, such as DNA-conservation score. The accuracy of this method was further verified by strongly negative correlation between predicted disease probabilities and the allele frequencies observed from 1000 genomes project.

Results of this dissertation contribute to a better understanding of the roles of protein structure and binding affinity in protein functions and disease-causing mutations. It also suggests profitable to expand our template-based method beyond protein-DNA, protein-RNA, and protein-carbohydrate binding. Moreover, simultaneous prediction of protein function and binding complexes allows a deeper understanding of binding mechanisms.

References

[1] Hunt DF, Yates JR, Shabanowitz J, Winston S & Hauer CR (1986) Protein sequencing by tandem mass spectrometry. Proc Natl Acad Sci U S A 83(17): p 6233–6237.

[2] Miyashita M, Presley JM, Buchholz BA, Lam KS, Lee YM, Vogel JS & Hammock BD (2001) Attomole level protein sequencing by edman degradation coupled with accelerator mass spectrometry. Proc Natl Acad Sci U S A 98(8): p 4403–4408.

[3] Chiang YS, Gelfand TI, Kister AE & Gelfand IM (2007) New classification of supersecondary structures of sandwich-like proteins uncovers strict patterns of strand assemblage. Proteins 68(4): p 915–921.

[4] Ishihama Y, Schmidt T, Rappsilber J, Mann M, Hartl FU, Kerner MJ & Frishman D (2008) Protein abundance profiling of the escherichia coli cytosol. BMC Genomics 9: p 102.

[5] Luk YY, Tingey ML, Dickson KA, Raines RT & Abbott NL (2004) Imaging the binding ability of proteins immobilized on surfaces with different orientations by using liquid crystals. J Am Chem Soc 126(29): p 9024–9032.

[6] Pabo CO & Sauer RT (1984) Protein-dna recognition. Annu Rev Biochem 53: p 293–321.

[7] Glisovic T, Bachorik JL, Yong J & Dreyfuss G (2008) RNA-binding proteins and post-transcriptional gene regulation. FEBS Lett 582: p 1977–1986.

[8] Moore MJ (2005) From birth to death: the complex lives of eukaryotic mRNAs. Science 309: p 1514–1518.

[9] Washburn MP, Wolters D & Yates JR (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. Nat Biotechnol 19(3): p 242–247.

[10] Fournier ML, Gilmore JM, Martin-Brown SA & Washburn MP (2007) Multidimensional separations-based shotgun proteomics. Chem Rev 107(8): p 3654–3686.

[11] Carey MF, Peterson CL & Smale ST (2012) Experimental strategies for the identification of dna-binding proteins. Cold Spring Harb Protoc 2012(1): p 18–33.

[12] Lu JY, Lin YY, Sheu JC, Wu JT, Lee FJ, Chen Y, Lin MI, Chiang FT, Tai TY, Berger SL, Zhao Y, Tsai KS, Zhu H, Chuang LM & Boeke JD (2011) Acetylation of yeast ampk controls intrinsic aging independently of caloric restriction. Cell 146(6): p 969–979.

[13] Hogan DJ, Riordan DP, Gerber AP, Herschlag D & Brown PO (2008) Diverse rna-binding proteins interact with functionally related sets of rnas, suggesting an extensive regulatory system. PLoS Biol 6(10): p e255.

[14] Tsvetanova NG, Klass DM, Salzman J & Brown PO (2010) Proteome-wide search reveals unexpected rna-binding proteins in saccharomyces cerevisiae. PLoS One 5(9).

[15] Scherrer T, Mittal N, Janga SC & Gerber AP (2010) A screen for rna-binding proteins in yeast indicates dual functions for many enzymes. PLoS One 5(11): p e15499.

[16] Butter F, Scheibe M, Mrl M & Mann M (2009) Unbiased rna-protein interaction screen by quantitative proteomics. Proc Natl Acad Sci U S A 106(26): p 10626–10631.

[17] Castello A, Fischer B, Eichelbaum K, Horos R, Beckmann BM, Strein C, Davey NE, Humphreys DT, Preiss T, Steinmetz LM, Krijgsveld J & Hentze MW (2012)

Insights into rna biology from an atlas of mammalian mrna-binding proteins. Cell 149(6): p 1393–1406.

[18] Somers WS, Tang J, Shaw GD & Camphausen RT (2000) Insights into the molecular basis of leukocyte tethering and rolling revealed by structures of p- and e-selectin bound to sle(x) and psgl-1. Cell 103(3): p 467–479.

[19] Angulo J, Rademacher C, Biet T, Benie AJ, Blume A, Peters H, Palcic M, Parra F & Peters T (2006) Nmr analysis of carbohydrate-protein interactions. Methods Enzymol 416: p 12–30.

[20] Vliegenthart JF (2001) Intramolecular carbohydrate-protein interaction. Adv Exp Med Biol 491: p 133–140.

[21] Lee YC (1997) Fluorescence spectrometry in studies of carbohydrate-protein interactions. J Biochem 121(5): p 818–825.

[22] Bhardwaj N, Langlois R, Zhao G & Lu H (2005) Structure based prediction of binding residues on dna-binding proteins. Conf Proc IEEE Eng Med Biol Soc 3: p 2611–2614.

[23] Pazos F & Sternberg MJE (2004) Automated prediction of protein function and detection of functional sites from structure. Proc Natl Acad Sci U S A 101(41): p 14754–14759.

[24] Ferrer-Costa C, Gelp JL, Zamakola L, Parraga I, de la Cruz X & Orozco M (2005) Pmut: a web-based tool for the annotation of pathological mutations on proteins. Bioinformatics 21(14): p 3176–3178.

[25] Lee D, Redfern O & Orengo C (2007) Predicting protein function from sequence and structure. Nat Rev Mol Cell Biol 8(12): p 995–1005.

[26] Shanahan HP, Garcia MA, Jones S & Thornton JM (2004) Identifying dna-binding proteins using structural motifs and the electrostatic potential. Nucleic Acids Res 32(16): p 4732–4741.

[27] Cai YD & Lin SL (2003) Support vector machines for predicting rRNA-, RNA-, and DNA-binding proteins from amino acid sequence Biochimica Biophusoca Acta Proteins and Proteomics 1648: p 127–133.

[28] Yu X, Cao J, Cai Y, Shi T & Li Y (2006) Predicting rrna-, rna-, and dna-binding proteins from primary structure with support vector machines. J Theor Biol 240(2): p 175–184.

[29] Shao X, Tian Y, Wu L, Wang Y, Jing L & Deng N (2009) Predicting dna- and rna-binding proteins from sequences with kernel methods. J Theor Biol 258(2): p 289–293.

[30] Spriggs RV, Murakami Y, Nakamura H & Jones S (2009) Protein function annotation from sequence: prediction of residues interacting with RNA. Bioinformatics 25: p 1492–1497.

[31] Gao M & Skolnick J (2008) DBD-Hunter: a knowledge-based method for the prediction of DNA-protein interactions. Nucleic Acids Res 36: p 3978–3992.

[32] Zhao H, Yang Y & Zhou Y (2010) Structure-based prediction of dna-binding proteins by structural alignment and a volume-fraction corrected dfire-based energy function. Bioinformatics 26: p 1857–1863.

[33] Zhou H & Zhou Y (2002) Distance-scaled, finite ideal-gas reference state improves structure-derived potentials of mean force for structure selection and stability prediction. Protein Sci 11: p 2714–2726.

[34] Zhao H, Yang Y & Zhou Y (2011) Structure-based prediction of rna-binding domains and rna-binding sites and application to structural genomics targets. Nucleic Acids Res 39(8): p 3017–3025.

[35] Gao M & Skolnick J (2009) A threading-based method for the prediction of dna-binding proteins with application to the human genome. PLoS Comput Biol 5(11): p e1000567.

[36] Zhao H, Yang Y & Zhou Y (2011) Highly accurate and high-resolution function prediction of rna binding proteins by fold recognition and binding affinity prediction. RNA Biol 8(6).

[37] Chothia C & Lesk AM (1986) The relation between the divergence of sequence and structure in proteins. EMBO J 5(4): p 823–826.

[38] Kolodny R, Petrey D & Honig B (2006) Protein structure comparison: implications for the nature of 'fold space', and structure and function prediction. Curr Opin Struct Biol 16(3): p 393–398.

[39] Hasegawa H & Holm L (2009) Advances and pitfalls of protein structural alignment. Curr Opin Struct Biol 19(3): p 341–348.

[40] Mizuguchi K & Go N (1995) Seeking significance in three-dimensional protein structure comparisons. Curr Opin Struct Biol 5(3): p 377–382.

[41] Zhang Y & Skolnick J (2004) Scoring function for automated assessment of protein structure template quality. Proteins 57(4): p 702–710.

