Prediction and Characterization of DNA and RNA Binding Residues from Protein Sequence: state-of-the-art, novel predictors and proteome-scale analysis

by

Jing Yan

A thesis submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Software Engineering and Intelligent Systems

Department of Electrical and Computer Engineering
University of Alberta

Abstract

Interactions between proteins and DNA/RNA play vital roles in many cellular processes and yet many of them remain to be found and characterized. Many computational methods have been developed to predict from protein sequences which parts of the proteins (so called interacting residues) are involved in these interactions. These methods can be used to find protein-RNA and protein-DNA interactions for the vast number of uncharacterized proteins. We review a comprehensive set of 30 such computational methods. We summarize them from several significant perspectives including their design, outputs and availability. We also perform empirical assessment of a subset of these methods that offer webservers using a new benchmark dataset characterized by a more complete annotation of interactions compared to the existing datasets. We show that the predictors of DNA-binding (RNA-binding) residues offer relatively strong predictive performance but they are unable to properly separate DNA- from RNA-binding residues. This substantial weakness motivates our research. Since the existing methods substantially vary in their architectures and predictions, they can be combined together to build consensuses that perhaps can offer improved predictive performance compared to the individual methods. We design and empirically assess several types of consensuses. We demonstrate that machine learning (ML)-based consensuses provide the improved predictive performance. We also formulate and execute first-of-its-kind study that targets combined prediction of DNA- and RNA-binding residues, with the goal of substantially reducing the cross predictions between DNA and RNA binding residues. We design and test three types of these novel consensuses and conclude that the approach that relies on

ML design provides better predictive quality than individual predictors and it also substantially improves discrimination between the two types of nucleic acids. As the only solution to solve the cross-prediction problem, this consensus is hard to use and time consuming to execute, given that it relies on the predictions from 8 methods that require long runtime. To this end, we develop a novel high-throughput method, DRNApred, that accurately and specifically predicts only DNA-binding and only RNA-binding residues from protein sequences. DRNApred is implemented using a new dataset with both DNA- and RNA-binding proteins, weight-based mechanism to penalize cross-predictions, and two-layered architecture. The predictions generated in both layers are based on logistic regression models constructed using a comprehensive set of sequence-derived information. We demonstrate that the novel design ideas utilized in DRNApred raise its predictive quality. DRNApred outperforms the other state-of-the-art representative methods for the prediction of DNA- or RNA-binding residues. Based on empirical test on a test dataset we show that our method substantially reduces the cross predictions. The false positives predicted by DRNApred have higher quality, since they are located nearby the native binding residues. Moreover, DRNApred outperforms the other methods for the prediction of DNA- or RNA-binding proteins. Application in human proteome confirms that DRNApred outperforms the only other runtime efficient existing method that can process such large number of proteins, BindN+, by substantially reducing the cross predictions. We show that the novel putative binding proteins predicted by DRNApred share similarities with the known annotated binding proteins indicating that DRNApred can be used to accurately discover novel DNA and RNA binding proteins in human.

Preface

This thesis is an original work conducted by Jing Yan. The research project, of which this thesis is a part, received funding from the Discovery grant (298328) from the Natural Sciences and Engineering Research Council (NSERC) of Canada to Dr. Lukasz Kurgan.

This thesis includes materials and results from the following publications (including the submitted works):

[1] **Yan J**, Friedrich S, Kurgan L. A comprehensive comparative review of sequence-based predictors of DNA- and RNA-binding residues. *Brief Bioinform*. 2016; 17(1):88-105.

[2] **Yan J**, Kurgan L. Consensus-Based Prediction of RNA and DNA Binding Residues from Protein Sequences. *6th International Conference on Pattern Recognition and Machine Intelligence*. July 2015; Warsaw, Poland, 501-511.

[3] **Yan J**, Kurgan L. A sequence-based high throughput computational method for prediction of DNA- and RNA- specific binding residues. (submitted)

Chapter 3 includes materials from Ref. [1]. Chapter 4 is based on Refs. [1] and [2]. The second author, Ms. Friedrich, contributed to the analysis of the logic based consensuses. The materials from Chapters 5 and 6 were submitted for publication. I was responsible for the data collection, data analysis, design of the other consensus models, analysis of results, and writing of the manuscripts across all Chapters.

Acknowledgements

Firstly, I would like to express my sincere thanks to my supervisor, Dr. Lukasz Kurgan, for his full support and encouraging guidance through my Ph.D. program. I am very grateful for his contributions of time and ideas in my research and all his help and advice in my future career. I could not have imaged having a better supervisor for my Ph.D. study.

I would like to thank my fellow lab mates for their collaboration and stimulating discussions. I would also like to thank all my friends for their companionship.

Last but not least, I would like to thank my parents. I could not have done this without their love and support.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AA – amino acid

AUC – area under the receiver operator characteristic curve

AULC – area under the low FPR value range in the receiver operator characteristic curve

AULRC – area under the low TPR value range of the ratio curve

AURC – area under the ratio curve

DNA – Deoxyribonucleic acid

FN – false negative

FP – false positive

FPR – false positive rate

GO – gene ontology

GO-CC – gene ontology cellular component

MCC – Matthews's correlation coefficient

ML – Machine learning

mRNA – messenger RNA

PBC – Point-biserial correlation coefficient

PCC – Pearson correlation coefficient

PDB – protein data bank

PSSM – position specific scoring matrix

RNA – Ribonucleic acid

ROC – receiver operator characteristic

rRNA – ribosomal RNA

RSA – relative solvent accessibility

SA – solvent accessibility

SS – secondary structure

SVM – Support Vector Machine

TM – template modeling

TN – true negative

TP – true positive

TPR – true positive rate

tRNA – transfer RNA

Chapter 1

Introduction

Interplay of proteins and the two types of nucleic acids: DNA and RNA, is very important since it defines and regulates many crucial cellular functions. DNA-binding proteins (i.e., proteins that interact with DNA) are driving regulation of gene expression and DNA transcription, replication and repair [1, 2]. The RNA-binding proteins that interact with several types of RNAs, such as mRNA, tRNA and rRNA, are involved in a variety of cellular functions including protein synthesis, regulation of gene expression, posttranscriptional modifications and posttranscriptional regulation [3-5]. The protein-nucleic acids interactions are studied primarily using structures of the corresponding complexes that are derived experimentally, typically with X-ray crystallography and nuclear magnetic resonance (NMR). Unfortunately, experimental methods are technically challenging and relatively expensive and thus only a small fraction of these interactions was characterized so far. In Protein Data Bank (PDB) database [6], which is the worldwide repository of structures of proteins and proteins in complex with other molecules, as of March 2016 there are only 5,438 structures on protein-DNA/RNA complexes. This is a low number compared to the several orders of magnitude larger number of known proteins, DNAs and RNAs. As of March 2016, the NCBI's RefSeq database [7] includes over 14 million of DNA and RNA transcripts and about 61 million non-redundant proteins from 58,776 organisms (source: http://www.ncbi.nlm.nih.gov/refseq/). To put these data into a context, the fraction of DNA-binding proteins among all proteins is relatively substantial and was estimated to be on average close to 3% in eukaryotic organisms and 5% in animals, which translates to about 800 proteins per an animal organism [2]. Similarly, the fraction of RNA-binding proteins was estimated to range between 2 and 8% of proteins in eukaryotic organisms [5]. A simple math reveals that assuming the most conservative estimates of 2% we should know 2% of 61 million = 1,220 thousand proteins that bind RNA and 3% of 61

million = 1,830 thousand proteins that bind DNA. The substantial and growing gap between the number of known and the number of yet to be learned DNA and RNA binding proteins motivates the need to increase the pace of the characterization of protein–DNA and protein–RNA interactions.

To this end, the existing experimental data are being used to develop time- and cost-efficient computational models that are utilized to perform automated prediction of these interactions for the millions of the uncharacterized proteins. Over the past several years a number of computational methods have been developed for the prediction of the protein-nucleic acids interactions. These methods can be categorized into two types according to the input information that they use: structure-based methods which predict the binding based on a known protein structure, and sequence-based methods which make the prediction solely from the protein sequence. Structure-based methods utilize input information derived from protein structure, typically based on shape and biophysical characteristics of the protein surface. However, structure is unknown for most of the proteins which limits utility of the structure-based methods. As of March 2016, there are only 117,240 protein structures in PDB, which is only a small fraction of the available sequence data. Therefore, it is necessary to develop reliable computational methods to identify binding from the sequence. There are two types of relevant sequence-based methods: those that predict DNA- or RNA- binding proteins and those that predict DNA- or RNA- binding residues in a protein sequence. The former type concerns a simple two-state prediction of whether a given protein sequence binds to DNA/RNA or not, while the latter is more useful and goes further by locating the binding residues (residues in contact with DNA/RNA) in the input sequence. Therefore, our focus is on the computational prediction of DNA- and RNA-binding residues from protein chains. These methods can be used to find the binding proteins in the vast sequence databases and to indicate sites of these interactions. A couple dozen of sequence-based methods that predict the DNA- or RNA- binding residues have been already published.

1.1 Motivation

The existing sequence-based methods are designed to predict either DNA-binding or RNA-binding residues. In other words, there are no methods that combine prediction of both DNA-binding and RNA-binding residues. Given that these methods were developed on dataset with only one type of binding residues, perhaps surprisingly they were never tested how well they differentiate the two types of the nucleic acid binding residues. Since DNA and RNA binding residues share similar biochemical properties, i.e., they are positively charged and have strong propensity to interact with the negatively charged phosphate backbone of DNA or RNA [8, 9], it is likely that these methods cross predict the other type of binding residues, i.e., methods for the prediction of the RNA-binding residues also predict DNA-binding residues and vice versa. This is an important problem because DNA and RNA binding residues carry out different cellular function and they should not be confused. Besides, most of the existing methods require a substantial amount of runtime, which makes it very difficult to apply them on large scale of thousands of proteins (human has ~70 thousand unique proteins). This necessitates the development of high-throughput (characterized by a low runtime) methods that specifically predicts one type of the nucleic acid-binding residues.

Moreover, the existing methods are designed on different datasets and assessed with different evaluation criteria, which makes it difficult for end users to understand and compare their predictive performance. Several efforts have been made to comparatively review the published predictors of the DNA-binding residues and the RNA-binding residues [10-14]. However, these reviews only summarize a small number of published methods and cover interactions with just one of the two nucleic acids types (Chapter 3 provides more details on this topic). Similarly, these comparative analyses focus solely on the prediction of one type of the nucleic acid-binding residues. Consequently, these studies do not consider how well the predictive methods separate between DNA and RNA interactions. Another drawback of the prior reviews is that their comparative analyses utilize datasets that are characterized by incomplete annotations of binding residues. This is because the annotations are based on a single structure of protein–DNA or protein–RNA complex, which could be incomplete if only a fragment of DNA or RNA

is considered in a given complex or if the same protein is involved in other binding events with nucleic acids.

Although many predictors exist, not much effort was made to exploit consensus designs, i.e., meta-methods that combine multiple predictors together. The use of consensuses was shown to result in an improved predictive performance when compared to the use of individual methods in related research area, such as the sequence-based prediction of secondary structure and intrinsic disorder [15-20]. The already considered consensuses of predictors of nucleic acids-binding residues [12, 14] use only simple designs (like a simple weighted average). These works did not compare and explore different ways to generate the consensus but just demonstrated that a given, one design is successful. Once again, these studies also did not investigate the potential problem with the cross prediction between DNA-binding and RNA-binding residues.

1.2 Goals

The overall objective of my thesis is to accurately and in high-throughput predict protein-nucleic acid interactions from protein sequences, particularly focusing on differentiating between DNA- and RNA-binding residues. To achieve this goal we address the following four goals:

1. **Assessment of predictive performance of existing sequence-based DNA- and RNA- binding residue predictors.** We review a comprehensive set of the sequence-based DNA-binding residue and RNA-binding residue predictors, access predictive quality of all available to the end user methods on new benchmark dataset with both DNA- and RNA- binding proteins, and focus our analysis on how well these predictors separate between DNA and RNA interactions. (Chapter 3)

2. **Development of novel consensus-based predictors to improve accuracy of the prediction of DNA- and RNA- binding residues.** Motivated by the availability of many predictors and success of consensuses in other related areas, we investigate the development of consensus predictors with the aim of improving the predictive performance. We consider a wide range of designs to build consensus-based predictor of

DNA-binding residues and another consensus for the RNA-binding residues by combining prediction from the available DNA- and RNA-binding residues methods, respectively. We also design a novel consensus for the combined prediction of DNA- and RNA-binding residues to improve discrimination between DNA- and RNA-binding residues. (Chapter 4)

3. **Development of DRNApred, a new high-throughput method that accurately and specifically predicts only DNA-binding and only RNA-binding residues.** Using information derived from the protein sequence, we design a new high-throughput predictor of the DNA- and RNA-binding residues. DRNApred is designed to offer good predictive performance and to solve the problem of cross prediction between DNA-binding and RNA-binding residues. Our method is also runtime efficient and can be applied on proteomic scale. (Chapter 5)

4. **Identification of known and novel DNA- and RNA-binding residues/proteins on proteomic-scale.** We apply the new high-throughput method to preform predictions on the entire set of all human proteins (human proteome). We assess predictive performance of our method by quantifying whether our method specifically targets each of the two types of nucleic acid binding residues. We also generate new putative RNA and DNA binding proteins and assess whether they are predicted accurately. (Chapter 6).

1.3 Thesis organization

This thesis is organized into seven chapters:

Chapter 2 introduces the sequence-based computational method for the prediction of DNA- and RNA-binding residues. It includes biological background concerning proteins, DNA, RNAs and their interactions. It also covers technologies that are used to determine the protein-DNA/RNA interactions, focusing on the recent studies that predict protein-DNA/RNA interactions from protein sequence. Finally, this chapter also includes the computational background, such as the principles of the design and evaluation of computational methods.

Chapter 3 provides a comprehensive assessment of predictive performance of the existing sequence-based DNA- and RNA-binding residues predictors focusing on the methods that are conveniently available to end users as webservers. In particular, we assess these methods based on how well they perform on datasets with just DNA-binding proteins, just RNA-binding proteins and both DNA and RNA binding proteins.

Chapter 4 concerns the design and evaluation of a consensus-based predictor that combines results of the DNA (RNA) binding residue predictors to improve predictive performance. It explores a comprehensive range of designs of consensuses including a simple logic based combination of methods, majority vote consensus, a more sophisticated machine learning based consensus, and a combined prediction of DNA- and RNA- binding residues. The predictive performance of these various consensuses is assessed on datasets with DNA-binding or RNA-binding proteins, as well as on a dataset with both DNA-binding and RNA-binding proteins.