[42] Yang Y, Zhan J, Zhao H & Zhou Y (2012) A new size-independent score for pairwise protein structure alignment and its application to structure classification and nucleic-acid binding prediction. Proteins 80(8): p 2080–2088.

[43] Dai L & Zhou Y (2011) Characterizing the existing and potential structural space of proteins by large-scale multiple loop permutations. J Mol Biol 408(3): p 585–595.

[44] Kihara D & Skolnick J (2003) The pdb is a covering set of small protein structures. J Mol Biol 334(4): p 793–802.

[45] Zhou H & Skolnick J (2010) Improving threading algorithms for remote homology modeling by combining fragment and template comparisons. Proteins 78(9): p 2041–2048.

[46] Peng J & Xu J (2010) Low-homology protein threading. Bioinformatics 26(12): p i294–i300.

[47] Wu S & Zhang Y (2008) Muster: Improving protein sequence profile-profile alignments by using multiple sources of structure information. Proteins 72(2): p 547–556.

[48] Lobley A, Sadowski MI & Jones DT (2009) pgenthreader and pdomthreader: new methods for improved protein fold recognition and superfamily discrimination. Bioinformatics 25(14): p 1761–1767.

[49] Yang Y, Faraggi E, Zhao H & Zhou Y (2011) Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of query and corresponding native properties of templates. Bioinformatics 27(15): p 2076–2082.

[50] Faraggi E, Xue B & Zhou Y (2009) Improving the prediction accuracy of residue solvent accessibility and real-value backbone torsion angles of proteins by guided-learning through a two-layer neural network. Proteins 74(4): p 847–856.

[51] Kono H & Sarai A (1999) Structure-based prediction of dna target sites by regulatory proteins. Proteins 35(1): p 114–131.

[52] Zhang C, Liu S, Zhu Q & Zhou Y (2005) A knowledge-based energy function for protein-ligand, protein-protein, and protein-dna complexes. J Med Chem 48: p 2325–2335.

[53] Xu B, Yang Y, Liang H & Zhou Y (2009) An all-atom knowledge-based energy function for protein-dna threading, docking decoy discrimination, and prediction of transcription-factor binding profiles. Proteins 76: p 718–730.

[54] Jaroszewski L, Li Z, Krishna SS, Bakolitsa C, Wooley J, Deacon AM, Wilson IA & Godzik A (2009) Exploration of uncharted regions of the protein universe. PLoS Biol 7(9): p e1000205.

[55] Burley SK (2000) An overview of structural genomics. Nat Struct Biol 7 Suppl: p 932–934.

[56] Ahmad S & Sarai A (2004) Moment-based prediction of dna-binding proteins. J Mol Biol 341(1): p 65–71.

[57] Punta M & Ofran Y (2008) The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function. PLoS Comput Biol 4(10): p e1000160.

[58] Sadowski MI & Jones DT (2009) The sequence-structure relationship and protein function prediction. Curr Opin Struct Biol 19(3): p 357–362.

[59] Watson JD, Laskowski RA & Thornton JM (2005) Predicting protein function from sequence and structural data. Curr Opin Struct Biol 15(3): p 275–284.

[60] Ferrer-Costa C, Shanahan HP, Jones S & Thornton JM (2005) Hthquery: a method for detecting dna-binding proteins with a helix-turn-helix structural motif. Bioinformatics 21(18): p 3679–3680.

[61] Mahony S, Benos PV, Smith TJ & Golden A (2006) Self-organizing neural networks to support the discovery of dna-binding motifs. Neural Netw 19(6-7): p 950–962.

[62] Stawiski EW, Gregoret LM & Mandel-Gutfreund Y (2003) Annotating nucleic acid-binding function based on protein structure. J Mol Biol 326(4): p 1065–1079.

[63] Tjong H, Qin S & Zhou HX (2007) Pi2pe: protein interface/interior prediction engine. Nucleic Acids Res 35(Web Server issue): p W357–W362.

[64] Lee H, Tu Z, Deng M, Sun F & Chen T (2006) Diffusion kernel-based logistic regression models for protein function prediction. OMICS 10(1): p 40–55.

[65] Kumar M, Gromiha MM & Raghava GPS (2008) Prediction of rna binding sites in a protein using svm and pssm profile. Proteins 71(1): p 189–194.

[66] Langlois RE, Carson MB, Bhardwaj N & Lu H (2007) Learning to translate sequence and structure to function: identifying dna binding and membrane binding proteins. Ann Biomed Eng 35(6): p 1043–1052.

[67] Szilgyi A & Skolnick J (2006) Efficient prediction of nucleic acid binding function from low-resolution protein structures. J Mol Biol 358(3): p 922–933.

[68] Yang Y & Zhou Y (2008) Ab initio folding of terminal segments with secondary structures reveals the fine difference between two closely related all-atom statistical energy functions. Protein Sci 17: p 1212–1219.

[69] Yang Y & Zhou Y (2008) Specific interactions for ab initio folding of protein terminal regions with secondary structures. Proteins 72: p 793–803.

[70] Li W (2006) A fast clustering algorithm for analyzing highly similar compounds of very large libraries. J Chem Inf Model 46(5): p 1919–1923.

[71] Lu XJ & Olson WK (2003) 3dna: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. Nucleic Acids Res 31(17): p 5108–5121.

[72] Cheatham TE, Cieplak P & Kollman PA (1999) A modified version of the cornell et al. force field with improved sugar pucker phases and helical repeat. J Biomol Struct Dyn 16(4): p 845–862.

[73] Angarica VE, Prez AG, Vasconcelos AT, Collado-Vides J & Contreras-Moreira B (2008) Prediction of tf target sites based on atomistic models of protein-dna complexes. BMC Bioinformatics 9: p 436.

[74] Consortium U (2010) The universal protein resource (uniprot) in 2010. Nucleic Acids Res 38(Database issue): p D142–D148.

[75] Engelhardt BE, Jordan MI, Srouji JR & Brenner SE (2011) Genome-scale phylogenetic function annotation of large and diverse protein families. Genome Res 21(11): p 1969–1980.

[76] Jones S, Daley DT, Luscombe NM, Berman HM & Thornton JM (2001) Protein-rna interactions: a structural analysis. Nucleic Acids Res 29(4): p 943–954.

[77] Bhardwaj N & Lu H (2007) Residue-level prediction of dna-binding sites and its application on dna-binding protein predictions. FEBS Lett 581(5): p 1058–1066.

[78] Lin WZ, Fang JA, Xiao X & Chou KC (2011) idna-prot: identification of dna binding proteins using random forest with grey model. PLoS One 6(9): p e24756.

[79] Kumar M, Gromiha MM & Raghava GPS (2007) Identification of dna-binding proteins using support vector machines and evolutionary profiles. BMC Bioinformatics 8: p 463.

[80] Nimrod G, Schushan M, Szilgyi A, Leslie C & Ben-Tal N (2010) idbps: a web server for the identification of dna binding proteins. Bioinformatics 26(5): p 692–693.

[81] Langlois RE & Lu H (2010) Boosting the prediction and understanding of dna-binding domains from sequence. Nucleic Acids Res 38(10): p 3149–3158.

[82] Huang HL, Lin IC, Liou YF, Tsai CT, Hsu KT, Huang WL, Ho SJ & Ho SY (2011) Predicting and analyzing dna-binding domains using a systematic approach to identifying a set of informative physicochemical and biochemical properties. BMC Bioinformatics 12 Suppl 1: p S47.

[83] Zhou W & Yan H (2011) Prediction of dna-binding protein based on statistical and geometric features and support vector machines. Proteome Sci 9 Suppl 1: p S1.

[84] Kumar KK, Pugalenthi G & Suganthan PN (2009) Dna-prot: identification of dna binding proteins from protein sequence information using random forest. J Biomol Struct Dyn 26(6): p 679–686.

[85] Jones S, Shanahan HP, Berman HM & Thornton JM (2003) Using electrostatic potentials to predict dna-binding sites on dna-binding proteins. Nucleic Acids Res 31(24): p 7189–7198.

[86] Ahmad S & Sarai A (2005) Pssm-based prediction of dna binding sites in proteins. BMC Bioinformatics 6: p 33.

[87] Tsuchiya Y, Kinoshita K & Nakamura H (2005) Preds: a server for predicting dsdna-binding site on protein molecular surfaces. Bioinformatics 21(8): p 1721–1723.

[88] Hwang S, Gou Z & Kuznetsov IB (2007) Dp-bind: a web server for sequence-based prediction of dna-binding residues in dna-binding proteins. Bioinformatics 23(5): p 634–636.

[89] Xiong Y, Xia J, Zhang W & Liu J (2011) Exploiting a reduced set of weighted average features to improve prediction of dna-binding residues from 3d structures. PLoS One 6(12): p e28440.

[90] Chen YC, Wright JD & Lim C (2012) Dr_bind: a web server for predicting dna-binding residues from the protein structure based on electrostatics, evolution and geometry. Nucleic Acids Res 40(Web Server issue): p W249–W256.

[91] Dey S, Pal A, Guharoy M, Sonavane S & Chakrabarti P (2012) Characterization and prediction of the binding site in dna-binding proteins: improvement of accuracy by combining residue composition, evolutionary conservation and structural parameters. Nucleic Acids Res 40(15): p 7150–7161.

[92] Comin M, Guerra C & Dellaert F (2009) Binding balls: fast detection of binding sites using a property of spherical fourier transform. J Comput Biol 16(11): p 1577–1591.

[93] Konc J & Janezic D (2010) Probis algorithm for detection of structurally similar protein binding sites by local structural alignment. Bioinformatics 26(9): p 1160–1168.

[94] Wang L & Brown SJ (2006) Bindn: a web-based tool for efficient prediction of dna and rna binding sites in amino acid sequences. Nucleic Acids Res 34(Web Server issue): p W243–W248.

[95] Wang L, Huang C, Yang MQ & Yang JY (2010) Bindn+ for accurate prediction of dna and rna-binding residues from protein sequence features. BMC Syst Biol 4 Suppl 1: p S3.

[96] Carson MB, Langlois R & Lu H (2010) Naps: a residue-level nucleic acid-binding prediction server. Nucleic Acids Res 38(Web Server issue): p W431–W435.

[97] Cai Y, He Z, Shi X, Kong X, Gu L & Xie L (2010) A novel sequence-based method of predicting protein dna-binding residues, using a machine learning approach. Mol Cells 30(2): p 99–105.

[98] Si J, Zhang Z, Lin B, Schroeder M & Huang B (2011) Metadbsite: a meta approach to improve protein dna-binding sites prediction. BMC Syst Biol 5 Suppl 1: p S7.

[99] Yan C, Terribilini M, Wu F, Jernigan RL, Dobbs D & Honavar V (2006) Predicting dna-binding sites of proteins from amino acid sequence. BMC Bioinformatics 7: p 262.

[100] Ahmad S, Gromiha MM & Sarai A (2004) Analysis and prediction of dna-binding proteins and their binding residues based on composition, sequence and structural information. Bioinformatics 20(4): p 477–486.

[101] Nimrod G, Szilgyi A, Leslie C & Ben-Tal N (2009) Identification of dna-binding proteins using structural, electrostatic and evolutionary features. J Mol Biol 387(4): p 1040–1053.

[102] Wang L, Yang MQ & Yang JY (2009) Prediction of dna-binding residues from protein sequence information using random forests. BMC Genomics 10 Suppl 1: p S1.

[103] Wu J, Liu H, Duan X, Ding Y, Wu H, Bai Y & Sun X (2009) Prediction of dna-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature. Bioinformatics 25(1): p 30–35.

[104] Frech K, Herrmann G & Werner T (1993) Computer-assisted prediction, classification, and delimitation of protein binding sites in nucleic acids. Nucleic Acids Res 21(7): p 1655–1664.

[105] Contreras-Moreira B & Collado-Vides J (2006) Comparative footprinting of dna-binding proteins. Bioinformatics 22(14): p e74–e80.

[106] Aloy P, Querol E, Aviles FX & Sternberg MJ (2001) Automated structure-based prediction of functional sites in proteins: applications to assessing the validity of inheriting protein function from homology in genome annotation and to protein docking. J Mol Biol 311(2): p 395–408.

[107] Skolnick J, Kihara D & Zhang Y (2004) Development and large scale benchmark testing of the prospector_3 threading algorithm. Proteins 56(3): p 502–518.

[108] Remmert M, Biegert A, Hauser A & Sding J (2012) Hhblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment. Nat Methods 9(2): p 173–175.

[109] Mariani V, Kiefer F, Schmidt T, Haas J & Schwede T (2011) Assessment of template based protein structure predictions in casp9. Proteins 79 Suppl 10: p 37–58.

[110] Nowotny M, Gaidamakov SA, Ghirlando R, Cerritelli SM, Crouch RJ & Yang W (2007) Structure of human rnase h1 complexed with an rna/dna hybrid: insight into hiv reverse transcription. Mol Cell 28(2): p 264–276.

[111] Mller CW, Rey FA, Sodeoka M, Verdine GL & Harrison SC (1995) Structure of the nf-kappa b p50 homodimer bound to dna. Nature 373(6512): p 311–317.

[112] Luo D, Xu T, Watson RP, Scherer-Becker D, Sampath A, Jahnke W, Yeong SS, Wang CH, Lim SP, Strongin A, Vasudevan SG & Lescar J (2008) Insights into rna unwinding and atp hydrolysis by the flavivirus ns3 protein. EMBO J 27(23): p 3209–3219.

[113] Xu T, Sampath A, Chao A, Wen D, Nanao M, Chene P, Vasudevan SG & Lescar J (2005) Structure of the dengue virus helicase/nucleoside triphosphatase catalytic domain at a resolution of 2.4 a. J Virol 79(16): p 10278–10288.

[114] Hu S, Xie Z, Onishi A, Yu X, Jiang L, Lin J, sool Rho H, Woodard C, Wang H, Jeong JS, Long S, He X, Wade H, Blackshaw S, Qian J & Zhu H (2009) Profiling the human protein-dna interactome reveals erk2 as a transcriptional repressor of interferon signaling. Cell 139(3): p 610–622.

[115] Lukong KE, Chang KW, Khandjian EW & Richard S (2008) RNA-binding proteins in human genetic disease. Trends Genet 24: p 416–425.

[116] Han LY, Cai CZ, Lo SL, Chung MCM & Chen YZ (2004) Prediction of RNA-binding proteins from primary sequence by a support vector machine approach. RNA 10: p 355–368.

[117] Shazman S & Mandel-Gutfreund Y (2008) Classifying rna-binding proteins based on electrostatic properties. PLoS Comput Biol 4: p e1000146.

[118] Jeong E, Chung I & Miyano S (2004) A neural network method for identification of rna-interacting residues in protein Genome Informatics 15: p 105–116.

[119] Terribilini M, Lee JH, Yan C, Jernigan RL, Honavar V & Dobbs D (2006) Prediction of RNA binding sites in proteins from amino acid sequence RNA 12: p 1450–1462.

[120] Terribilini M, Sander JD, Lee JH, Zaback P, Jernigan RL, Honavar V & Dobbs D (2007) RNABindR: a server for analyzing and predicting RNA-binding sites in proteins Nucl Acids Res 35: p W578–W584.

[121] Wang Y, Xue Z, Shen G & Xu J (2008) PRINTR: Prediction of RNA binding sites in proteins using SVM and profiles Amino Acids 35: p 295–302.

[122] Cheng CW, Su ECY, Hwang JK, Sung TY & Hsu WL (2008) Predicting rna-binding sites of proteins using support vector machines and evolutionary information. BMC Bioinformatics 9 Suppl 12: p S6.

[123] Kumar M, Gromiha AM & Raghava GPS (2008) Prediction of RNA binding sites in a protein using SVM and PSSM profile Proteins 71: p 189–194.

[124] Tong J, Jiang P & Lu Zh (2008) RISP: A web-based server for prediction of RNA-binding sites in proteins Comput Methods and Programs in Biomed 90: p 148–153.

[125] Chen YC & Lim C (2008) Predicting rna-binding sites from the protein structure based on electrostatics, evolution and geometry. Nucleic Acids Res 36(5): p e29.

[126] Maetschke SR & Yuan Z (2009) Exploiting structural and topological information to improve prediction of rna-protein binding sites. BMC Bioinformatics 10: p 341.

[127] Li Q, Cao Z & Liu H (2010) Improve the prediction of RNA-binding residues using structural neighbours Protein Peptide Lett 17: p 287–296.

[128] Liu ZP, Wu LY, Wang Y, Zhang XS & Chen L (2010) Prediction of protein-RNA binding sites by a random forest method with combined features Bioinformatics 26: p 1616–1622.

[129] Perez-Cano L & Fernandez-Recio J (2010) Optimal Protein-RNA Area, OPRA: A propensity-based method to identify RNA-binding sites on proteins Proteins 78: p 25–35.

[130] Zheng S, Robertson TA & Varani G (2007) A knowledge-based potential function predicts the specificity and relative binding energy of rna-binding proteins. FEBS J 274(24): p 6378–6391.