Chapter 5 introduces a new method, DRNApred, which accurately, specifically, and in high throughput fashion predicts DNA-binding and RNA-binding residues. We describe the novel dataset that we collected to design this method, summarize how this method was designed, and evaluate the predictive quality and runtime of our method.

Chapter 6 summarizes results of a large scale application of our method to identify DNA- and RNA-binding residues/proteins in the entire human proteome. We assess predictive quality of these results using the already known binding proteins and the newly predicted binding proteins.

Chapter 7 summarizes the thesis and lists conclusions and major contributions.

Chapter 2

Background

2.1 DNA, RNAs and proteins

Deoxyribonucleic acid (DNA) is a double-stranded macromolecule that stores genetic information. It is composed of four types of nucleotides: adenine (A), guanine (G), cytosine (C) and thymine (T). Gene is a segment of DNA that contains the genetic information that defines a protein. There are between several and over a dozens of thousands of genes in a given DNA molecule, depending on a complexity of the corresponding organism. To encode proteins, gene information is transcribed into a messenger RNA (mRNA) in a process called transcription, see Figure 2.1.

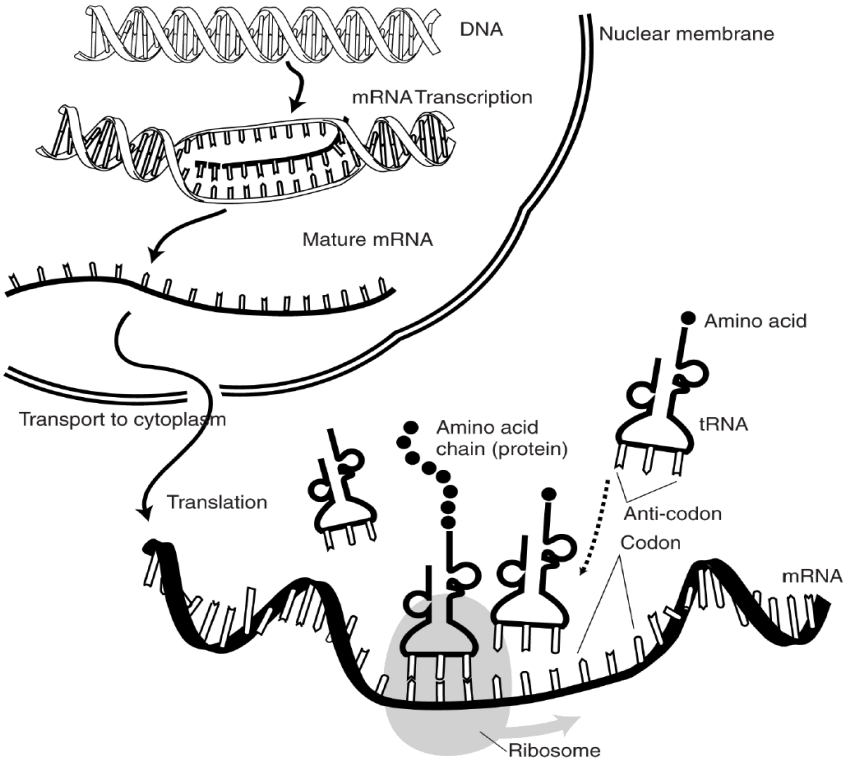


Figure 2.1. Diagram that summarizes how proteins are generated from the information encoded in genes (source: http://www.contexo.info/DNA_Basics/Protein_synthesis.htm).

Ribonucleic acid (RNA) is a single-stranded macromolecule. Similar to DNA, RNA is also made up of four nucleotides, but thymine (T) is replaced with uracil (U). RNA is more versatile than DNA and it comes in a variety of shapes and types. The three most known types of RNA are: messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA). mRNA carries the genetic information copied from genes. tRNA helps to translate the nucleotide sequence from mRNA into amino acid sequence that makes up proteins. rRNA associates with a set of proteins to form ribosome that implements the translation process (process of translating mRNA into proteins). As shown in Figure 2.1, with ribosome moving along the mRNA, amino acids are added one by one to form an amino acid chain of the corresponding protein. The different order of nucleotide sequences in genes translates to different sequences of amino acid that are held together by peptide bounds.

Table 2.1. Table of 20 amino acids along with their abbreviation names and selected physiochemical properties.

Amino acid	Abbreviation		Charge	Polarity	Hydrophobicity
	3-letter	1-letter			
Alanine	Ala	A	neutral	nonpolar	hydrophobic
Arginine	Arg	R	positive	polar	hydrophilic
Asparagine	Asn	N	neutral	polar	hydrophilic
Aspartic Acid	Asp	D	negative	polar	hydrophilic
Cysteine	Cys	C	neutral	nonpolar	hydrophobic
Glutamic Acid	Glu	E	negative	polar	hydrophilic
Glutamine	Gln	Q	neutral	polar	hydrophilic
Glycine	Gly	G	neutral	nonpolar	hydrophobic
Histidine	His	H	positive	polar	hydrophilic
Isoleucine	Ile	I	neutral	nonpolar	hydrophobic
Leucine	Leu	L	neutral	nonpolar	hydrophobic
Lysine	Lys	K	negative	polar	hydrophilic
Methionine	Met	M	neutral	nonpolar	hydrophobic
Phenylalanine	Phe	F	neutral	nonpolar	hydrophobic
Proline	Pro	P	neutral	nonpolar	hydrophobic
Serine	Ser	S	neutral	polar	hydrophilic
Threonine	Thr	T	neutral	polar	hydrophobic
Tryptophan	Trp	W	neutral	nonpolar	hydrophobic
Tyrosine	Tyr	Y	neutral	polar	hydrophobic
Valine	Val	V	neutral	nonpolar	hydrophobic

Proteins are macromolecules consisting of one or more sequences (linear chains) of amino acids. There are 20 types of amino acids and each has different chemical structure and properties. Table 2.1 lists the names of the 20 amino acids, their three- and one-letter abbreviations, and some of their physiochemical properties, such as charge, polarity and hydrophobicity. Different linear combination of the 20 amino acids fold into different three-dimensional shapes, which in turn define how these proteins interact with other molecules to carry out their functions. There are four distinct levels of protein structure. The primary structure is defined by the order of amino acids in the protein sequence. The secondary structure refers to spatially local arrangements of the sequences into three major types of regular structures: alpha-helix, beta-sheet and coil. The alpha-helix structure looks like a coiled spring, where the protein chain is assembled along a helical path. The beta-sheet structure is composed of pairs of strands (linear segments of protein chain) that lie alongside each other to form as a sheet. The coil is highly dynamic and does not have one specific and stable structure like the alpha-helix or beta-sheet. Its primary role is to connect the helices and sheets together. The tertiary structure is a spatial arrangement of these secondary structures and is defined by position of each amino acid (and all its atoms) in the three-dimensional space. Multiple sequences can aggregate together to form the quaternary structure. The primary and secondary structures determine the tertiary structure, i.e., a protein with a given primary and secondary structure will always fold into the same tertiary structure. In turn, a given tertiary structure determines what other molecules (such as nucleic acids, other proteins, nucleotides, peptides, etc.) can interact this this protein. Some proteins are characterized by lack of a fixed or stable secondary and tertiary structure, and they are called intrinsically disordered proteins. Many of these disordered proteins can undergo transitions to ordered states upon interacting with (binding) other molecules (e.g. DNA, RNA). The structural flexibility of disordered proteins facilitates their ability to form multiple conformations (three-dimensional shapes) and the corresponding ability to bind to different targets. Thus, the disordered proteins are usually enriched in proteins participating in binding related functions. Another structural property of proteins is the solvent accessibility (SA). Some amino acids are buried inside a protein, while some other amino acids are on the surface of a protein. Since proteins inhabit aqueous

environment the amino acids on their surface are exposed to the solvent. Solvent accessibility measures the area of the exposed surface of each amino acid in a given protein. Relative solvent accessibility (RSA) is the relative exposure calculated by normalizing the solvent accessibility of the amino acid in the structure by the maximal possible solvent accessibility that this amino acid can take. Both SA and RSA can be used to identify amino acids as being either exposed to solvent or buried in the structure. These are important structural properties as the binding sites (site of the interaction with the other molecules) are typically on the protein surface and thus they are composed of the solvent exposed amino acids.

2.2 Protein-DNA/RNA interactions

Proteins implement and regulate all cellular processes but they rarely act alone. Vast majority of protein functions happens via interactions with other molecules, like DNA, RNA, and other proteins. Protein-DNA interactions (a protein binds either a single or double stranded DNA) play essential roles in a variety of biological processes, such as activation or repression of gene expression, and chromosome packaging in the cell nucleus that involves interactions between DNA and histone proteins. Figure 2.2 shows the sequence (primary structure) together with the annotation of binding residues and the corresponding three-dimensional structure of the DNA-binding protein Hnt3. This protein is involved in repair of breaks in single-stranded and double-stranded DNA and base excision repair [21]. The protein-RNA interactions (a protein binds either single or double stranded RNA) are more diverse compared to the protein-DNA interactions. This is because the structure of RNA varies more widely than that of DNA. Protein-RNA interactions play key roles in the post-translational processes, such as splicing, mRNA transcript stability, mRNA localization and translation.

Given the importance of the protein-DNA/RNA interactions, various techniques have been developed to study them. Determining whether proteins interact with DNA/RNA or not and which amino acid in the interacting protein bind to DNA/RNA are vital to understanding principles underlining protein-nucleic acid binding and help us to understand the various roles these interactions play in regulating cellular processes.

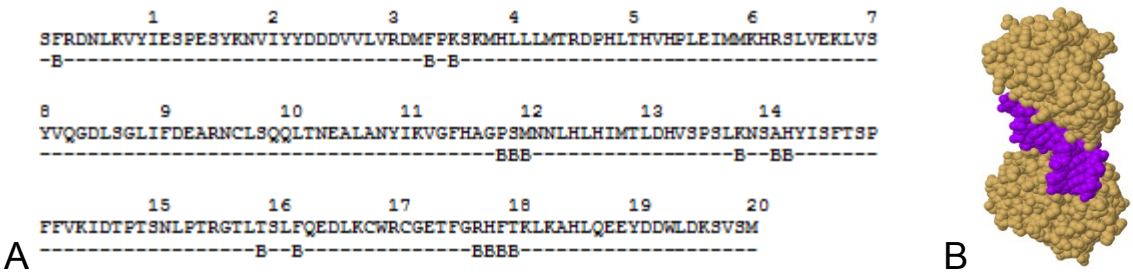


Figure 2.2. Interaction of DNA with aprataxin ortholog Hnt3 (PDB ID: 3SPD).

Panel A shows the protein-DNA interaction in the primary structure. Protein Hnt3 has 200 (disordered regions are omitted) amino acids coded with single-letter. We split the sequence into three rows where the first line is the residue number, second line is the amino acid sequence, and third line gives annotation of DNA-binding residues where B is for binding residue and – for nonbinding residue. Panel B shows the interaction between protein and DNA in the 3D molecular view. DNA is shown in purple and protein is shown in bronze.

2.2.1 Experimental technologies to determine protein-DNA/RNA interactions

A number of wet lab techniques have been developed to study the complex interactions of proteins with DNA/RNA. They include electrophoretic mobility shift assays (EMSAs) [22], conventional chromatin immunoprecipitation (ChIP) [23], MicroChIP [24, 25], Fast ChIP [26], X-ray crystallography [27], and nuclear magnetic resonance (NMR) spectroscopy [28]. The most definitive way to identity and study protein-DNA and protein-RNA interactions is to determine the structure of their complex using X-ray crystallography.

The X-ray crystallography determines the coordinates of all atoms of a crystal in 3D space. This technique usually involves three steps as shown in Figure 2.3. The first step is to generate the protein crystal. A crystal is a solid form in which all atoms are arranged in a highly ordered structure. A protein crystal that includes a large number of protein molecules arranged into a regular structure is required since single protein molecule is far too small to be detected in the X-ray diffraction experiment. Crystal is used to magnify the signal since multiple copies of the same molecule can be used to increase accuracy. This step is the most challenging part of X-ray crystallography as growing protein crystal requires quite pure protein samples and highly restricted aqueous environment. In the next step, an intense beam of X-ray is shot through a protein crystal diffracting it into many specific directions, and consequently producing a specific diffraction pattern on a

film. A crystal is rotated so that X-ray hits all its sides and all angles of diffraction patterns can be recorded. These recorded data are combined into a 3D diffraction pattern. Finally, the electron density map is calculated based on the diffraction pattern to reconstruct the crystal structure. This map is a plot the electron clouds that can be used to determine average positions of atoms in the crystal. Next, these atomic positions combined with the underlying knowledge of the amino acid sequence and exact atomic composition of each amino acid, are utilize to derive 3D atomic model of the protein. A series of refinements are often further carried out to perfect the model. Resolution measured in unit of Angstrom (Å) is a primary measurement of the accuracy of the model. The higher the resolution is (lower value in Å), the more precise the model structure is. X-ray crystallography is widely used to generate protein-DNA and protein-RNA complexes since it can provide a highly detailed atomic view of the two interacting molecules. This is the main method that was used to generate structures in PDB that we use in this work. However, determination of structures of these complexes is costly (between about $20 thousand to a few hundreds of thousands of dollars, depending on the size and complexity of the complex) and time consuming. The number of protein-DNA, protein-RNA complexes in the PDB has grown rapidly in recently years, but still lags far behind the number of known protein sequences. Hence, there is a pressing need to develop accurate computational methods to predict protein-RNA and protein-DNA binding residues from protein sequences, which are cheaper and faster to use compared to the X-ray crystallography.