[131] Perez-Cano L, Solernou A, Pons C & Fernandez-Recio J (2010) Structural prediction of protein-RNA interaction by computational docking with propensity-based statistical potentials Pac Symp Biocomput 15: p 293–301.

[132] Zhou H, Xue B & Zhou Y (2007) Ddomain: Dividing structures into domains using a normalized domain-domain interaction profile. Protein Sci 16: p 947–955.

[133] Murzin AG, Brenner SE, Hubbard T & Chothia C (1995) Scop: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247(4): p 536–540.

[134] Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W & Lipman DJ (1997) Gapped blast and psi-blast: a new generation of protein database search programs Nucleic Acids Research 25(17): p 3389–3402.

[135] Wang G & Dunbrack RL (2003) PISCES: a protein sequence culling server. Bioinformatics 19: p 1589–1591.

[136] Myers EW & Miller W (1988) Optimal alignments in linear space. Comput Appl Biosci 4: p 11–17.

[137] Cl?ry A, Blatter M & Allain FHT (2008) Rna recognition motifs: boring? not quite. Curr Opin Struct Biol 18(3): p 290–298.

[138] Maris C, Dominguez C & Allain FHT (2005) The rna recognition motif, a plastic rna-binding platform to regulate post-transcriptional gene expression. FEBS J 272(9): p 2118–2131.

[139] Zhang Y & Skolnick J (2005) Tm-align: a protein structure alignment algorithm based on the tm-score. Nucleic Acids Res 33(7): p 2302–2309.

[140] Tahirov TH, Temiakov D, Anikin M, Patlan V, McAllister WT, Vassylyev DG & Yokoyama S (2002) Structure of a T7 RNA polymerase elongation complex at 2.9 a resolution. Nature 420: p 43–50.

[141] Nowotny M, Gaidamakov SA, Crouch RJ & Yang W (2005) Crystal structures of RNase H bound to an RNA/DNA hybrid: substrate specificity and metal-dependent catalysis. Cell 121: p 1005–1016.

[142] Liker E, Fernandez E, Izaurralde E & Conti E (2000) The structure of the mrna export factor tap reveals a cis arrangement of a non-canonical rnp domain and an lrr domain. EMBO J 19(21): p 5587–5598.

[143] Ho DN, Coburn GA, Kang Y, Cullen BR & Georgiadis MM (2002) The crystal structure and mutational analysis of a novel rna-binding domain found in the human tap nuclear mrna export factor. Proc Natl Acad Sci U S A 99(4): p 1888–1893.

[144] Ponting CP & Russell RR (2002) Natural history of protein domains Annu Rev Biophys Biomol Struct 31: p 45–71.

[145] Bonin M, Zhu R, Klaue Y, Oberstrass J, Oesterschulze E & Nellen W (2002) Analysis of RNA flexibility by scanning force spectroscopy. Nucleic Acids Res 30: p e81.

[146] Bork P & Koonin EV (1998) Predicting functions from protein sequences–where are the bottlenecks? Nat Genet 18(4): p 313–318.

[147] Friedberg I (2006) Automated protein function prediction–the genomic challenge. Brief Bioinform 7(3): p 225–242.

[148] Perez-Iratxeta C, Palidwor G & Andrade-Navarro MA (2007) Towards completion of the earth's proteome. EMBO Rep 8(12): p 1135–1141.

[149] Kumar M, Gromiha MM & Raghava GPS (2011) Svm based prediction of rna-binding proteins using binding residues and evolutionary information. J Mol Recognit 24(2): p 303–313.

[150] Dennis G, Sherman BT, Hosack DA, Yang J, Gao W, Lane HC & Lempicki RA (2003) David: Database for annotation, visualization, and integrated discovery. Genome Biol 4(5): p P3.

[151] Ha SC, Lokanath NK, Quyen DV, Wu CA, Lowenhaupt K, Rich A, Kim YG & Kim KK (2004) A poxvirus protein forms a complex with left-handed z-dna: crystal structure of a yatapoxvirus zalpha bound to dna. Proc Natl Acad Sci U S A 101(40): p 14367–14372.

[152] Herbert A, Alfken J, Kim YG, Mian IS, Nishikura K & Rich A (1997) A z-dna binding domain present in the human editing enzyme, double-stranded rna adenosine deaminase. Proc Natl Acad Sci U S A 94(16): p 8421–8426.

[153] Placido D, Brown BA, Lowenhaupt K, Rich A & Athanasiadis A (2007) A left-handed rna double helix bound by the z alpha domain of the rna-editing enzyme adar1. Structure 15(4): p 395–404.

[154] Kawai J, Shinagawa A, Shibata K, Yoshino M, Itoh M, Ishii Y, Arakawa T, Hara A, Fukunishi Y, Konno H, Adachi J, Fukuda S, Aizawa K, Izawa M, Nishi K, Kiyosawa H, Kondo S, Yamanaka I, Saito T, Okazaki Y, Gojobori T, Bono H, Kasukawa T, Saito R, Kadota K, Matsuda H, Ashburner M, Batalov S, Casavant T, Fleischmann W, Gaasterland T, Gissi C, King B, Kochiwa H, Kuehl P, Lewis S, Matsuo Y, Nikaido I, Pesole G, Quackenbush J, Schriml LM, Staubli F, Suzuki R, Tomita M, Wagner L, Washio T, Sakai K, Okido T, Furuno M, Aono H, Baldarelli R, Barsh G, Blake J, Boffelli D, Bojunga N, Carninci P, de Bonaldo MF, Brownstein MJ, Bult C, Fletcher C, Fujita M, Gariboldi M, Gustincich S, Hill D, Hofmann M, Hume DA, Kamiya M, Lee NH, Lyons P, Marchionni L, Mashima J, Mazzarelli J, Mombaerts P, Nordone P, Ring B, Ringwald M, Rodriguez I, Sakamoto N, Sasaki H, Sato K, Schnbach C, Seya T, Shibata Y, Storch KF, Suzuki H, Toyo-oka K, Wang KH, Weitz C, Whittaker C, Wilming L, Wynshaw-Boris A, Yoshida K, Hasegawa Y, Kawaji H, Kohtsuki S, Hayashizaki Y, Team RIKENGERGPI & the FANTOM Consortium (2001)

Functional annotation of a full-length mouse cdna collection. Nature 409(6821): p 685–690.

[155] Okazaki Y, Furuno M, Kasukawa T, Adachi J, Bono H, Kondo S, Nikaido I, Osato N, Saito R, Suzuki H, Yamanaka I, Kiyosawa H, Yagi K, Tomaru Y, Hasegawa Y, Nogami A, Schnbach C, Gojobori T, Baldarelli R, Hill DP, Bult C, Hume DA, Quackenbush J, Schriml LM, Kanapin A, Matsuda H, Batalov S, Beisel KW, Blake JA, Bradt D, Brusic V, Chothia C, Corbani LE, Cousins S, Dalla E, Dragani TA, Fletcher CF, Forrest A, Frazer KS, Gaasterland T, Gariboldi M, Gissi C, Godzik A, Gough J, Grimmond S, Gustincich S, Hirokawa N, Jackson IJ, Jarvis ED, Kanai A, Kawaji H, Kawasawa Y, Kedzierski RM, King BL, Konagaya A, Kurochkin IV, Lee Y, Lenhard B, Lyons PA, Maglott DR, Maltais L, Marchionni L, McKenzie L, Miki H, Nagashima T, Numata K, Okido T, Pavan WJ, Pertea G, Pesole G, Petrovsky N, Pillai R, Pontius JU, Qi D, Ramachandran S, Ravasi T, Reed JC, Reed DJ, Reid J, Ring BZ, Ringwald M, Sandelin A, Schneider C, Semple CAM, Setou M, Shimada K, Sultana R, Takenaka Y, Taylor MS, Teasdale RD, Tomita M, Verardo R, Wagner L, Wahlestedt C, Wang Y, Watanabe Y, Wells C, Wilming LG, Wynshaw-Boris A, Yanagisawa M, Yang I, Yang L, Yuan Z, Zavolan M, Zhu Y, Zimmer A, Carninci P, Hayatsu N, Hirozane-Kishikawa T, Konno H, Nakamura M, Sakazume N, Sato K, Shiraki T, Waki K, Kawai J, Aizawa K, Arakawa T, Fukuda S, Hara A, Hashizume W, Imotani K, Ishii Y, Itoh M, Kagawa I, Miyazaki A, Sakai K, Sasaki D, Shibata K, Shinagawa A, Yasunishi A, Yoshino M, Waterston R, Lander ES, Rogers J, Birney E, Hayashizaki Y, Consortium FANTOM & Team RIKENGERGPII (2002) Analysis of the mouse transcriptome based on functional annotation of 60,770 full-length cdnas. Nature 420(6915): p 563–573.

[156] Consortium ENCODEP, Birney E, Stamatoyannopoulos JA, Dutta A, Guig R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE, Kuehn MS, Taylor CM, Neph S, Koch CM, Asthana S, Malhotra A, Adzhubei I, Greenbaum JA, Andrews RM, Flicek P, Boyle PJ, Cao H, Carter NP, Clelland

GK, Davis S, Day N, Dhami P, Dillon SC, Dorschner MO, Fiegler H, Giresi PG, Goldy J, Hawrylycz M, Haydock A, Humbert R, James KD, Johnson BE, Johnson EM, Frum TT, Rosenzweig ER, Karnani N, Lee K, Lefebvre GC, Navas PA, Neri F, Parker SCJ, Sabo PJ, Sandstrom R, Shafer A, Vetrie D, Weaver M, Wilcox S, Yu M, Collins FS, Dekker J, Lieb JD, Tullius TD, Crawford GE, Sunyaev S, Noble WS, Dunham I, Denoeud F, Reymond A, Kapranov P, Rozowsky J, Zheng D, Castelo R, Frankish A, Harrow J, Ghosh S, Sandelin A, Hofacker IL, Baertsch R, Keefe D, Dike S, Cheng J, Hirsch HA, Sekinger EA, Lagarde J, Abril JF, Shahab A, Flamm C, Fried C, Hackermller J, Hertel J, Lindemeyer M, Missal K, Tanzer A, Washietl S, Korbel J, Emanuelsson O, Pedersen JS, Holroyd N, Taylor R, Swarbreck D, Matthews N, Dickson MC, Thomas DJ, Weirauch MT, Gilbert J, Drenkow J, Bell I, Zhao X, Srinivasan KG, Sung WK, Ooi HS, Chiu KP, Foissac S, Alioto T, Brent M, Pachter L, Tress ML, Valencia A, Choo SW, Choo CY, Ucla C, Manzano C, Wyss C, Cheung E, Clark TG, Brown JB, Ganesh M, Patel S, Tammana H, Chrast J, Henrichsen CN, Kai C, Kawai J, Nagalakshmi U, Wu J, Lian Z, Lian J, Newburger P, Zhang X, Bickel P, Mattick JS, Carninci P, Hayashizaki Y, Weissman S, Hubbard T, Myers RM, Rogers J, Stadler PF, Lowe TM, Wei CL, Ruan Y, Struhl K, Gerstein M, Antonarakis SE, Fu Y, Green ED, Karaz U, Siepel A, Taylor J, Liefer LA, Wetterstrand KA, Good PJ, Feingold EA, Guyer MS, Cooper GM, Asimenos G, Dewey CN, Hou M, Nikolaev S, Montoya-Burgos JI, Lytynoja A, Whelan S, Pardi F, Massingham T, Huang H, Zhang NR, Holmes I, Mullikin JC, Ureta-Vidal A, Paten B, Seringhaus M, Church D, Rosenbloom K, Kent WJ, Stone EA, Program NISCCS, of Medicine Human Genome Sequencing Center BC, Center WUGS, Institute B, Institute CHOR, Batzoglou S, Goldman N, Hardison RC, Haussler D, Miller W, Sidow A, Trinklein ND, Zhang ZD, Barrera L, Stuart R, King DC, Ameur A, Enroth S, Bieda MC, Kim J, Bhinge AA, Jiang N, Liu J, Yao F, Vega VB, Lee CWH, Ng P, Shahab A, Yang A, Moqtaderi Z, Zhu Z, Xu X, Squazzo S, Oberley MJ, Inman D, Singer MA, Richmond TA, Munn

KJ, Rada-Iglesias A, Wallerman O, Komorowski J, Fowler JC, Couttet P, Bruce AW, Dovey OM, Ellis PD, Langford CF, Nix DA, Euskirchen G, Hartman S, Urban AE, Kraus P, Calcar SV, Heintzman N, Kim TH, Wang K, Qu C, Hon G, Luna R, Glass CK, Rosenfeld MG, Aldred SF, Cooper SJ, Halees A, Lin JM, Shulha HP, Zhang X, Xu M, Haidar JNS, Yu Y, Ruan Y, Iyer VR, Green RD, Wadelius C, Farnham PJ, Ren B, Harte RA, Hinrichs AS, Trumbower H, Clawson H, Hillman-Jackson J, Zweig AS, Smith K, Thakkapallayil A, Barber G, Kuhn RM, Karolchik D, Armengol L, Bird CP, de Bakker PIW, Kern AD, Lopez-Bigas N, Martin JD, Stranger BE, Woodroffe A, Davydov E, Dimas A, Eyras E, Hallgrmsdttir IB, Huppert J, Zody MC, Abecasis GR, Estivill X, Bouffard GG, Guan X, Hansen NF, Idol JR, Maduro VVB, Maskeri B, McDowell JC, Park M, Thomas PJ, Young AC, Blakesley RW, Muzny DM, Sodergren E, Wheeler DA, Worley KC, Jiang H, Weinstock GM, Gibbs RA, Graves T, Fulton R, Mardis ER, Wilson RK, Clamp M, Cuff J, Gnerre S, Jaffe DB, Chang JL, Lindblad-Toh K, Lander ES, Koriabine M, Nefedov M, Osoegawa K, Yoshinaga Y, Zhu B & de Jong PJ (2007) Identification and analysis of functional elements in 1by the encode pilot project. Nature 447(7146): p 799–816.

[157] Iioka H, Loiselle D, Haystead TA & Macara IG (2011) Efficient detection of rna-protein interactions using tethered rnas. Nucleic Acids Res 39(8): p e53.

[158] Galante PAF, Sandhu D, de Sousa Abreu R, Gradassi M, Slager N, Vogel C, de Souza SJ & Penalva LOF (2009) A comprehensive in silico expression analysis of rna binding proteins in normal and tumor tissue: Identification of potential players in tumor formation. RNA Biol 6(4): p 426–433.

[159] Zhang C & Darnell RB (2011) Mapping in vivo protein-rna interactions at single-nucleotide resolution from hits-clip data. Nat Biotechnol 29(7): p 607–614.

[160] Sherman BT, Huang DW, Tan Q, Guo Y, Bour S, Liu D, Stephens R, Baseler MW, Lane HC & Lempicki RA (2007) David knowledgebase: a gene-centered

database integrating heterogeneous gene annotation resources to facilitate high-throughput gene functional analysis. BMC Bioinformatics 8: p 426.

[161] Zhou H & Zhou Y (2003) Predicting the topology of transmembrane helical proteins using mean burial propensity and a hidden-markov-model-based method. Protein Sci 12(7): p 1547–1555.

[162] Kanehisa M, Goto S, Kawashima S, Okuno Y & Hattori M (2004) The kegg resource for deciphering the genome. Nucleic Acids Res 32(Database issue): p D277–D280.

[163] Robins P, Pappin DJ, Wood RD & Lindahl T (1994) Structural and functional homology between mammalian dnase iv and the 5'-nuclease domain of escherichia coli dna polymerase i. J Biol Chem 269(46): p 28535–28538.

[164] Ganesan S, Silver DP, Greenberg RA, Avni D, Drapkin R, Miron A, Mok SC, Randrianarison V, Brodie S, Salstrom J, Rasmussen TP, Klimke A, Marrese C, Marahrens Y, Deng CX, Feunteun J & Livingston DM (2002) Brca1 supports xist rna concentration on the inactive x chromosome. Cell 111(3): p 393–405.

[165] Hattori H, Liu YC, Tohnai I, Ueda M, Kaneda T, Kobayashi T, Tanabe K & Ohtsuka K (1992) Intracellular localization and partial amino acid sequence of a stress-inducible 40-kda protein in hela cells. Cell Struct Funct 17(1): p 77–86.

[166] Matsuoka S, Ballif BA, Smogorzewska A, McDonald ER, Hurov KE, Luo J, Bakalarski CE, Zhao Z, Solimini N, Lerenthal Y, Shiloh Y, Gygi SP & Elledge SJ (2007) Atm and atr substrate analysis reveals extensive protein networks responsive to dna damage. Science 316(5828): p 1160–1166.

[167] Rikova K, Guo A, Zeng Q, Possemato A, Yu J, Haack H, Nardone J, Lee K, Reeves C, Li Y, Hu Y, Tan Z, Stokes M, Sullivan L, Mitchell J, Wetzel R, Macneill J, Ren JM, Yuan J, Bakalarski CE, Villen J, Kornhauser JM, Smith B, Li D, Zhou X, Gygi SP, Gu TL, Polakiewicz RD, Rush J & Comb MJ (2007)

Global survey of phosphotyrosine signaling identifies oncogenic kinases in lung cancer. Cell 131(6): p 1190–1203.