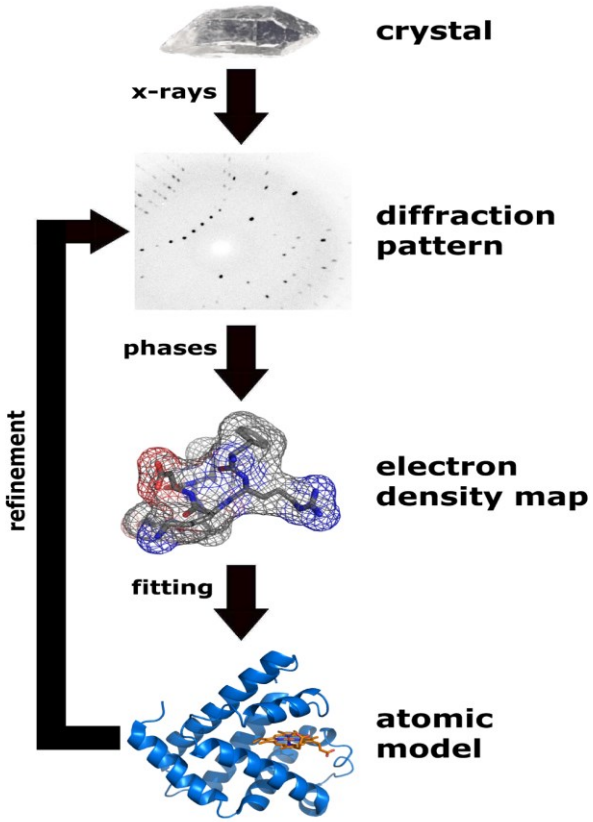


Figure 2.3. The workflow of how X-ray crystallography is used to solve the 3D structure of a protein molecule (source: https://en.wikipedia.org/wiki/X-ray_crystallography).

2.3 Prediction of protein-DNA/RNA binding residues

The experimentally solved protein-DNA/RNA complex structures provide a rich source of data that can be are used to analyze structural and sequence characteristics of the interacting residues (interface). These findings are in turn used to build and test computational method that predicts DNA- or RNA- binding residues in unbound proteins. Based on the input information used, there are two types of computational methods: structure-based methods and sequence-based methods.

2.3.1 Structure-based method

Structure-based methods perform prediction by employing knowledge of protein structure, in particular surface of the protein, to find whether this protein interacts with a

nucleic acid. The methods utilize information concerning charge on the surface, evolutionary conservation (some residues are preserved in the sequences of the same protein in different organisms, which points to their functional importance), geometry of the surface, contacts (spatial proximity) between residues on the surface, etc. In principle, there are two types of structure-based methods for the prediction of the DNA- or RNA-binding residues. The template-based methods utilize structural alignment protocols to detect significant structural similarity between the query protein structure and a set of proteins that are known to bind DNA/RNA. The premise here is that similar proteins share similar functions. An example method, SPOT-struc that was developed by Zhou and colleagues [29], aligns structure of the input/query protein to a library of structures of proteins that are known to bind RNA, and the query protein is predicted to bind RNA if the structural similarity between the query and any protein from the library is higher than a certain threshold. The second type of the structure-based methods is template-free. The template-free methods do not perform a direct structural comparison but instead they use structural features extracted from the structure of the query protein and machine-learning algorithms to scan the surface of the query protein and predict surface residues as either binding or non-binding. Various machine learning (ML) methods are used in the template-free methods including SVM [30-32], neural network [33], and random forest [34]. Research concerning the structure-based method is outside of the scope of this thesis, so we do not further elaborate on this class of methods. Additional details can be found in recent review papers on this topic [11, 35-37]. These review articles summarize the existing structure-based methods and discuss the corresponding benchmark datasets, architectures of these methods (structural features and machine learning algorithms), evaluation protocols, and assessment of their predictive performance.

2.3.2 Sequence-based method

Sequence-based methods perform the predictions solely from the sequence, without the need for the expensive and time-consuming process of experimental determination of protein structure. The sequence-based prediction methods are the only choice for majority of the proteins that do not have structures. This is why we focus on the analysis and development of this class of methods. We review the existing sequence-based predictors

of DNA- and RNA-binding residues. We discuss how they define binding residues, overview their predictive models and summarize their outputs.

The sequence-based methods published before 2015 include 14 methods for the prediction of DNA-binding residues and 16 for the RNA-binding residues; they are summarized in Table 2.2. Perhaps their most striking characteristic is that these predictors define binding residues in different ways. We note that binding residues are typically defined based on the structure of the protein-RNA or protein-DNA complexes, but only the sequence is used to perform predictions. Virtually all predictors, except for DNABindR [38, 39] and PRINTR [40], define a given residue as binding if at least one of its atoms is closer than a cutoff distance from an atom of the RNA/DNA molecule. However, the cutoff values vary widely between 3.5 Å and 6 Å. The most commonly used value is 3.5 Å and such close proximity typically involves a formation of a bond between protein and nucleic acid. Similarly, prior comparative reviews [10-14] also most often considered values of 3.5 Å and 5.0 Å. Consequently, we define the binding residues based on the 3.5 Å cutoff.

The existing predictive models can be divided into two types: 'sequence-only' models that perform predictions using solely the sequence and sequence-derived one-dimensional descriptors [41], such as secondary structure and solvent accessibility; and 'template-based' models that rely on a library of structural templates. The latter group of methods uses the input sequence to find a structure in complex with DNA or RNA that has similar sequence, and they use this structure to perform predictions. The two 'template-based' approaches, DBD-Threader [42] for the prediction of DNA-binding residues and SPOT-Seq [43] for the RNA-binding residues provide accurate predictions but they also require relatively long runtime; our tests using their webservers show runtime values up to several hours per protein for DBD-Threader and 20 min to a few hours for SPOT-Seq. Interestingly, SPOT-Seq was shown to discriminate between RNA- and DNA-binding proteins [43]. In the next chapter we investigate whether this could be also accomplished with the sequence-only models.

Table 2.2. Summary of predictors of DNA- and RNA- binding residues. Methods used in our empirical assessment are shown in bold.

	Method	Year	Ref[1]	Cut-off	Considered types of input features[2]											Prediction model[3]	Webserver[4]		
					AC	PP	PA	PS	SA	PSSM	Max Hom	Wild Span	StL	SeL	WS		URL	Output bin	Output pr
Predictors of DNA binding residues	DBS-pred	2004	[44]	3.5	✓					✓					3	NN	www.abren.net/dbs-pred/	✓	
	DBS-PSSM	2005	[45]	3.5						✓					5	NN	dbspssm.netasa.org	✓	✓
	BindN	2006	[46]	3.5		✓									11	SVM	bioinfo.ggc.org/bindn/	✓	✓
	Ho et al.	2007	[47]	3.5						✓					7	SVM	N/A		
	DP-Bind	2006,2007	[8, 48]	4.5						✓					7	SVM, KLR, PLR	lcg.rit.albany.edu/dp-bind	✓	✓
	DISIS	2007	[49]	6.0			✓	✓	✓		✓				3, 9	NN & SVM	cubic.bioc.columbia.edu/services/disis*		
	DNABindR	2006,2008	[38, 50]	1.0[5]	✓										21	Naïve Bayes	turing.cs.iastate.edu/PredDNA/index.html*		
	BindN-RF	2009	[51]	3.5		✓			✓	✓					11	RF	bioinfo.ggc.org/bindn-rf/*		
	DBindR	2009	[39]	3.5	✓			✓		✓					11	RF	www.cbi.seu.edu.cn/DBindR/DBindR.htm*		
	DBD-Threader	2009	[42]	4.5									✓		N/A	Template-based	cssb.biology.gatech.edu/skolnick/webservice/DBD-Threader/index.html	✓	
	ProteDNA	2009	[52]	4.5[6]						✓				✓	11	SVM	protedna.csie.ntu.edu.tw/method.php	✓	
	BindN+	2010	[53]	3.5		✓			✓	✓					11	SVM	bioinfo.ggc.org/bindn+/	✓	✓
	NAPS	2010	[54]	4.5	✓	✓				✓					7	C4.5	proteomics.bioengr.uic.edu/NAPS/*		
	DNABR	2012	[55]	3.5	✓	✓				✓					9	RF	www.cbi.seu.edu.cn/DNABR/*	✓	✓
Predictors of RNA binding residues	Jeong et al.	2004	[56]	6.0	✓			✓							41	NN	N/A		
	Jeong et al.	2006	[57]	6.0						✓					15	NN	N/A		
	BindN	2006	[46]	3.5		✓									11	SVM	bioinfo.ggc.org/bindn/	✓	✓
	PRINTR	2008	[58]	ENTANGLE				✓		✓					15	SVM	210.42.106.80/printr/*		
	RISP	2008	[59]	3.5						✓					7	SVM	grc.seu.edu.cn/RISP*		
	Pprint	2008	[60]	6.0						✓					11,13,15	SVM	www.imtech.res.in/raghava/pprint/	✓	✓
	RNAProB	2008	[40]	6, 5,3.5						✓					25	SVM	N/A		
	BindN+	2010	[53]	3.5		✓			✓	✓					11	SVM	bioinfo.ggc.org/bindn+/	✓	✓
	PiRaNhA	2009,2010	[61, 62]	3.9		✓	✓			✓					23	SVM	www.bioinformatics.sussex.ac.uk/PIRANHA*		
	NAPS	2010	[54]	4.5	✓	✓				✓					7	C4.5	proteomics.bioengr.uic.edu/NAPS*		
	ProteRNA	2010	[63]	5.0				✓		✓		✓			23	SVM	N/A		
	RBRpred	2010	[64]	6.0	✓		✓	✓		✓					15	SVM	N/A		
	Wang et al	2011	[65]	6.0		✓	✓			✓					15	SVM	N/A		
	PRBR	2011	[66]	3.5		✓		✓		✓					11	RF	www.cbi.seu.edu.cn/PRBR/*	✓	✓
	SPOT-Seq	2011	[43]	4.5									✓		N/A	Template-based	sparks.informatics.iupui.edu	✓	
	RNABindR	2006,2007,2012	[9, 13, 67]	5.0						✓					25	SVM	einstein.cs.iastate.edu/RNABindR/	✓	✓

1 Ref – reference

2 AC – amino acid composition; PP – physiochemical properties of amino acids; PA – predicted solvent accessibility (ASA); PS – predicted secondary structure; SA – sequence alignment; PSSM – position-specific scoring matrix; MaxHom – MaxHom algorithm [68]; WildSpan – WildSpan algorithm[69]; StL –template library of structures; SeL –template library of sequences; WS – window size;

3 NN – neural network; SVM – support vector machine; KLR – kernel logistic regression; PLR – penalized logistic regression; RF – random forest; C4.5 – decision tree

4 bin – outputs binary prediction; pr – outputs numeric propensity score;

5 An amino acid is a DNA-binding residue if its ASA computed in the protein-DNA complex using NACCESS is smaller than its ASA in the unbounded protein by at least 1Å

6 A residue is regarded as involved in sequence-specific binding with the DNA if one or more heavy atoms in its side chain fall within 4.5Å from the nucleobases of the DNA

* Webserver was not available as of December 2013 when the predictions were collected.

The predictive strategy used by the 'sequence-only' methods consists of two steps. First, each residue in the input protein sequence is encoded into a vector of numerical features. Next, these features are used as inputs to a predictive model that outputs a binary value (binding versus nonbinding) and, for some methods, also a numeric score that quantifies propensity for the binding (Table 2.2). The information used to compute features for a given residue is collected from a window of residues that are adjacent to this residue in the sequence. The sizes of this window vary widely between methods, ranging from 3 (one residue on each side of the predicted residue) to 41; the most frequently used value is 11 (Table 2.2). The sequence-only predictors use a variety of designs that vary both on the information that is used to generate the features and the predictive models used. The input features include information derived directly from the protein sequence including amino acid composition (counts of specific types of amino acids), and physiochemical properties of the input amino acids, such as pKa value of side chains, hydrophobicity, molecular mass and charge. Some features are also computed from one-dimensional structural characteristics that are predicted from the sequence, such as secondary structure and solvent accessibility. The most common input is based on the results of multiple sequence alignment of the input chain into a large sets of protein sequences (such as the *nr* database), primarily in the form of the evolutionary profile quantified with the position-specific scoring matrix (PSSM). This is related to the fact that PSSM can be used to quantify conservation of residues and the binding residues were shown to be conserved in the sequence [68, 70, 71]. Two predictors substitute PSSM with another way to find conserved residues. ProteRNA method [63] uses the WildSpan algorithm [69], while DISIS [49] uses MaxHom [68] algorithm. The predictive models are exclusively implemented based on a variety of ML algorithms (described in detail in section 2.4) including neural networks, SVMs, Naïve Bayes and decision trees. The SVM is used most often, which is motivated by empirical results that demonstrate that this type of model usually provides strong predictive performance [12, 46]. However, we note that different methods were trained and tested on different datasets, which vary in terms of their release date, size, resolution of structures used to generate annotation of binding, sequence similarity within the dataset and definition of binding annotation. Moreover, they were evaluated using different protocols (e.g. using test datasets and a variety of

cross-validation types) and the predictive performance was assessed using different measures. Therefore, we could not use the results reported in the original articles to directly compare predictive quality of these methods. Our tests of methods that offer webserver indicate that DBS-pred [44] and BindN [46] are among the fastest methods that complete the prediction of DNA-, RNA-binding residues for an average-sized protein with about 200 amino acids in <1 s.

Recent studies also investigated development of consensus approaches, which combine multiple predictions into one prediction. The premise is that if individual predictions are different from each other and they complement each other, then combining them together would lead to an improved predicting performance when compared to each of the individual methods. Si [14] have implemented a consensus method MetaDBSite that integrates predictions from six DNA-binding predictors: DBS-pred [44], BindN [46], DP-Bind [48], DISIS [49], DNABindR [38] and BindN-RF [51]. The results of these predictors are combined using the SVM model, and the resulting consensus was shown to outperform each individual predictor. Similarly, Puton [12] assessed predictive quality of seven sequence-based methods for prediction of RNA-binding residues and developed a consensus that combines predictions from the top three predictors: PiRaNhA [62], Pprint [60] and BindN+ [53]. The outputs of these methods were merged together using weighted average where the weights correspond to the predictive quality of these methods on a benchmark dataset. Again, their empirical results show that their consensus outperforms the results generated by each of the three single predictors. These results motivated us to further investigate feasibility of building accurate consensus-based approaches.

2.4 Computational background

Generally speaking, computational prediction of the DNA- or RNA- binding residues usually considers prediction as a binary classification problem, i.e., each amino acid in the input protein sequence is classified as either binding or nonbinding. This involves two steps: construction of a classification model and use of this model to perform predictions. First, a set of labelled amino acids (binding vs. non-binding) is used by a machine

learning algorithms to generate a classification model by exploiting the underlining data patterns (amino acid properties that distinguish binding from non-binding). Next, the model is used to make classifications/predictions on new data (amino acid) whose class label information is unknown. Some models also produce a probability/confidence that is associated with the binary prediction or which is used to generate the binary prediction. In the latter case the residues with confidence > given threshold are assumed as binding and otherwise they are assumed as nonbinding.