[168] Irwin S, Vandelft M, Pinchev D, Howell JL, Graczyk J, Orr HT & Truant R (2005) Rna association and nucleocytoplasmic shuttling by ataxin-1. J Cell Sci 118(Pt 1): p 233–242.

[169] Tsubahara A, Chino N, Ishii H, Akaboshi K, Takahashi H & Saitoh M (1992) Objective parameters in magnetic resonance imaging (mri) for neuromuscular diagnosis: preliminary findings. Disabil Rehabil 14(4): p 163–167.

[170] Jnson L, Vikesaa J, Krogh A, Nielsen LK, vO Hansen T, Borup R, Johnsen AH, Christiansen J & Nielsen FC (2007) Molecular composition of imp1 ribonucleoprotein granules. Mol Cell Proteomics 6(5): p 798–811.

[171] Mittal V & Hernandez N (1997) Role for the amino-terminal region of human tbp in u6 snrna transcription. Science 275(5303): p 1136–1140.

[172] Httelmaier S, Illenberger S, Grosheva I, Rdiger M, Singer RH & Jockusch BM (2001) Raver1, a dual compartment protein, is a ligand for ptb/hnrnpi and microfilament attachment proteins. J Cell Biol 155(5): p 775–786.

[173] Hallier M, Tavitian A & Moreau-Gachelin F (1996) The transcription factor spi-1/pu.1 binds rna and interferes with the rna-binding protein p54nrb. J Biol Chem 271(19): p 11177–11181.

[174] Hughes CM, Rozenblatt-Rosen O, Milne TA, Copeland TD, Levine SS, Lee JC, Hayes DN, Shanmugam KS, Bhattacharjee A, Biondi CA, Kay GF, Hayward NK, Hess JL & Meyerson M (2004) Menin associates with a trithorax family histone methyltransferase complex and with the hoxc8 locus. Mol Cell 13(4): p 587–597.

[175] Singh A, Olowoyeye A, Baenziger PH, Dantzer J, Kann MG, Radivojac P, Heiland R & Mooney SD (2008) Mutdb: update on development of tools for

the biochemical analysis of genetic variation. Nucleic Acids Res 36(Database issue): p D815–D819.

[176] Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TKB, Gronborg M, Ibarrola N, Deshpande N, Shanker K, Shivashankar HN, Rashmi BP, Ramya MA, Zhao Z, Chandrika KN, Padma N, Harsha HC, Yatish AJ, Kavitha MP, Menezes M, Choudhury DR, Suresh S, Ghosh N, Saravana R, Chandran S, Krishna S, Joy M, Anand SK, Madavan V, Joseph A, Wong GW, Schiemann WP, Constantinescu SN, Huang L, Khosravi-Far R, Steen H, Tewari M, Ghaffari S, Blobe GC, Dang CV, Garcia JGN, Pevsner J, Jensen ON, Roepstorff P, Deshpande KS, Chinnaiyan AM, Hamosh A, Chakravarti A & Pandey A (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res 13(10): p 2363–2371.

[177] Willig TN, Draptchinskaia N, Dianzani I, Ball S, Niemeyer C, Ramenghi U, Orfali K, Gustavsson P, Garelli E, Brusco A, Tiemann C, Prignon JL, Bouchier C, Cicchiello L, Dahl N, Mohandas N & Tchernia G (1999) Mutations in ribosomal protein s19 gene and diamond blackfan anemia: wide variations in phenotypic expression. Blood 94(12): p 4294–4306.

[178] Angelini M, Cannata S, Mercaldo V, Gibello L, Santoro C, Dianzani I & Loreni F (2007) Missense mutations associated with diamond-blackfan anemia affect the assembly of ribosomal protein s19 into the ribosome. Hum Mol Genet 16(14): p 1720–1727.

[179] Bruening W, Bardeesy N, Silverman BL, Cohn RA, Machin GA, Aronson AJ, Housman D & Pelletier J (1992) Germline intronic and exonic mutations in the wilms' tumour gene (wt1) affecting urogenital development. Nat Genet 1(2): p 144–148.

[180] Jeanpierre C, Denamur E, Henry I, Cabanis MO, Luce S, Ccille A, Elion J, Peuchmaur M, Loirat C, Niaudet P, Gubler MC & Junien C (1998)

Identification of constitutional wt1 mutations, in patients with isolated diffuse mesangial sclerosis, and analysis of genotype/phenotype correlations by use of a computerized mutation database. Am J Hum Genet 62(4): p 824–833.

[181] Schumacher V, Schrer K, Whl E, Altrogge H, Bonzel KE, Guschmann M, Neuhaus TJ, Pollastro RM, Kuwertz-Brking E, Bulla M, Tondera AM, Mundel P, Helmchen U, Waldherr R, Weirich A & Royer-Pokora B (1998) Spectrum of early onset nephrotic syndrome associated with wt1 missense mutations. Kidney Int 53(6): p 1594–1600.

[182] Alexander RP, Fang G, Rozowsky J, Snyder M & Gerstein MB (2010) Annotating non-coding regions of the genome. Nat Rev Genet 11(8): p 559–571.

[183] Knig J, Zarnack K, Luscombe NM & Ule J (2011) Protein-rna interactions: new genomic technologies and perspectives. Nat Rev Genet 13(2): p 77–83.

[184] Price SR, Ito N, Oubridge C, Avis JM & Nagai K (1995) Crystallization of rna-protein complexes. i. methods for the large-scale preparation of rna suitable for crystallographic studies. J Mol Biol 249(2): p 398–408.

[185] Ke A & Doudna JA (2004) Crystallization of rna and rna-protein complexes. Methods 34(3): p 408–414.

[186] Perederina A & Krasilnikov AS (2012) Crystallization of rna-protein complexes: from synthesis and purification of individual components to crystals. Methods Mol Biol 905: p 123–143.

[187] Chen Y & Varani G (2005) Protein families and rna recognition. FEBS J 272(9): p 2088–2097.

[188] Morozova N, Allers J, Myers J & Shamoo Y (2006) Protein-rna interactions: exploring binding patterns with a three-dimensional superposition analysis of high resolution structures. Bioinformatics 22(22): p 2746–2752.

[189] Ellis JJ, Broom M & Jones S (2007) Protein-rna interactions: structural analysis and functional classes. Proteins 66(4): p 903–911.

[190] Treger M & Westhof E (2001) Statistical analysis of atomic contacts at rna-protein interfaces. J Mol Recognit 14(4): p 199–214.

[191] Shulman-Peleg A, Nussinov R & Wolfson HJ (2009) Rsitedb: a database of protein binding pockets that interact with rna nucleotide bases. Nucleic Acids Res 37(Database issue): p D369–D373.

[192] Puton T, Kozlowski L, Tuszynska I, Rother K & Bujnicki JM (2012) Computational methods for prediction of protein-rna interactions. J Struct Biol 179(3): p 261–268.

[193] Walia RR, Caragea C, Lewis BA, Towfic F, Terribilini M, El-Manzalawy Y, Dobbs D & Honavar V (2012) Protein-rna interface residue prediction using machine learning: an assessment of the state of the art. BMC Bioinformatics 13: p 89.

[194] Ahmad S & Sarai A (2011) Analysis of electric moments of rna-binding proteins: implications for mechanism and prediction. BMC Struct Biol 11: p 8.

[195] Peng CR, Liu L, Niu B, Lv YL, Li MJ, Yuan YL, Zhu YB, Lu WC & Cai YD (2011) Prediction of rna-binding proteins by voting systems. J Biomed Biotechnol 2011: p 506205.

[196] Fujishima K, Komasa M, Kitamura S, Suzuki H, Tomita M & Kanai A (2007) Proteome-wide prediction of novel dna/rna-binding proteins using amino acid composition and periodicity in the hyperthermophilic archaeon pyrococcus furiosus. DNA Res 14(3): p 91–102.

[197] Todd AE, Orengo CA & Thornton JM (2001) Evolution of function in protein superfamilies, from a structural perspective. J Mol Biol 307(4): p 1113–1143.

[198] Rost B (2002) Enzyme function less conserved than anticipated. J Mol Biol 318(2): p 595–608.

[199] Weinhold N, Sander O, Domingues FS, Lengauer T & Sommer I (2008) Local function conservation in sequence and structure space. PLoS Comput Biol 4(7): p e1000105.

[200] Jeong E, Chung IF & Miyano S (2004) A neural network method for identification of rna-interacting residues in protein. Genome Inform 15(1): p 105–116.

[201] Kim OTP, Yura K & Go N (2006) Amino acid residue doublet propensity in the protein-rna interface and its application to rna interface prediction. Nucleic Acids Res 34(22): p 6450–6460.