2.4.1 Development of computational methods for the prediction of protein-DNA/RNA interactions

As shown in Figure 2.4, the development of a prediction model includes two main activities: training and testing. First, the experimental data (experimentally solved DNA/RNA-binding protein sequences and their corresponding binding annotations) are split into training and test sets to train and to test the prediction model, respectively. The data is split per protein, i.e., all residues from a given protein are placed into one of these two datasets. During the training process, the prediction model learns to distinguish between the already known binding residues (binding class) and nonbinding residues (nonbinding class) using a training dataset of proteins. Each amino acid in the training protein sequence is encoded with a set of numerical features. These numerical features are utilized as input that is fed into a machine learning algorithm to build the prediction model. Then the model is used to generate prediction output. Prediction accuracy is assessed by comparing the prediction output with the actual/true output, and this information is used to guide the learning process to maximize prediction accuracy. Often, this process involves finding a well-performing set of features and parameterization of machine learning algorithm. This is performed to maximize predictive performance on the training dataset, and to avoid overfitting a cross validation is applied. Overfitting means the model fits (memorizes) the training data very well but fails to generalize on new data. This may happen when the model is large and complex, and as a result it ends up describing noise or errors rather than the underlying data patterns in the training dataset. We use cross validation, which is discussed in section 2.4.3, to reduce a chance of overfitting the training dataset. During the testing, we apply our trained on the training

dataset predictive model to generate prediction for the test sequences. Next, the predicted outcomes are compared with the actual/true outcomes for these test proteins to measure predictive performance of the method.

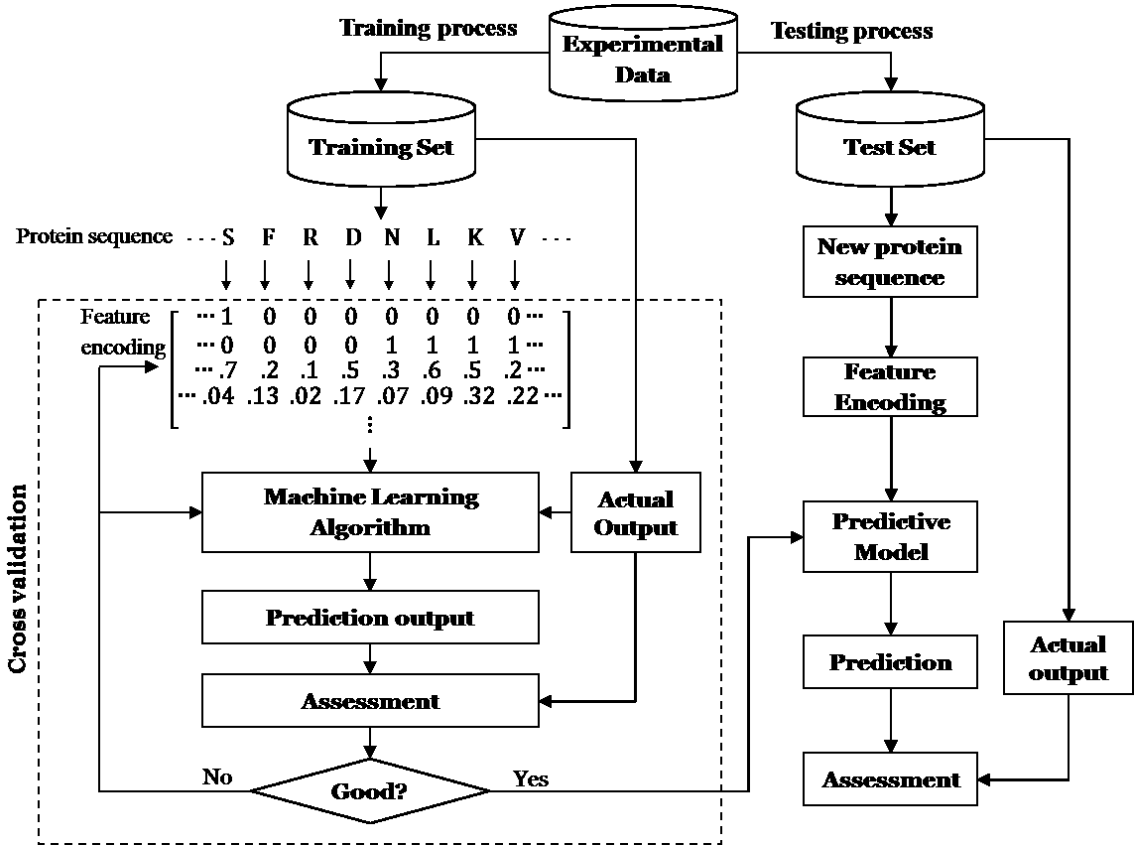


Figure 2.4. Flowchart of the process to develop and test the computational prediction methods.

A typical training process consists of the following three steps:

1. Feature generation

First, one needs to design a set of features that represent the input data. In this study, each input amino acid is encoded with numerical features that quantify structural and physicochemical properties of these amino acids that can be potentially used to discriminate between binding and non-binding amino acids. Based on the underlying feature type, these features can be grouped as binary (e.g. "exposed" or "buried" that denote the solvent accessibility state), categorical (e.g. "Helix", "Strand", or "Coil" that represent the secondary structure state) and real-valued (e.g. area of solvent accessible

surface). These features are derived either directly from the input protein sequence (e.g. the type of amino acid) or computed using outputs generated by other software that predicts structural or functional characteristics of the protein from the sequence (e.g. predicted secondary structure).

2. Feature selection

The number of the generated features could be large and some of them might be redundant or irrelevant to a given classification task. If not removed, they might harm predictive performance (model would describe the weak or noisy information from the irrelevant features) and will unnecessarily increase the runtime required to build the model. Feature selection aims to select a subset of relevant and non-redundant features. We consider two types of feature selections, depending on how feature selection search is combined with the construction of the prediction model: filter methods and wrapper methods. The filter methods treat the feature subset selection as independent of the model construction. They typically involve computation of a score for each feature that measures usefulness of that feature for the prediction. A commonly used filter technique uses correlation-based score. It estimates relevance of a given feature using its correlation with the outcome (class label). We use the point biserial correlation coefficient (PBC) which is suitable to quantify correlation of our features with our binary outcome: binding vs. non-binding class labels. Features with low values of the correlation (irrelevant features) are removed. Next, we remove the redundant features by using Pearson correlation coefficient (PCC) that quantifies correlation between features. We remove one of the features (the one with lower PBC) in each pairs of features that have high PCC value; this way we remove redundant features. The filter methods are usually computationally fast. The wrapper methods embed the model construction within the feature selection process. They explore various subsets of features that are generated using a search algorithm. Each considered subset of features is used to train and test the prediction model. The feature subset that returns the best prediction performance on the training set is selected. Wrapper methods are computational more expensive when compared to the filter methods, but they also tend to provide a feature set that secures better predictive performance on the training dataset.

3. Model training

Using the selected set of features, a machine learning algorithms learns a mapping of the values of these features into the corresponding output (class labels). Many machine learning algorithms can be used including Support Vector Machine (SVM), Logistic Regression, Neural Network, Naïve Bayes, and Random Forest, among others. In this study, we use logistic regression which is described in detail in the next section. This type of algorithm has been already successfully used in a related study [72]. Moreover, further motivation to use this particular type of algorithm is because it generates simple linear models, which are less likely to overfit training dataset compared to more complex models that use a larger number of parameters. This algorithm is also fast to generate the model on the training dataset, which is an important advantage given a relatively large size of our dataset. Finally, models generated by regression are also fast to make prediction on new data, which is crucial given our goal to developing a runtime efficient (high throughput) predictor. Most of the machine learning algorithms require the user to determine values of certain control parameters. These parameters can be used to optimize the predictive performance and generalize the resulting model, so that it performs well on test dataset. Determining the best set of values of these parameters, called parameterization of the learning algorithm, is important. A popular method to select parameters is grid search. This is an exhaustive search through a manually specified grid of parameters. One train the algorithm with each set of considered parameter values from the grid using a given training dataset and evaluates the corresponding predictive performance. The set of parameter values which returns the best prediction performance is selected. The easy of this parameterization is another advantage of the logistic regression as this algorithm uses only one parameter: the ridge λ (detail information see in section 2.4.2). Most other algorithms use multiple parameters, which results in much larger computational cost of the parameterization. With the regression, we first define a set of considered ridge values $\lambda \in \{10^{-10}, 10^{-9}, \ldots, 10^{5}\}$. After using the grid search we compute predictive performance for each ridge value on the training dataset, we select ridge value with the best predictive performance.

After the entire training process is completed, the designed predictive model (using selected features and parameters of the algorithm) is generated using the training dataset and applied to make predictions on new unseen test data.

2.4.2 Logistic regression

Logistic regression is a type of regression that predicts the probability of occurrence of an event by fitting data to a logistic function. Suppose that we have *m* observations (in our study these are amino acids) represented by (X, Y), where $Y = [Y_1, Y_2, \dots, Y_m]$ and $X = [X_1, X_2, \dots, X_m]$. For each observation (amino acid), $Y_i \in \{0,1\}$ is the binary outcome representing the binding vs. non-binding state of this amino acid, and X_i is an n-dimensional vector of input features. Logistic regression finds a linear fit of X_i:

$$f(X_i) = \theta_0 + \theta_1 X_i^1 + \cdots + \theta_n X_i^n = \sum_{k=0}^{n} \theta_k X_i^k = \theta^T X_i \,,$$

where $\theta = [\theta_0, \theta_1, \dots, \theta_n]$ is an n-dimensional vector of coefficients that need to be calculated, $f(X_i)$ is the decision boundary according to which input data X_i can be assigned to one of the two outcomes:

$$Y_i = 1, if\ f(X_i) > 0$$

$$Y_i = 0, otherwise$$

Then we apply logistic function on $f(X_i)$:

$$h(X_i) = g(f(X_i)) = \frac{1}{1+e^{-\theta^T X_i}} \,,$$

so that the function value $h(X_i)$ is bound to unit interval [0, 1], and could be interpreted as the probability of $Y_i = 1$ knowing X_i, that is:

$$P(Y_i = 1 | X_i; \theta) = h(X_i)$$

$$P(Y_i = 0 | X_i; \theta) = 1 - h(X_i)$$

Combining them together, the probability function can be written as:

$$P(Y_i | X_i; \theta) = (h(X_i))^{Y_i} (1 - h(X_i))^{1-Y_i}$$

Finding the optimal θ is usually done using the maximum likelihood estimation with the input data points (training dataset). The likelihood function is then

$$L(\theta) = \prod_{i=1}^{m} P(Y_i|X_i;\theta) = \prod_{i=1}^{m} (h(X_i))^{Y_i}(1-h(X_i))^{1-Y_i}$$

Log-likelihood function turns products into sums:

$$l(\theta) = logL(\theta) = \sum_{i=1}^{m} Y_i \log h(X_i) + (1-Y_i)\log(1-h(X_i))$$

Log-likelihood function with penalty is defined as

$$l^{\lambda}(\theta) = l(\theta) - \lambda\|\theta\|^2$$

where the ridge parameter λ controls the amount of shrinkage of the norm of θ. The quality of the fit of ridge logistic regression depends on the selection of proper ridge parameter λ.

2.4.3 Cross validation

Cross validation is a method that is used to evaluate predictive model. It aims to mimic testing on the test dataset (use of data were not used in model building to test a given model) using the training dataset. This method is used to estimate predictive performance that is in turn used during the training process to guide the feature selection and machine learning algorithm parameterization. There are several types of cross validation: holdout method (2-fold cross validation), *k*-fold cross validation where $2 < k < m$ and leave-one-out cross validation where $k = m$. In the *k*-fold cross validation, the whole training dataset is randomly split into *k* equally sized subsets (also called folds). One fold is used as a test set, and all the remaining *k*-1 folds are combined together to form a training set. The training set is used to train a model, while the test set is utilized to perform the evaluation of this model. This is repeated *k* times, and each time we choose a different fold as the test set. Eventually, we average the prediction performance over all the *k* test folds to produce a single result which gives an estimate of how well this model performs over the entire training dataset.

2.4.4 Evaluation criteria

Evaluation of predictive quality is performed for the two types of predictions: binary prediction (binding vs. nonbinding) and the real-valued scores that quantify the propensity that a given residue binds a given type of nucleic acid. The binary predictions can be assessed using the following four measures:

$$\text{Accuracy} = \frac{TP+TN}{TP+FN+FP+TN}$$

$$\text{Sensitivity} = \frac{TP}{TP+FN}$$

$$\text{Specificity} = \frac{TN}{FP+TN}$$

$$\text{Matthews correlation coefficient (MCC)} = \frac{TP \times TN - FN \times FP}{\sqrt{(TP+FN) \times (TP+FP) \times (TN+FP) \times (TN+FN)}}$$

where TP is the number of true positives (correctly predicted binding residues), FN is the number of false negatives (incorrectly predicted binding residues), FP is the number of false positives (incorrectly predicted nonbinding residues) and TN is the number of true negatives (correctly predicted nonbinding residues). These four measures were used in similar works that addressed prediction of DNA or RNA binding residues [10-14, 38, 53, 59].

The predicted propensities are evaluated using receiver operating curve (ROC), which is a plot of false-positive rate (FPR), which equals 1-specificity, against the true-positive rate (TPR), which is the same as sensitivity. These two rates are computed by binarizing the propensities using thresholds (we use all unique values of the predicted propensities as thresholds). We report the area under the ROC curve (AUC) and the same value was reported in other studies [12, 13, 53, 55]. In our training dataset the fraction of the DNA- (RNA-) binding residues is 8.2% (4.8%), i.e., majority of the residues are nonbinding. Thus, even a small FPR = 0.2 corresponds to the prediction where the binding residues are over-predicted 2.5 (4) times compared to their actual number. Therefore, we focus our assessment of the predictive performance on the part of the ROC where number of FPs is no bigger than the number of actual positives (native binding residues). This corresponds

to the lower (left side) part of ROC where FPR≤8.2% for DNA and ≤4.8% for RNA. Consequently, we also report the area under this lower part of curve (AULC).

Since DNA-binding and RNA-binding residues share similar biochemical properties, it is likely that the existing methods cross-predict between these two types of nucleic acid binding residues, i.e., method for prediction of RNA-binding residues also predict DNA-binding residues and vice versa. We introduce a new measure, called Ratio, to quantify the amount of cross-prediction between DNA- and RNA-binding residues. Ratio is defined as the fraction of native DNA-binding residues that are predicted as RNA-binding, and the fraction of native RNA-binding residues that are predicted as DNA-binding. Moreover, we introduce a Ratio curve, which is a plot of Ratio against the TPR. These two values are calculated by binarizing the propensities using thresholds (we use all unique values of the predicted propensities as thresholds). We report the area under the entire Ratio curve (AURC). Given the low numbers of binding residues, we also quantify the area under the lower part of the curve where TPR≤0.5 (AULRC). The larger values of TPR are less interesting because such prediction would generate very high FPR.

2.4.5 Statistical test

We used statistical tests to assess whether differences in two sets of numeric results are significant. The null hypothesis was defined as the median or mean of the two sets of results being equal. The difference is identified as statistically significant if the null hypothesis is unlikely to be observed at a given level of significance. In this study, we apply statistical tests to evaluate the significance of differences in predictive quality between the best-performing prediction method and each of the other considered methods. First, we estimate the predictive quality (e.g. MCC) on a representative and sufficiently large sample of the input data (proteins). We cannot compare the results using individual proteins since a single protein does not offer a sufficient amount of information to calculate the predictive quality. For instance, some proteins do not have positives (binding residues) and some will have very few positives. This is the case in the prediction of DNA-binding residues, where all residues in the RNA-binding proteins do not bind DNA. Therefore, instead of comparing results on individual proteins, we compare pairs of methods on several randomly selected datasets of proteins. More

specifically, we randomly select 70% of proteins in a given test dataset and measure the predictive quality for the considered methods. The choice of 70% allows us to have sufficiently large dataset to obtain high-quality estimates of predictive performance while still generating substantially diverse subpopulations of the complete test dataset. We repeat this process 10 times and collect the 10 results for each method. Next, we compare the 10 pairs of results. If the results follow normal distribution, as tested using the Anderson–Darling test [73] with 0.05 significance, we use the paired *t*-test to investigate significance; otherwise we use the Wilcoxon rank sum test (described in detail in the following sections). The difference between a given pair of predictors is assumed statistically significant if p-value <0.05.

Student's t-test (paired)

A paired student's t-test is used to compare two groups of data (values of predictive performance) in which individual data points in one group are paired with data points in the other group (predictive performance for two methods in the same randomly chosen dataset), and determine whether their means are significantly different from each other. Student's t-test is commonly applied when data in each group follow a normal distribution.

Suppose we have two groups of data with size n (n = 10 in our work),

$$X_1 = \begin{bmatrix} x_{1,1} \\ x_{1,2} \\ \vdots \\ x_{1,n} \end{bmatrix}, \qquad X_2 = \begin{bmatrix} x_{2,1} \\ x_{2,2} \\ \vdots \\ x_{2,n} \end{bmatrix}$$

where $x_{1,i}$ and $x_{2,i}$ are the matched pairs for $i \in [1, n]$. For example, $x_{1,i}$ is the value of MCC of one method on a given randomly chosen dataset, and $x_{2,i}$ is the value of MCC of the other method on the same dataset and $i = 1,2, \ldots ,10$ is the index of the dataset. Paired t-test is carried out by first setting the null hypothesis:

H_0: the means of the two groups are equal.

Then the t value is calculated:

$$t = \frac{\bar{X}_D}{\frac{s_D}{\sqrt{n}}},$$

$$d.f. = n - 1,$$

where $\bar{X}_D = \sum_{i=1}^{n}(x_{1,i} - x_{2,i})/n$ is the average difference of all pairs of data, s_D is the standard deviation of those differences, and $d.f.$ is the degree of freedom.

Once the *t* value is calculated, a *p*-value can be determined from a table of values from the student's t-distribution using the corresponding $d.f.$ value. P-value is the probability of obtaining a result equal to or more extreme than what was actually observed, assuming the null hypothesis is true. A low *p*-value (typically <0.05) suggests that the data provides enough evidence to reject the null hypothesis and to accept an alternative hypothesis that is there is a significant difference between the means of the two groups of data.

Wilcoxon rank test

The Wilcoxon rank-sum test is a nonparametric alternative to the paired t-test in the case when the data is not normally distributed. It is used to determine the difference between the medians of two groups of data (values of predictive performance). Wilcoxon test is used to test the null hypothesis:

H_0: the median of the two groups are equal.

This test is carried out in the following steps:

a) Calculate the difference for each pair of data, $d_i = x_{1,i} - x_{2,i}, i \in [1, n]$.
b) Exclude the pairs where d_i=0, which leaves n_r difference.
c) Rank the absolute difference, $|d_i|$, from smallest to largest for the remaining pairs, assign rank 1 to the smallest $|d_i|$, rank 2 to the next, etc.
d) Add sign to each rank according to the sign of d_i. Assign a "+" sign when d_i>0, and a "-"sign when d_i<0.
e) Calculate W, which is the sum of the signed ranks. It includes: W_+, the sum of the ranks of the positive d_i, and W_-, the sum of the ranks of the negative d_i.

f) Choose $W = \min(W_+, W_-)$

g) Find the *p*-value for the computed W from Wilcoxon reference table.

If the *p*-value is smaller than a threshold (typically 0.05), then we reject the null hypothesis H_0 and support an alternative hypothesis that the median between these two groups of data differ significantly from each other.

Anderson-Darling test

The selection of the student's t-test or Wilcoxon rank test is based on the normality of the underlying data, which we determine with the Anderson-Darling test [73]. This test checks whether a given group of data came from a specific distribution (e.g. normal, lognormal, exponential, etc.). In this study, we consider the normal distribution. Suppose we have a group of data *X* with size *n*, the null hypothesis of the Anderson-Darling test is

H_0: The data follows a normal distribution.

To compute the test, we first sort all data in X from smallest to largest, $x_1 < x_2 < \cdots < x_n$, and then calculate

$$A^2 = -n - S,$$

where

$$S = \sum_{i=1}^{n} \frac{(2i-1)}{n} [\ln(F(x_i)) + \ln(1 - F(x_{n+1-i}))],$$

and *F* is the cumulative distribution function of the normal distribution. Then the test statistic *A* is compared against critical values from the table for normal distribution. If the *p*-value found in the table is bigger than a threshold (typically 0.05), then we accept the null hypothesis that the data are from a normal distribution, otherwise we reject the null hypothesis.

Chapter 3

Goal 1: Assessment of predictive performance of existing sequence-based DNA- and RNA- binding residue predictors

Many methods predicting the protein–DNA/–RNA interactions from the protein sequence and structure have been published and reviewed in the literature over the past several years [11, 74-76] Table 3.1 summarizes recent comparative reviews of the predictors of DNA-binding residues [10, 14] and RNA-binding residues [11-13]. These comparative analyses provide useful clues about the predictive performance of various predictors and help the end users to select a suitable method from among many available choices. However, these reviews and the corresponding predictive models focus solely on the prediction of interactions with just one of the two nucleic acids types. They do not consider how well they separate between DNA and RNA interactions, which is an important oversight. Another drawback of the prior comparative reviews is that they consider datasets with incomplete annotations of binding residues. This is because the annotations are based on a single structure of protein–DNA or protein–RNA complex, which could be partial if only a fragment of DNA or RNA is considered in a given complex or if the same protein is involved in other binding events.

Table 3.1. Summary and comparison of recent reviews concerning prediction of DNA- and RNA- binding residues from protein sequences.

Review article (year published)	Scope of descriptive component					Scope of empirical component						
	Coverage	# methods	Year published of newest method	Defines binding	Discusses outputs of methods	Benchmark dataset used				Considers cross prediction (RNA-binding on DNA-binding proteins and vice versa)	Evaluates consensus methods	Considers combined DNA and RNA binding prediction
						Year collected	# proteins	Cutoff(s) to define binding	Complete annotations of binding			
This study	DNA and RNA	30 (14+16)	2012	yes	yes	2013	531	3.5; 5	yes	yes	yes	yes
[11] (2013)	RNA	10	2011	no	no	2012	106	undefined	no	no	no	no
[10] (2013)	DNA	11	2011	no	no	2012	301	3.5	no	no	no	no
[13] (2012)	RNA	13	2011	yes	no	2010	198	5	no	no	no	no
[12] (2012)	RNA	7	2011	no	no	2011	44	3.5	no	no	yes	no
[14] (2011)	DNA	6	2009	yes	no	2010	232	3.5; 4; 4.5; 5; 5.5; 6	no	no	yes	no

We perform empirical assessment of methods that offer webservers using a new benchmark dataset characterized by a more complete annotation that includes binding residues transferred from the same or similar proteins. We also investigate the ability of these methods to discriminate DNA-binding residues from RNA-binding residues. The complete review was published in [77], while here we discuss a selection of arguably the most interesting findings.

3.1 Benchmark datasets

Similar to other studies [10-14], our benchmark datasets were extracted from structures of protein–DNA and protein–RNA complexes collected from PDB [6]; these data were obtained in September 2013. The definitions of the binding residues differ between prior studies in this area, with the most prevalent approach based on the cutoff distance [44]. Table 3.1 ['cutoff(s) to define binding' column] and Table 2.2 ('cutoff' column) reveal that 29 of 30 predictors of binding residues use this definition, although the cutoff values used vary considerably. We apply 3.5 Å to define binding, since this cutoff value is used most often when designing the prediction methods (13 of 30 methods in Table 2.2). We collected total of 1082 high-quality X-ray structures (resolution better than 2.5 Å) of protein–DNA complexes, 271 protein–RNA complexes and 4 complexes that include both DNA and RNA. These complexes are split into chains and the chains that have no binding residues or are shorter than 30 amino acids in length are removed. As a result, we obtained 1935 DNA-binding chains and 981 RNA-binding chains for distance cutoff of 3.5 Å.

Motivated by a recent work that evaluated predictive quality of methods that find small ligand binding pockets on the protein surface [78], we improve the annotations of binding residues by transferring these annotations between similar proteins. This similarity stems from the fact that the structures of protein–DNA and protein–RNA complexes could concern paralogs, similar or the same proteins in different organisms, and structures of the same proteins solved at different resolutions or with different co-factors. Using the procedure introduced in [78], we first find proteins that are similar in their structure and sequence. The structural similarity is expressed with the template

modeling (TM) score [79]. The similarity in the sequence is measured with the sequence identity expressed as a fraction of aligned residues over the length of the shorter sequence, where the alignment is calculated using bl2seq [80] with default parameters; we only consider the aligned proteins for which *e*-value <0.001. The two similarity scores are used to perform clustering of protein chains where two chains are assigned into the same cluster if their TM score >0.5 and the sequence identity >80% [78]. The chains in the same cluster are assumed to be sufficiently similar and are represented by one chain with the largest number of binding residues. The annotations of binding residues of the remaining chains in the same cluster are transferred into this chain. This is done based on the alignment with bl2seq (e-value <0.001) where annotations are transferred for positions that are matched in the alignment. As a result of the transfer, the numbers of annotated DNA-binding residues and RNA-binding residues were enlarged by 13.7 and 9.7%, respectively.

The original redundant datasets were reduced after the clustering to the non-redundant (proteins are different from each other in both sequence and structure) dataset of 356 DNA-binding proteins, and 175 RNA-binding proteins. The non-redundancy is important since this way we avoid a bias towards certain overrepresented types of proteins. We split them into training and test proteins based on their release date. We observe that the datasets used by the considered predictors of DNA- and RNA-binding residues were collected before September 2010. Correspondingly, the binding proteins released before September 2010 constitute the training set, which we use to select and compute consensuses. The proteins released after September 2010 are less likely to be used to train the published methods. Furthermore, we reduce this set of proteins by excluding those that are similar to the training proteins. Using CD-HIT [81], we clustered all training and test protein chains sharing ≥30% sequence similarity and we removed all test proteins that are in the same cluster with any of the training chains. The remaining test proteins share <30% sequence similarity with training proteins; this assures that these test proteins are sufficiently different from proteins used to build predictive models to perform unbiased tests. Consequently, our training dataset contains 293 DNA-binding proteins and 149 RNA-binding proteins. The test proteins were used to establish the following datasets: ‘DNA_T’ test dataset that includes 47 DNA-binding proteins,

'RNA_T' test dataset that contains 17 RNA-binding proteins, and the combined 'COMB_T' test dataset that has 64 nucleic acid-binding proteins; 'T' denotes the fact that the annotations were transferred.

3.2 Selection of methods included in the empirical assessment

The empirical assessment includes sequence-based methods for the prediction of DNA- and RNA-binding residues that were selected from the comprehensive list of 30 methods shown in Table 2.2. We selected nine predictors that were available as webservers as of December 2013 when the predictions were collected and which are runtime-efficient, i.e. they predict an average size protein sequence with 200 residues in under 10 min; this assures that we cover methods that are convenient to use for the end users. We use the most recent versions of methods that have multiple versions. We include four predictors of DNA-binding residues, DBS-PSSM [45], DP-Bind [8, 48], ProteDNA [52] and BindN+ [53], and three for the predictions of RNA-binding residues, Pprint [60], BindN+ [53] and RNABindR [9, 13, 67]. DP-Bind implements a family of methods that includes three ML models, support vector machine (SVM), kernel logistic regression (KLR) and penalized logistic regression, and two types of consensuses of these models [48]. We consider the default KLR classifier-based model, DP-Bind(klr), and the default majority-vote-based consensus, DP-Bind(maj). ProteDNA offers predictions in two modes, one with high-precision and another balanced; we use the latter version, ProteDNA(B), that provides a better balance between sensitivity and specificity [52]. We also consider two recent consensus-based approaches, which combine predictions of multiple methods: MetaDBSite [14] for the DNA-binding and the consensus by Puton et al. [12] for the RNA-binding. In total we examine 10 predictors, including three consensus-based approaches, which cover a comprehensive range of designs. These methods include a variety of predictive algorithms (Table 2.2), such as neural networks, SVMs, regression, Bayesian classifiers and consensuses, and they make use of several different types of inputs, such as evolutionary profiles, sequence alignment, composition of amino acids and physiochemical properties of amino acids.

From the list of recent methods we exclude DBS-pred [44] and BindN [46], which were superseded by DBS-PSSM and BindN+, respectively; DBD-Threader [42] and SPOT-Seq [43] that rely on libraries of structures of protein–DNA and protein–RNA complex and took excessive amount of time to run; and several methods that do not offer a webserver including the predictor by Ho [47], by Jeong [56, 57], RNAProB [40], ProteRNA [63], RBRpred [64], and method by Wang [65]. We also could not consider DISIS [49], DNABindR [38, 50], BindN-RF [51], DBindR [39], NAPS [54], DNABR [55], PRINTR [58], RISP [59], PiRaNhA [61, 62] and PRBR [66] because their webservers were either no longer maintained or unavailable at the time of our experiment.