[202] Towfic F, Caragea C, Gemperline DC, Dobbs D & Honavar V (2010) Struct-nb: predicting protein-rna binding sites using structural features. Int J Data Min Bioinform 4(1): p 21–43.

[203] Li Q, Cao Z & Liu H (2010) Improve the prediction of rna-binding residues using structural neighbours. Protein Pept Lett 17(3): p 287–296.

[204] Zhang T, Zhang H, Chen K, Ruan J, Shen S & Kurgan L (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): p 609–628.

[205] Murakami Y, Spriggs RV, Nakamura H & Jones S (2010) Piranha: a server for the computational prediction of rna-binding residues in protein sequences. Nucleic Acids Res 38(Web Server issue): p W412–W416.

[206] Ma X, Guo J, Wu J, Liu H, Yu J, Xie J & Sun X (2011) Prediction of rna-binding residues in proteins from primary sequence using an enriched random forest model with a novel hybrid feature. Proteins 79(4): p 1230–1239.

[207] Chen Y, Kortemme T, Robertson T, Baker D & Varani G (2004) A new hydrogen-bonding potential for the design of protein-rna interactions predicts specific contacts and discriminates decoys. Nucleic Acids Res 32(17): p 5147–5162.

[208] Gabb HA, Jackson RM & Sternberg MJ (1997) Modelling protein docking using shape complementarity, electrostatics and biochemical information. J Mol Biol 272(1): p 106–120.

[209] Katchalski-Katzir E, Shariv I, Eisenstein M, Friesem AA, Aflalo C & Vakser IA (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): p 2195–2199.

[210] Prez-Cano L, Solernou A, Pons C & Fernndez-Recio J (2010) Structural prediction of protein-rna interaction by computational docking with propensity-based statistical potentials. Pac Symp Biocomput *: p 293–301.

[211] Zacharias M (2005) Attract: protein-protein docking in capri using a reduced protein model. Proteins 60(2): p 252–256.

[212] Setny P & Zacharias M (2011) A coarse-grained force field for protein-rna docking. Nucleic Acids Res 39(21): p 9118–9129.

[213] Lensink MF & Wodak SJ (2010) Docking and scoring protein interactions: Capri 2009. Proteins 78(15): p 3073–3084.

[214] Nakahara S & Raz A (2008) Biological modulation by lectins and their ligands in tumor progression and metastasis. Anticancer Agents Med Chem 8(1): p 22–36.

[215] Hakomori S (1996) Tumor malignancy defined by aberrant glycosylation and sphingo(glyco)lipid metabolism. Cancer Res 56(23): p 5309–5318.

[216] Franois KO & Balzarini J (2012) Potential of carbohydrate-binding agents as therapeutics against enveloped viruses. Med Res Rev 32(2): p 349–387.

[217] Brown A & Higgins MK (2010) Carbohydrate binding molecules in malaria pathology. Curr Opin Struct Biol 20(5): p 560–566.

[218] Fukui S, Feizi T, Galustian C, Lawson AM & Chai W (2002) Oligosaccharide microarrays for high-throughput detection and specificity assignments of carbohydrate-protein interactions. Nat Biotechnol 20(10): p 1011–1017.

[219] Tateno H, Mori A, Uchiyama N, Yabe R, Iwaki J, Shikanai T, Angata T, Narimatsu H & Hirabayashi J (2008) Glycoconjugate microarray based on an evanescent-field fluorescence-assisted detection principle for investigation of glycan-binding proteins. Glycobiology 18(10): p 789–798.

[220] Rillahan CD & Paulson JC (2011) Glycan microarrays for decoding the glycome. Annu Rev Biochem 80: p 797–823.

[221] Someya S, Kakuta M, Morita M, Sumikoshi K, Cao W, Ge Z, Hirose O, Nakamura S, Terada T & Shimizu K (2010) Prediction of carbohydrate-binding proteins from sequences using support vector machines. Adv Bioinformatics 1.

[222] Shionyu-Mitsuyama C, Shirai T, Ishida H & Yamane T (2003) An empirical approach for structure-based prediction of carbohydrate-binding sites on proteins. Protein Eng 16(7): p 467–478.

[223] Kulharia M, Bridgett SJ, Goody RS & Jackson RM (2009) Inca-sitefinder: a method for structure-based prediction of inositol and carbohydrate binding sites on proteins. J Mol Graph Model 28(3): p 297–303.

[224] Tsai KC, Jian JW, Yang EW, Hsu PC, Peng HP, Chen CT, Chen JB, Chang JY, Hsu WL & Yang AS (2012) Prediction of carbohydrate binding sites on protein surfaces with 3-dimensional probability density distributions of interacting atoms. PLoS One 7(7): p e40846.

[225] Taroni C, Jones S & Thornton JM (2000) Analysis and prediction of carbohydrate binding sites. Protein Eng 13(2): p 89–98.

[226] Malik A & Ahmad S (2007) Sequence and structural features of carbohydrate binding in proteins and assessment of predictability using a neural network. BMC Struct Biol 7: p 1.

[227] Nassif H, Al-Ali H, Khuri S & Keirouz W (2009) Prediction of protein-glucose binding sites using support vector machines. Proteins 77(1): p 121–132.

[228] Malik A, Firoz A, Jha V & Ahmad S (2010) Procarb: A database of known and modelled carbohydrate-binding protein structures with sequence-based prediction tools. Adv Bioinformatics *: p 436036.

[229] Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NST, Abeysinghe S, Krawczak M & Cooper DN (2003) Human gene mutation database (hgmd): 2003 update. Hum Mutat 21(6): p 577–581.

[230] Ball EV, Stenson PD, Abeysinghe SS, Krawczak M, Cooper DN & Chuzhanova NA (2005) Microdeletions and microinsertions causing human genetic disease: common mechanisms of mutagenesis and the role of local dna sequence complexity Human mutation 26(3): p 205–13.

[231] Mullaney JM, Mills RE, Pittard WS & Devine SE (2010) Small insertions and deletions (indels) in human genomes Human molecular genetics 19(R2): p R131–6.

[232] Mills RE, Luttig CT, Larkins CE, Beauchamp A, Tsui C, Pittard WS & Devine SE (2006) An initial map of insertion and deletion (indel) variation in the human genome Genome research 16(9): p 1182–90.

[233] Kondrashov AS & Rogozin IB (2004) Context of deletions and insertions in human coding sequences Human mutation 23(2): p 177–85.

[234] Clark TG, Andrew T, Cooper GM, Margulies EH, Mullikin JC & Balding DJ (2007) Functional constraint and small insertions and deletions in the encode regions of the human genome Genome biology 8(9): p R180.

[235] Ng PC & Henikoff S (2002) Accounting for human polymorphisms predicted to affect protein function Genome research 12(3): p 436–46.

[236] Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS & Sunyaev SR (2010) A method and server for predicting damaging missense mutations Nature methods 7(4): p 248–9.

[237] Li B, Krishnan VG, Mort ME, Xin F, Kamati KK, Cooper DN, Mooney SD & Radivojac P (2009) Automated inference of molecular mechanisms of disease from amino acid substitutions Bioinformatics 25(21): p 2744–50.

[238] Ng PC & Henikoff S (2006) Predicting the effects of amino acid substitutions on protein function Annual review of genomics and human genetics 7: p 61–80.

[239] Gonzalez-Castejon M, Marin F, Soler-Rivas C, Reglero G, Visioli F & Rodriguez-Casado A (2011) Functional non-synonymous polymorphisms prediction methods: current approaches and future developments Current medicinal chemistry 18(33): p 5095–103.

[240] Mah JT, Low ES & Lee E (2011) In silico snp analysis and bioinformatics tools: a review of the state of the art to aid drug discovery Drug discovery today 16(17-18): p 800–9.

[241] Thusberg J, Olatubosun A & Vihinen M (2011) Performance of mutation pathogenicity prediction methods on missense variants Human mutation 32(4): p 358–68.

[242] Zia A & Moses AM (2011) Ranking insertion, deletion and nonsense mutations based on their effect on genetic information BMC Bioinformatics 12: p 299.

[243] Hu J & Ng PC (2012) Predicting the effects of frameshifting indels Genome biology 13(2): p R9.

[244] Clarke L, Zheng-Bradley X, Smith R, Kulesha E, Xiao C, Toneva I, Vaughan B, Preuss D, Leinonen R, Shumway M, Sherry S & Flicek P (2012) The 1000

genomes project: data management and community access Nature methods 9(5): p 459–462.

[245] Zhang T, Faraggi E, Xue B, Dunker AK, Uversky VN & Zhou Y (2012) Spine-d: Accurate prediction of short and long disordered regions by a single neural-network based method J Biomol Struc Dyan 28(4): p 799–813.

[246] Lupski JR, Belmont JW, Boerwinkle E & Gibbs RA (2011) Clan genomics and the complex architecture of human disease Cell 147(1): p 32–43.

[247] Garcia-Blanco MA, Baraniak AP & Lasda EL (2004) Alternative splicing in disease and therapy Nature biotechnology 22(5): p 535–46.

[248] Monastyrskyy B, Fidelis K, Moult J, Tramontano A & Kryshtafovych A (2011) Evaluation of disorder predictions in casp9 Proteins Structure Function and Bioinformatics 79 (S10): p 107–118.

[249] Faraggi E, Zhang T, Yang Y, Kurgan L & Zhou Y (2011) Spine x: Improving protein secondary structure prediction by multi-step learning coupled with prediction of solvent accessible surface area and backbone torsion angles J Computational Chemistry 33: p 259–263.

[250] Blake JD & Cohen FE (2001) Pairwise sequence alignment below the twilight zone Journal of Molecular Biology 307(2): p 721–35.

[251] Dai L, Yang Y, Kim HR & Zhou Y (2010) Improving computational protein design by using structure-derived sequence profile Proteins Structure Function and Bioinformatics 78(10): p 2338 – 2348.

[252] Huang J, Ellinghaus D, Franke A, Howie B & Li Y (2012) 1000 genomes-based imputation identifies novel and refined associations for the wellcome trust case control consortium phase 1 data European journal of human genetics EJHG *.

[253] Hu J & Yan C (2008) Identification of deleterious non-synonymous single nucleotide polymorphisms using sequence-derived information BMC Bioinformatics 9: p 297.

[254] Marth GT, Yu F, Indap AR, Garimella K, Gravel S, Leong WF, Tyler-Smith C, Bainbridge M, Blackwell T, Zheng-Bradley X, Chen Y, Challis D, Clarke L, Ball EV, Cibulskis K, Cooper DN, Fulton B, Hartl C, Koboldt D, Muzny D, Smith R, Sougnez C, Stewart C, Ward A, Yu J, Xue Y, Altshuler D, Bustamante CD, Clark AG, Daly M, DePristo M, Flicek P, Gabriel S, Mardis E, Palotie A & Gibbs R (2011) The functional spectrum of low-frequency coding variation Genome biology 12(9): p R84.

[255] Tennessen JA, Bigham AW, O'Connor TD, Fu W, Kenny EE, Gravel S, McGee S, Do R, Liu X, Jun G, Kang HM, Jordan D, Leal SM, Gabriel S, Rieder MJ, Abecasis G, Altshuler D, Nickerson DA, Boerwinkle E, Sunyaev S, Bustamante CD, Bamshad MJ & Akey JM (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes Science *.

[256] Subramanian S (2012) Quantifying harmful mutations in human populations European journal of human genetics EJHG *.

[257] Gorlov IP, Gorlova OY, Frazier ML, Spitz MR & Amos CI (2011) Evolutionary evidence of the effect of rare variants on disease etiology Clinical genetics 79(3): p 199–206.

[258] Nilsson J, Grahn M & Wright AP (2011) Proteome-wide evidence for enhanced positive darwinian selection within intrinsically disordered regions in proteins Genome biology 12(7): p R65.

[259] Mort M, Evani US, Krishnan VG, Kamati KK, Baenziger PH, Bagchi A, Peters BJ, Sathyesh R, Li BA, Sun YN, Xue B, Shah NH, Kann MG, Cooper DN, Radivojac P & Mooney SD (2010) In silico functional profiling of human

disease-associated and polymorphic amino acid substitutions Human Mutation 31(3): p 335–346.

[260] Kumar S, Sanderford M, Gray VE, Ye J & Liu L (2012) Evolutionary diagnosis method for variants in personal exomes Nat Methods 9(9): p 855–6.

[261] Dor O & Zhou Y (2007) Achieving 80secondary structure prediction by large-scale training Proteins Structure Function and Bioinformatics 66(4): p 838–45.

[262] Kvikstad EM, Chiaromonte F & Makova KD (2009) Ride the wavelet: A multiscale analysis of genomic contexts flanking small insertions and deletions Genome research 19(7): p 1153–64.

[263] Tanay A & Siggia ED (2008) Sequence context affects the rate of short insertions and deletions in flies and primates Genome biology 9(2): p R37.

[264] Cooper GM & Shendure J (2011) Needles in stacks of needles: finding disease-causal variants in a wealth of genomic data Nature reviews Genetics 12(9): p 628–40.

[265] Zhao H, Yang Y, Lin H, Zhang X, Mort M, Copper DN, Liu Y & Zhou Y (2013) Ddig-in: discriminating between disease-associated and neutral non-frameshifting micro-indels. Genome Biol 14(3): p R23.

CURRICULUM VITAE

HUIYING ZHAO

EDUCATION

- Aug 2009-Aug 2013 Doctor of Philosophy

  School of Informatics Indiana University-Purdue University at Indianapolis (IUPUI)

- Jul 2007 Doctor of Engineering

  University of Science and Technology of China

- Jul 2004 Master of Science

  Department of Safety Engineering North China University

- Jul 2001 Bachelor of Science

  Department of Safety Engineering North China University

PROFESSIONAL EXPERIENCES

- *Research Assistant, 2009-2013*

  School of Informatics, Indiana University-Purdue University at Indianapolis (IUPUI)

- *Research Assistant, 2007-2009*

  School of Engineering and Technology, Indiana University-Purdue University at Indianapolis (IUPUI)

- *Research Assistant, 2004-2007*

  Department of Mechanical Engineering, University of Science and Technology of China

- *Teaching Assistant, 2001-2004*

  Department of Safety Engineering North China University

RESEARCH INTERESTS

- ♢ Classification of Human Genetic Variations
- ♢ Protein Function Prediction from Structure/Sequence
- ♢ Protein interaction with protein, DNA/RNA, Carbohydrate and small molecules.

TECHNICAL SUMMARY

Programming Languages: C/C++, Python, Shell, Perl, MPI/openMP, R, Matlab, SAS, SPSS

PEER-REVIEWED PUBLICATIONS

Google Scholar: http://scholar.google.com/citations?user=ei4g2_gAAAAJ

1. **Zhao H**†, Yang Y†, Zhou Y. Structure-based prediction of DNA-binding proteins by structural alignment and a volume-fraction corrected DFIRE-based energy function. Bioinformatics. 2010 Aug 1;26(15):1857-63.
2. **Zhao H**†, Yang Y†, Zhou Y. Highly accurate and high-resolution function prediction of RNA binding proteins by fold recognition and binding affinity prediction. RNA Biology. 2011 Nov 1;8(6).
3. Yang Y, Faraggi E, **Zhao H**, Zhou Y. Improving protein fold recognition and template-based modeling by employing probabilistic-based matching between predicted one-dimensional structural properties of the query and corresponding native properties of templates. Bioinformatics 2011 Aug 1;27(15):2076-82.
4. **Zhao H**†, Yang Y†, Zhou Y. Structure-based prediction of RNA-binding domains and RNA-binding sites and application to structural genomics targets. Nucleic Acids Research. 2011 Apr;39(8):3017-25.

5. Yang Y, Zhan J, **Zhao H**, Zhou Y. A new size-independent score for pairwise protein structure alignment and its assessment by structure classification and nucleic-acid binding prediction. Proteins. 2012 Aug;80(8):2080-8.

6. **Zhao H**†, Yang Y†, Hai Lin, Xinjun Zhang, Matthew Mort, David N. Cooper, Yunlong Liu*, and Yaoqi Zhou*. Discriminating disease-causing from neutral non-frameshifting micro-INDELs by support vector machines with integrated sequence- and structure-based features. Genome Biology, 2013 Mar 13;14(3):R23

7. Yang Y, **Zhao H**, Wang J and Zhou Y*. Sequence-based prediction of RNA binding proteins by fold recognition and binding affinity prediction. Methods in Molecular Biology.(Accepted, Book Chapter).

8. **Huiying Zhao**, Yuedong Yang, and Yaoqi Zhou*. Prediction of RNA binding proteins comes of age from low resolution to high resolution. MOL BIOSYST(Accepted, review article).

9. **Huiying Zhao**, Yuedong Yang, Sarath Chandra Janga, Cheng Kao, and Yaoqi Zhou*. Charting the unexplored RNA-binding protein atlas of the human genome by combining structure and binding predictions. (Submitted).

10. **Huiying Zhao**, Yuedong Yang, Yaoqi Zhou. Template-based prediction of carbohydrate-binding proteins, binding residues and complex structures by structural alignment and binding affinity prediction. (In preparation).