3.3 Results and discussion

We perform empirical assessment of the 10 selected computationally efficient sequence-only predictors that are available as webservers on the test datasets: DNA_T (with DNA-binding proteins only), RNA_T (with RNA-binding proteins only) and COMB_T (with DNA- and RNA-binding proteins). These datasets include binding annotations that were transferred between similar proteins, which results in a more complete set of annotations when compared to prior comparative studies.

3.3.1 Predictive performance on the datasets with DNA-binding or RNA-binding proteins

Table 3.2 reveals that predictive performance of the individual predictors of DNA-binding residues [DBS-PSSM, DP-Bind(maj), DP-Bind(klr) and BindN+] on the DNA_T dataset is relatively similar, with MCC values ranging between 0.293 and 0.307, and AUC ranging between 0.795 and 0.797. The only exception is the ProteDNA method that is characterized by lower predictive quality on this test dataset. A likely explanation is the fact that this method was designed to find binding residues specifically in the transcription factors, which are a subset of our dataset that also includes other types of the DNA-binding proteins protein. This is corroborated by the relatively low value of sensitivity that was obtained by this predictor. Interestingly, the MetaDBSite consensus is

also underperforming when compared with the results reported by the authors [14]. The reason is that four methods that this consensus was originally designed to use are no longer maintained. Consequently, instead of combining results of six predictors the current version of MetaDBSite is a simple ensemble of BindN and DP-Bind based on the logical AND, i.e., a given residue is predicted as DNA-binding if both methods predict it as such. Analysis of the results concerning prediction of the RNA-binding residues leads to similar observations (Table 3.2). Predictive performance of the three considered predictors (BindN+, RNABindR and Pprint) vary between 0.141 and 0.219 in MCC, and between 0.681 and 0.738 in AUC on the RNA_T test dataset. The Meta2 consensus is not performing as well as previously reported [12]. This is because some of the methods Meta2 was originally designed to combine are no longer available.

Table 3.2. Results of empirical assessment of predictors of the DNA- or RNA-binding residues on the DNA_T or RNA_T datasets, respectively.

Significance of the difference in MCC and AUC values between the best performing method and other methods on a given dataset was assessed based on 10 tests that utilize 70% of randomly chosen proteins; + (=) in the Sig column denotes that the difference was (was not) significant at p-value <0.05. AUC values could not be computed for DP-Bind(maj), MetaDBSite, ProteDNA(B), Meta2, and the four new consensuses since these methods provide only the binary predictions. Methods are sorted by the MCC value.

Datasets	Methods	Sensitivity	Specificity	MCC	Sig	AUC	Sig
DNA binding on DNA_T dataset	Machine learning consensus	0.478	0.916	0.354		0.831	
	Majority vote (BindN+(DNA), DBS_PSSM, ProteDNA(B))	0.447	0.907	0.314	+		
	DBS-PSSM	0.721	0.753	0.307	+	0.796	+
	Logic consensus (BindN+ AND DBS-PSSM)	0.424	0.912	0.305	+		
	DP-Bind(maj)	0.598	0.823	0.301	+		
	DP-Bind(klr)	0.590	0.824	0.297	+	0.795	+
	BindN+	0.482	0.879	0.293	+	0.797	+
	MetaDBSite consensus	0.325	0.935	0.267	+		
	ProteDNA(B)	0.093	0.982	0.142	+		
RNA binding on RNA_T dataset	Machine learning consensus	0.242	0.962	0.234		0.755	
	BindN+	0.399	0.891	0.219	=	0.738	+
	Majority vote (BindN+(RNA), RNAbindR, Pprint)	0.457	0.854	0.212	+		
	Meta2 consensus	0.526	0.812	0.211	+		
	Logic consensus (BindN+ AND RNABindR AND Pprint)	0.244	0.950	0.203	+		
	RNABindR	0.575	0.739	0.178	+	0.724	+
	Pprint	0.433	0.796	0.141	+	0.681	+

Overall, we conclude that methods for the prediction of DNA-binding (RNA-binding) residues are characterized by relatively good predictive performance measured by their values of MCC and AUC when tested on the dissimilar (in the sequence) proteins that bind DNA (RNA). Their AUC is at about 0.8 (0.7), and their predictions have modest correlation with the native annotations at about 0.3 (0.2). They offer relatively high specificity coupled with modest sensitivity, which means that they predict a subset of native binding residues with high predictive quality while missing the remaining binding residues.

3.3.2 Predictive performance on the dataset with DNA- and RNA-binding proteins

We are the first to comprehensively assess predictive performance of the considered predictors on the COMB_T dataset that combines DNA- and RNA-binding proteins, see Table 3.3. We observe a drop in MCC when compared with the results in Table 3.2. This is a universal pattern, irrespective of whether we assess predictors of DNA- or RNA-binding residues, and it reveals that these methods confuse the two types of binding residues. Sensitivity stays the same, as the annotation of the binding residues does not change compared with when we consider prediction of DNA- or RNA-binding residues; we just introduce additional nonbinding residues.

Considering individual predictors of the DNA-binding residues, the MCC on the COMB_T dataset (Table 3.3) is lower by 3.8–5.5% when compared with the results on the DNA_T dataset (Table 3.2). The only exception is ProteDNA, which has low sensitivity and MCC and which predicts a relatively small number of residues that selectively bind transcription factors. Ratio, which quantifies fraction of RNA-binding residues that are predicted to be DNA-binding, reveals that at least 28.9% and as many as 48.7% of the RNA-binding residues are mispredicted. Similarly, assessment of the predictors of the RNA-binding residues on the COMB_T dataset demonstrates that the results are worse when compared with the results on the RNA_T dataset. Specifically, MCC is lower by 5.7–10.5%; AUC by 1.2–3.1%; and specificity by 1.4–3.7%. Most importantly, the identical sensitivity coupled with the lower specificity indicates that predictors of RNA-binding residues mispredict the DNA-binding residues as RNA-

binding, which is further confirmed by the large values of Ratio. Ratio tells that these methods mispredict between 47.8 and 64.3% DNA-binding residues as RNA-binding.

Table 3.3. Results of empirical assessment of predictors of the DNA- or RNA-binding residues on the COMB_T dataset.

Significance of the difference in MCC and AUC values between the best performing method and other methods on a given dataset was assessed based on 10 tests that utilize 70% of randomly chosen proteins; + (=) in the Sig column denotes that the difference was (was not) significant at p-value <0.05. AUC values could not be computed for DP-Bind(maj), MetaDBSite, ProteDNA(B), Meta2, and the four new consensuses since these methods provide only the binary predictions. Methods are sorted by the MCC value.

Binding type	Methods	Sensitivity	Specificity	Ratio	MCC	Sig	AUC	Sig
DNA binding	Machine learning consensus	0.478	0.922	0.267	0.311		0.841	
	Majority vote (BindN+(DNA), DBS_PSSM, ProteDNA(B))	0.447	0.916	0.232	0.277	+		
	Logic consensus (BindN+ AND DBS-PSSM)	0.424	0.919	0.232	0.267	+		
	DBS-PSSM	0.721	0.774	0.487	0.266	+	0.810	+
	BindN+	0.482	0.888	0.289	0.256	+	0.806	+
	DP-Bind(maj)	0.598	0.823	0.467	0.247	+		
	DP-Bind(klr)	0.590	0.828	0.445	0.246	+	0.794	+
	MetaDBSite consensus	0.325	0.933	0.230	0.221	+		
	ProteDNA(B)	0.093	0.990	0.000	0.158	+		
RNA binding	Machine learning consensus	0.242	0.945	0.240	0.128		0.730	
	Majority vote (BindN+(RNA), RNAbindR, Pprint)	0.457	0.821	0.551	0.116	+		
	Meta2 consensus	0.526	0.774	0.616	0.116	+		
	BindN+	0.399	0.854	0.498	0.114	+	0.706	+
	Logic consensus (BindN+ AND RNABindR AND Pprint)	0.244	0.933	0.279	0.113	+		
	RNABindR	0.575	0.718	0.643	0.105	+	0.712	+
	Pprint	0.433	0.782	0.478	0.084	+	0.667	+

The results on the COMB_T, DNA_T and RNA_T datasets (Tables 3.2 and 3.3) indicate that current methods that predict DNA-binding or RNA-binding residues are characterized by good predictive performance. However, although these predictors perform well on their own type of binding, they also substantially overpredict the other type of binding residues, i.e. predictors of RNA-binding (DNA-binding) residues also predict a large number of DNA-binding (RNA-binding) residues as RNA-binding (DNA-binding). This means that they tend to predict nucleic acids-binding residues rather than more specific DNA- or RNA-binding residues.

3.4 Conclusions

In this chapter we performed a comparative evaluation of predictive quality of runtime-efficient and conveniently available as webservers predictors of the DNA-binding (RNA-binding) residues on well-designed benchmark datasets of the DNA-binding (RNA-binding) proteins. Our empirical assessment reveals that they are characterized by acceptable levels of predictive performance. They have AUCs at about 0.7–0.8 and MCCs between 0.1 and 0.3 when measured on a hard dataset of proteins characterized by low sequence similarity to the proteins used to design these methods. However, when tested on the test data set that include both RNA- and DNA-binding proteins, we found that these predictors are guilty of substantial amounts of cross prediction, i.e. they predict RNA-binding residues as DNA-binding and vice versa. In other words, they are unable to properly separate DNA from RNA binding residues. This is likely the results of use of similar input features by the predictors of DNA and RNA binding residues and the fact that these methods were trained based on data sets that use either only DNA-binding or only RNA-binding proteins. The two existing consensus methods, MetaDBSite and Meta2, are underperforming on the corresponding DNA-binding and RNA-binding proteins respectively. This is because some of the individual methods utilized in these consensus methods are no longer available, which directly affects the predictive quality of the resulting consensus. Besides, the two consensus methods also have the cross prediction problem introduced by the individual methods when tested on both DNA-binding and RNA-binding proteins. They mis-predict a large fraction of RNA- (DNA-) binding residues as DNA- (RNA-) binding residues.

Chapter 4

Goal 2: Development of novel consensus-based predictors to improve accuracy of the prediction of DNA- and RNA- binding residues

Review of the existing sequence-based methods in Table 2.2 reveals that the current methods differ in their inputs and predictive models. Empirical assessment of their predictive performance (Tables 3.2 and 3.3) shows that these methods make different predictions on the same datasets of proteins, e.g., their values of sensitivity and specificity are widely different. These substantial differences in their designs and their predictions can be exploited to build consensus-based approaches. The articles that have introduced the two existing consensuses report that they offer improved predictive performance when compared to the use of the corresponding individual methods [12, 14]. However, to date no effort has been made to explore and empirically compare different ways to generate consensuses. Moreover, the current consensuses also cross predict DNA and RNA binding resides, as shown in Table 3.3 (high value of Ratio in the MetaDBSite and Meta2 lines). Consequently, we designed and empirically assessed several types of consensus approaches. These results were published in [77, 82], and here we summarize these findings.

4.1 Methods

The datasets used to build and test the consensuses are described in detail in Section 3.1. They include the training dataset that is used to select designs that offer the highest predictive performance, and the test dataset that is used to compare the selected design

with the existing methods including the predictors that constitute inputs to our consensuses.

We consider a comprehensive range of designs of consensuses and empirically assess their predictive performance. We are the first to investigate logic-based consensuses, which are selected as the best-performing (according to the MCC score on the training dataset) combination of k methods, $k = 1, 2, \ldots, N$ where N is the total number of predictors of RNA- or DNA-binding residues that we consider in our empirical assessment; selection of the considered methods is described in Section 3.2. The predictions of the k methods are combined using two logic-based approaches, based on logical OR and logical AND operators. Specifically, the AND-based consensus assumes that a given residue is predicted as binding only if all k methods predict it as binding; otherwise this residue is predicted as nonbinding. The OR-based approach predicts a given residue as binding if any of the k methods predict it as binding. We also considered a majority vote-based consensus predictor. This consensus predicts a residue as binding only if over half of the input methods predict so. This design generates the number of predicted binding residues that is lower than a consensus that uses only the logical OR and higher than if only the logical AND is used given that the same input predictors are used. These types of consensuses are simple to implement by an end user and do not involve any parameterization, which reduces risk of over-fitting into a given training dataset.

We also extend these relatively simple consensuses to a more sophisticated ML consensus using linear logistic regression. This meta-model implements weighted average of the input predictions and uses both the binary predictions and the propensity scores generated by the individual DNA-binding or RNA-binding predictors. We generate the regression model on the training dataset and assess its predictions on a given test dataset. Since the number of the nonbinding residues is substantially larger than the number of the binding residues in our training set, we under-sampled the nonbinding residues. For each training chain, we randomly sampled without replacement 25% (15%) of the nonbinding residues, and as a result, their number is about twice larger than the number of DNA-binding (RNA-binding) residues. The propensity scores generated by

the regression model are binarized using the cutoff that corresponds to the maximal values of MCC on the training dataset.

DNA and RNA binding amino acids share similar biochemical properties, but the corresponding interactions are associated with very different cellular functions. Thus, a given predictor should be able to separate DNA-binding residues from the RNA-binding residues. This is perhaps as crucial as the ability to separate RNA-/DNA-binding residues from the nonbinding residues. We are the first to consider the prediction with four outcomes: DNA&RNA-binding residue that binds to both DNA and RNA, DNA-binding residue (which does not bind to RNA), RNA-binding residue (which does not bind to DNA) and nonbinding residue. Such setup for the prediction should potentially address the cross-prediction between DNA and RNA binding residues. The results of the two-outcome-based predictions of the DNA binding and of the RNA binding can be combined to obtain the four outcomes as shown in Table 4.1. We implement and empirically test first-of-its-kind method for the predictions of DNA- and RNA-binding residues based on the four outcomes. We considered three different approaches: 'single consensus', 'multiple consensus' and 'machine learning consensus'. The 'single consensus' combines outputs generated by a single DNA-binding predictor and a single RNA-binding predictor. We use the best-performing, according to the MCC score on the training dataset, predictors and apply the rules from Table 5 to merge their predictions. Since consensuses of RNA-binding (DNA-binding) predictors outperform individual predictors on the training dataset, the 'multiple consensus' approach extends the single consensus by integrating results of multiple predictors of RNA-binding residues or multiple predictors of DNA-binding residues. In other words, this approach combines outputs generated by a consensus of DNA-binding predictors and a consensus of RNA-binding predictors. We examine the combination of the two logic-based consensuses (multiple consensus logic) and two majority-vote-based consensuses (multiple consensus majority vote). Also, we combine the DNA-binding residue predictions with the RNA-binding residue predictions generated by the corresponding two ML consensus predictors (multiple consensus ML). Finally, we design and test a novel consensus that combines predictions generated by all considered predictors of DNA-binding and RNA-binding residues using the logistic regression model (DNA and RNA ML consensus). This is a

single regression model rather than a combination of two regression models that is implemented in the multiple consensus ML. All these consensuses were build using only the training dataset, i.e., the specific combinations of methods used in the multiple consensuses were selected based on maximizing the MCC value on the training dataset, and the regression model for the DNA and RNA ML consensus was also generated on the training dataset.

Table 4.1. The conversion of the prediction of DNA-binding residues and the prediction of RNA-binding residues into the combined prediction of the DNA- and RNA-binding residues.

Outcome		**Two outcome predictions of RNA binding**	
		RNA-binding	Nonbinding
Two outcome predictions of DNA binding	DNA-binding	DNA&RNA-binding	DNA-binding only
	Nonbinding	RNA-binding only	Nonbinding

4.2 Results and discussion

First, we assess our consensuses and compare them with existing predictors on test datasets that include either DNA or RNA-binding proteins. Next, we perform tests on the dataset that includes both DNA and RNA-binding proteins to assess the extent of cross-prediction between RNA and DNA binding residues. Finally, we evaluate our novel design of the consensus that combines prediction of DNA and RNA binding to find out whether this approach results in a reduced cross-prediction, when compared to the other designs of consensuses and other existing predictors.

4.2.1 Predictive performance of the consensus-based predictors of DNA-binding and RNA-binding residues on the datasets with DNA-binding or RNA-binding proteins

We designed two types of consensuses: logic-based consensus, which combines individual predictors of DNA-binding residues (BindN+, DBS-PSSM, DP-Bind and ProteDNA) using all permutations of logical OR and logical AND operators; and majority vote consensus, which combines them using a majority voting rule. The best logic-based consensus on the training dataset (consensus that secures the highest MCC on the training dataset) combines BindN+ and DBS-PSSM using logical AND, which means

that a given residue is predicted as binding only if both methods predict it as binding. The best majority vote consensus on the training dataset combines BindN+, DBS-PSSM and ProteDNA, which means that a given residue is predicted as binding only if at least two of these methods predict it as binding. The results show that although the logic-based and majority vote consensuses improve the prediction performance on the training dataset, they do not deliver these improvements on the test dataset (Table 3.2). The logic-based approach provides similar MCC when compared with the best-performing individual predictor, DBS-PSSM, on the test dataset. Majority vote consensus only slightly improves MCC by 0.7%. The reason for the lack of improvement is that the test dataset is dissimilar to the training dataset (<30% sequence similarity) and such simple combinations of individual predictors did not translate well between these two datasets. Motivated by this, we extended these designs of the consensuses into a more advanced ML consensus that applies linear logistic regression. The ML consensus outperforms all single predictors by at least 4.7% in MCC, 3.4% in AUC, and these differences are statistically significant (Table 3.2). Analysis of the results concerning prediction of the RNA-binding residues leads to similar observations (Table 3.2). The logic-based consensus, which outperforms other considered consensuses on the training dataset, integrates predictions from BindN+, RNABindR and Pprint using logical AND. The majority vote consensus also combines these three individual predictors. Similar to the results for the DNA-binding, these two types of simple consensuses do not perform well on the test dataset. They only achieve equivalent or slightly worse MCC compared with the best-performing predictor, BindN+, on this dataset. However, the ML-based consensus outperforms all the individual predictors. More specifically, its MCC is higher by at least 1.5%, AUC by at least 1.8% and specificity by at least 7.2%.

To sum up, our analysis reveals that a simple consensus based on majority vote or logic does not improve predictive performance when applied to predict proteins that are dissimilar to the proteins that were used to develop this consensus. However, a more sophisticated logistic regression-based consensus outperforms all individual methods in the prediction of DNA-binding and RNA-binding residues, even for the dissimilar protein chains.

4.2.2 Predictive performance of the consensus-based predictors of DNA-binding and RNA-binding residues on the dataset with DNA- and RNA-binding proteins

Evaluation of the predictors of the DNA-binding residues shows that the majority vote and logic consensuses do not offer improved MCC when compared with their input methods on this test dataset that includes both DNA and RNA-binding proteins, but their Ratio values are reduced to 23.2% (see the upper half of Table 3.3). This means that the individual predictors do not agree on the misprediction of the RNA-binding residues as DNA-binding for a substantial number of cases, i.e. they mispredict different residues. The ML-based consensus that we designed again outperforms all other predictors on this dataset, i.e., it provides improvements on the datasets with only RNA or only DNA binding proteins and on the dataset with both RNA and DNA binding proteins. It secures the highest MCC equal 0.311 and also the highest AUC of 0.841, and these improvements are statistically significant (Table 3.3). Figure 4.1A shows the ROCs for the ML consensus and all the individual predictors that generate the propensity scores on the COMB_T datasets. Notably, the TPR of our ML consensus is higher than the TPR of any individual predictors for almost the entire range of FPR values. However, this consensus still has a problem of substantial levels of mispredictions between DNA and RNA binding residues, which is demonstrated by the moderate values of Ratio (Table 3.3). We attempt to solve this problem by proposing a new design of the ML consensus that combines prediction of both DNA- and RNA-binding residues.

Similarly, assessment of the predictors of the RNA-binding residues on the COMB_T dataset demonstrates that logic-based and majority vote consensuses do not improve predictive performance when compared with their input predictors on this dataset, the former consensus provides relatively low values of Ratio (see lower half of Table 3.3). However, the ML-based consensus outperforms all the individual predictors. It secures the highest MCC, AUC and specificity and the lowest (best) Ratio. The ROCs of this consensus and all the individual predictors are shown in Figure 4.1B. Overall, the ML consensus achieves the best performance when considering the entire range of FPR values. Pprint performs well at low FPR (<0.05) value, while its TPR drops substantially

for higher values of FPR. RNABindR curve overlaps with our ML consensus curve at larger values of FPR, but this method has lower TPR when FPR <0.45. The only weakness of the ML consensus is the relatively high values of Ratio, in spite of the fact that it is lower than the Ratio of the other methods.

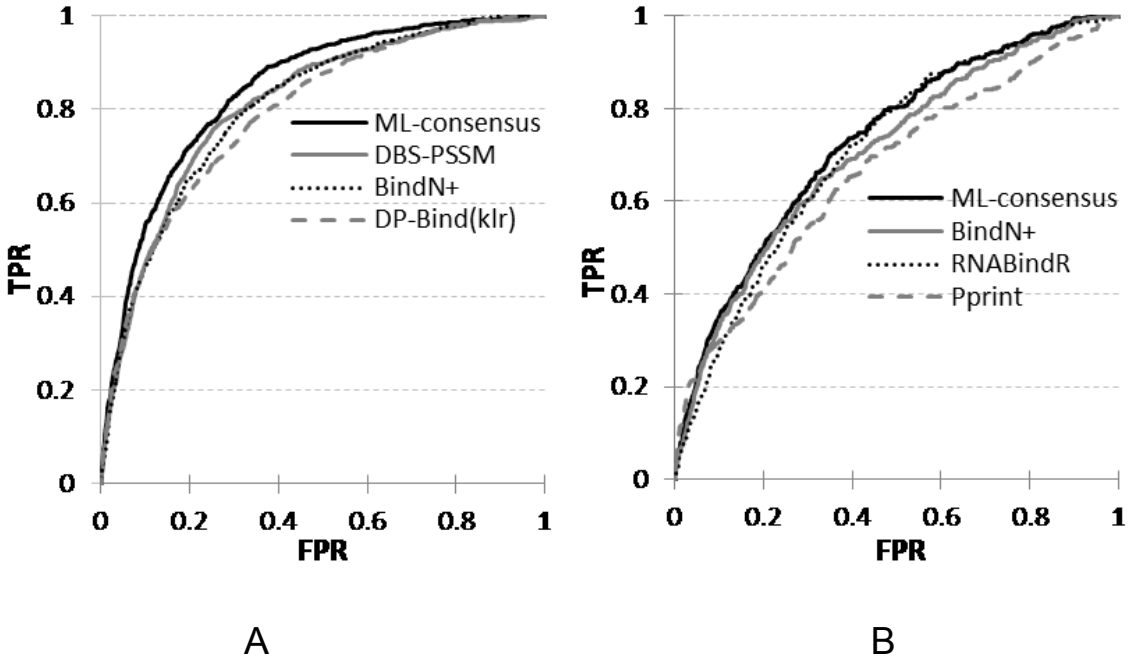


Figure 4.1. The ROCs for the machine learning consensuses and the individual predictors of DNA- and RNA-binding residues on the COMB_T dataset.

Panels A and B compare the DNA-binding predictors and the RNA-binding predictors, respectively.

4.2.3 Predictive performance of the consensus-based combined predictor of DNA- and RNA-binding residues

The published predictors were designed specifically to target either protein–DNA or protein–RNA interactions. The results on the COMB_T, DNA_T and RNA_T datasets (Tables 3.2 and 3.3) indicate that these methods are characterized by good predictive performance. However, although these predictors perform well on their own type of binding, they also overpredict the other type of binding residues, i.e. predictors of RNA-binding (DNA-binding) residues also predict a large number of DNA-binding (RNA-binding) residues as RNA-binding (DNA-binding). This means that they tend to predict nucleic acids-binding residues rather than more specific DNA- or RNA-binding residues.

One way to potentially alleviate this drawback is to redefine these two prediction tasks as a single prediction with four outcomes: DNA&RNA-binding, DNA-binding, RNA-binding and nonbinding residue. We are the first to design such predictors and comprehensively assess their predictive performance. As we explain in section 4.1, our designs integrate multiple predictors of DNA- and RNA-binding residues based on three types of consensuses: single consensus, multiple consensus and ML consensus. The single consensus combines the best-performing (i.e. providing the highest MCC) on the training dataset predictor of DNA-binding residues, BindN+ (DNA version), with the best-performing predictor of RNA-binding residues, BindN+ (RNA version). The multiple consensus approach combines multiple predictors of DNA-binding residues and RNA-binding residues. We consider three designs of the multiple consensuses: multiple consensus logic, multiple consensus majority vote and multiple consensus ML. Moreover, the DNA and RNA ML consensus combines predictions generated by all considered predictors of DNA-binding and RNA-binding residues using the logistic regression model. We assess these methods on the COMB_T test dataset. There are no DNA&RNA-binding residues in this dataset, so we cannot compute sensitivity and MCC for this outcome.

All multiple consensuses outperform the single consensus in MCC for the combined prediction of DNA and RNA binding (Table 4.2). Moreover, the single consensus substantially overpredicts the RNA&DNA outcome with the corresponding specificity at 0.908. The multiple consensuses reduce this overprediction obtaining specificities between 0.922 and 0.957. The result of this overprediction for both single and multiple consensuses is the relatively low sensitivity for the prediction of DNA binding and the prediction of RNA binding, i.e. many of the RNA or DNA binding residues are predicted to bind both RNA and DNA. However, the RNA and DNA ML consensus, which is inherently designed to predict the four outcomes, correctly does not predict the DNA&RNA binding residues (specificity = 1) and secures high values of specificity and MCC. Its MCC is higher by 7 and 3.4% for the prediction of DNA-binding residues and RNA-binding residues, respectively, when compared with the best multiple consensus. This result demonstrates that the RNA and DNA ML consensus provides improved predictive performance when compared with the other consensuses.

Table 4.2. Results of empirical assessment of consensus-based methods on the COMB_T dataset when considering prediction of combined DNA- and RNA-binding residues and individual prediction of DNA- or RNA-binding residues.

There are no DNA&RNA binding residues in this dataset and thus we cannot compute sensitivity and MCC for this outcome. Values of Ratio cannot be computed for the combined prediction of RNA and DNA binding. The "multiple consensus logic" utilizes the two best-performing logic-based consensuses that we built for the prediction of DNA-binding residues and RNA-binding residues, respectively; "multiple consensus majority vote" combines the two best-performing majority vote-based consensuses for the prediction of DNA- and RNA-binding residues, respectively; "multiple consensus machine learning" is the combination of the two machine learning consensus for the prediction of DNA- and RNA-binding residues, respectively; and "DNA and RNA machine learning consensus" combines predictions generated by all considered predictors of DNA-binding and RNA-binding residues using logistic regression model.

		Prediction of DNA and RNA binding				**Prediction of DNA or RNA binding**	
		DNA&RNA	**DNA**	**RNA**	**non-DNA & non-RNA**	**DNA vs. non-DNA**	**RNA vs. non-RNA**
Sensitivity	Single consensus	N/A	0.101	0.164	0.839	0.482	0.399
	Multiple consensus logic	N/A	0.207	0.103	0.899	0.424	0.244
	Multiple consensus majority vote	N/A	0.085	0.259	0.821	0.447	0.457
	Multiple consensus machine learning	N/A	0.261	0.078	0.914	0.478	0.242
	RNA and DNA machine learning consensus	N/A	0.392	0.125	0.929	0.392	0.125
Specificity	Single consensus	0.908	0.962	0.942	0.552	0.888	0.854
	Multiple consensus logic	0.957	0.951	0.974	0.438	0.919	0.933
	Multiple consensus majority vote	0.922	0.976	0.895	0.590	0.916	0.821
	Multiple consensus machine learning	0.955	0.956	0.986	0.451	0.922	0.945
	RNA and DNA machine learning consensus	1.000	0.941	0.981	0.409	0.941	0.981
MCC	Single consensus	N/A	0.074	0.072	0.277	0.256	0.114
	Multiple consensus logic	N/A	0.159	0.076	0.281	0.267	0.113
	Multiple consensus majority vote	N/A	0.086	0.081	0.280	0.277	0.116
	Multiple consensus machine learning	N/A	0.220	0.084	0.318	0.311	0.128
	RNA and DNA machine learning consensus	N/A	0.290	0.118	0.315	0.290	0.118
Ratio	Single consensus	N/A	N/A	N/A	N/A	0.289	0.498
	Multiple consensus logic	N/A	N/A	N/A	N/A	0.232	0.279
	Multiple consensus majority vote	N/A	N/A	N/A	N/A	0.232	0.551
	Multiple consensus machine learning	N/A	N/A	N/A	N/A	0.267	0.240
	RNA and DNA machine learning consensus	N/A	N/A	N/A	N/A	0.183	0.064

We applied the considered consensuses to predict DNA-binding residues and RNA-binding residues separately (the two right-most columns in Table 4.2). The predictions of the consensuses that consider four outcomes are converted into prediction of DNA-binding residues as follows: 'DNA&RNA-binding' and 'DNA-binding' are assigned as 'DNA-binding'; 'RNA-binding' and 'nonbinding' are assigned as 'nonbinding'. For the prediction of RNA-binding residues, the conversion assumes 'RNA-binding' for the 'DNA&RNA-binding' and 'RNA-binding' predictions, and 'nonbinding' for the 'DNA-binding' and 'nonbinding' predictions. Table 4.2 shows that the two ML consensuses outperform the other types of consensuses having higher values of MCC and specificity. The main observation is that the RNA and DNA ML consensus offers substantially reduced values of Ratio, at 0.183 and 0.064 for the DNA and for the RNA binding, respectively, compared with the second best Ratios of 0.232 and 0.240. This means that this novel type of consensus generates predictions with lower rate of mispredictions between DNA- and RNA-binding residues.

We compare results generated by the two ML consensuses for the prediction of DNA-binding residues with the considered predictors of DNA-binding, see Figure 4.2. The DNA and RNA ML consensus obtains MCC of 0.290, which is lower than MCC of 0.311 of the multiple consensus ML for the prediction of DNA-binding residues (black bars in Figure 4.2). However, the former consensus has by far the lowest values of Ratio at only 0.183 (gray bars in Figure 4.2), except for the ProteDNA that predicts a small subset of DNA-binding residues and has the lowest MCC. Similar conclusions are true when considering prediction of the RNA-binding residues (Figure 4.2). The DNA and RNA ML consensus secures MCC of 0.118, which is lower compared with the best MCC of 0.128 obtained by the multiple consensus ML. It also boasts the lowest value of Ratio at 0.064 compared with the second lowest value at 0.240. Most importantly, the novel DNA and RNA ML consensus improves over all individual predictors having higher MCC while providing much lower Ratio for prediction of the RNA and the DNA binding residues (Figure 4.2). These results suggest that the development of consensuses for the combined prediction of DNA- and RNA-binding residues could offer a viable solution to generate high-quality prediction of DNA- or RNA-binding residues where the cross-predictions are substantially reduced.

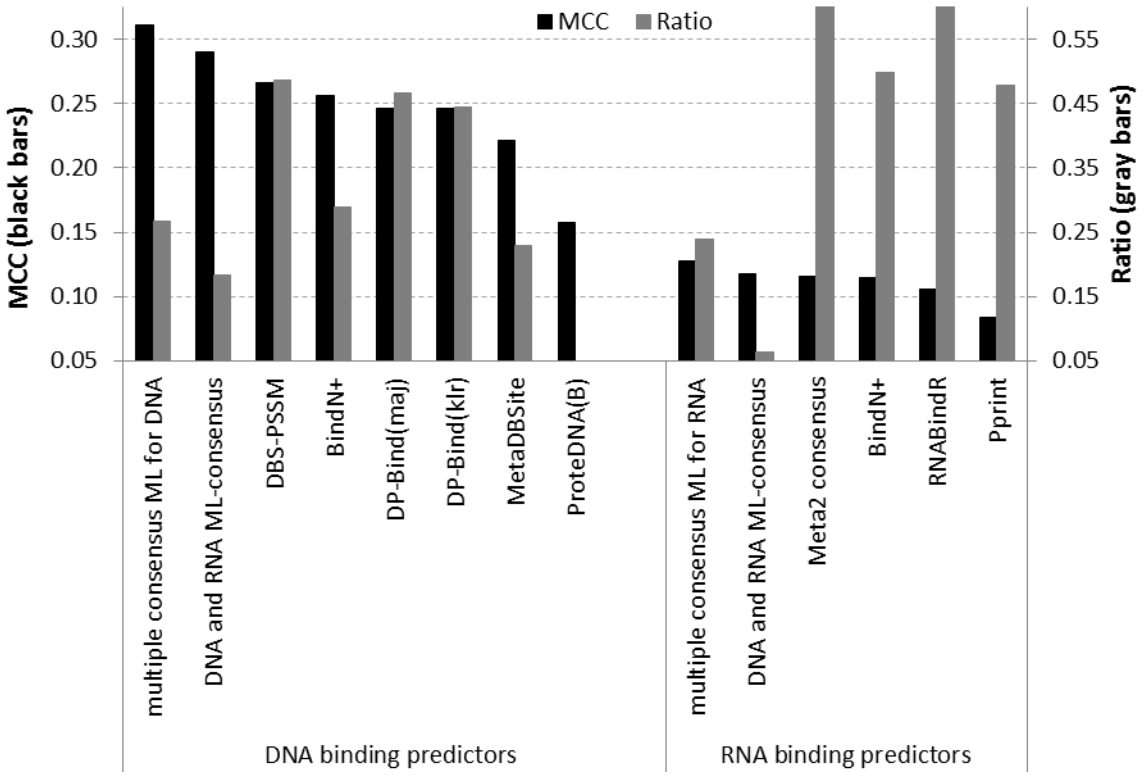


Figure 4.2. Comparison between the DNA and RNA machine learning (ML) consensus that targets combined prediction of DNA- and RNA-binding residues and the considered predictors of DNA- or RNA-binding residues on the COMB_T test dataset.

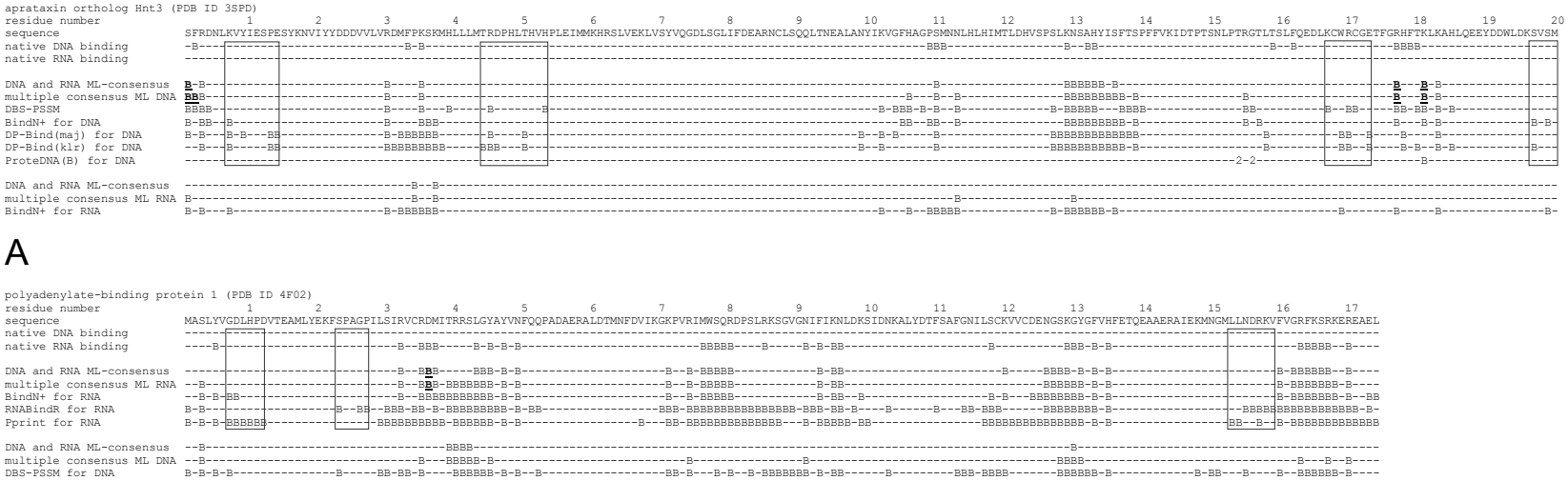
The predictors of DNA- or RNA-binding residues include the two machine learning based DNA- or RNA- binding consensuses. The evaluation considers prediction of DNA-binding residues (left side of the figure) and prediction of RNA-binding residues (right side of the figure) on the COMB_T test dataset.

4.3 Case studies

We illustrate predictions of the most successful in our tests ML consensuses and all considered individual predictors of DNA- and RNA-binding residues on two proteins selected from the test dataset. The overall predictive performance measured with MCC for the consensuses on these two proteins is similar to the value on the whole test dataset. Figure 4.3A compares predictions for the DNA-binding aprataxin ortholog Hnt3 (PDB ID: 3SPD). We observe that virtually all binding regions (except for the residues near position 160) were captured by most predictors. Both ML consensuses for the prediction of DNA-binding residues filter FP predictions (nonbinding residues predicted as binding) at both termini (shown using boxes in Figure 4.3A). These boxed regions are relatively far away from the native binding regions. Moreover, they annotate a few binding residues

that were predicted by a subset of individual predictors (shown in bold and underline in Figure 4.3A) which are either correct predictions or immediately adjacent to the native binding residues. The RNA and DNA ML consensus reduces some of the FP generated by the multiple consensus ML, particularly near position 135. The best performing in our tests individual method that predicts RNA-binding residues (last line in Figure 4.3A) generate FP that generally line up with the location of the DNA binding residues. However, the ML consensuses, in particular the novel DNA and RNA ML consensus, substantially reduce these mispredictions. Similar observations are true for the predictions for the RNA-binding polyadenylate-binding protein 1 (PDB ID: 4F02) shown in Figure 4.3B. The two ML consensuses filter out FP generated by the individual predictors of RNA binding residues in the boxed regions that are relatively far from the native binding regions. They also correctly locate binding residue at position 36 that was missed by one of the individual RNA-binding predictors. Moreover, the best performing in our tests predictor of the DNA binding residues incorrectly predicts relatively many DNA binding residues (last line in Figure 4.3B) which again align with the native RNA binding residues. The ML approaches for the prediction of DNA binding residues reduce the number of these mispredictions by a large factor.

Overall, the case studies demonstrate that the ML consensuses successfully reduce some of the FP generated by the individual predictors and correctly predict binding residues even if some of the individual predictors do not. The novel DNA and RNA ML consensus further reduces some of the FP generated by the multiple consensus ML.



A

B

Figure 4.3. Two case studies that illustrate the working of the machine learning consensuses. Panel A concerns the DNA-binding aprataxin ortholog Hnt3 (PDB ID: 3SPD) and Panel B show the RNA-binding polyadenylate-binding protein 1 (PDB ID: 4F02).

'B' denotes binding residues and '-' denotes the non-binding residues. Boxes and bold denote results that are discussed in the text. Disordered regions in these two proteins (regions with no coordinates) are omitted. Panel A includes the following lines (from top to bottom): residue number, sequence, native annotation of DNA binding residues, native annotation of RNA-binding residues, blank line, predictions from the DNA and RNA machine learning model for the DNA binding residues, machine learning model for DNA-binding, DBS-PSSM, BindN+(DNA), DP-Bind(maj), DP-Bind(klr), ProteDNA(B), blank line, predictions from the DNA and RNA machine learning model for the RNA binding residues, machine learning model for the RNA-binding residues, and the best individual RNA-binding residue predictor BindN+(RNA). Panel B includes the following lines (from top to bottom): residue number, sequence, native annotation of DNA binding residues, native annotation of RNA-binding residues, blank line, predictions from the DNA and RNA machine learning model for the RNA binding residues, machine learning model for the RNA-binding residues, BindN+, RNABindR, Pprint, blank line, predictions from the DNA and RNA machine learning model for the DNA binding residues, machine learning model for the DNA-binding residues, and the best individual DNA-binding residue predictor DBS-PSSM.

4.4 Conclusions

Motivated by the prior success in building consensus-based predictors, we designed and empirically tested simple logic-based consensuses based on combinations of logical OR and logical AND operators, a majority vote consensus, and a more sophisticated ML consensus. We show that the logic and majority-vote-based consensuses do not offer improvements when tested on the hard (dissimilar to the training dataset) test dataset. However, the ML consensuses provide improved predictive performance when compared with the individual methods for the prediction of DNA-binding residues and for the prediction of RNA-binding residues on the same hard test dataset. We also performed first-of-its-kind study concerning combined prediction of DNA- and RNA-binding residues. We designed three types of consensuses to address this prediction, including a ML-based approach. The ML consensus offers strong predictive performance in the combined prediction and, most importantly, also for the prediction of DNA-binding or RNA-binding residues individually. We empirically show that this consensus provides higher values of MCC compared with the best-performing individual predictors while it also substantially reduces the cross-prediction. Finally, we illustrate these empirical results using two case studies. They demonstrate that the ML consensuses filter out false predictions of the binding residues generated by individual predictors that are located relatively far from the native binding residues.

Chapter 5

Goal 3: Development of DRNApred, a new high-throughput method that accurately and specifically predicts only DNA-binding and only RNA-binding residues

In Chapter 3 we found that cross prediction between RNA and DNA binding residues is a widespread and substantial problem. Although the ML-consensuses for the prediction of DNA/RNA-binding residues built in Chapter 4 help to improve the predictive quality in term of AUC and MCC, they still confuse DNA-binding with RNA binding residues. The DNA and RNA ML-consensus (consensus with the 4 outcomes) is so far the only approach that provides a working solution to reduce the amount of these mis-predictions. However, this consensus is inconvenient to use and is not runtime efficient since it combines 8 individual predictors (5 for DNA and 3 for RNA) for which the predictions have to be retrieved from webservers. To this end, in this chapter we aim to build a new high-throughput method that offers good predictive quality and solves the problem of cross prediction between DNA-binding and RNA-binding residues.

5.1 Benchmark dataset

We collected new DNA/RNA-binding complexes from PDB to expand our existing datasets that were described in Section 3.1. We collected complexes that were solved with resolution <2.5 Å and that were released after the date that the existing datasets have been collected. They include 564 protein-DNA, 72 protein-RNA, and 16 protein-DNA-

RNA binding complexes. After extracting protein chains, we have 892 DNA-binding and 145 RNA-binding sequences. We combine these new chains with our existing datasets and consequently obtain total of 2827 DNA-binding and 1125 RNA-binding chains. Next, following the same protocol as described in Section 3.1 we transfer binding annotations between similar proteins. We cluster proteins that share ≥80% sequence similarity and ≥0.5 TM scores, and then transfer annotations between proteins in the same cluster. However, this time we first combine DNA-binding and RNA-binding sequence together and then cluster them. Note that in Chapter 3 we transferred annotations separately for the DNA binding proteins and separately for the RNA binding proteins, which is less accurate compared to transferring annotations for the combined set of DNA and RNA binding proteins. In this way a given cluster may contain both DNA-binding and RNA-binding proteins. We transfer both types of annotations from all chains into the representative chain (with the largest number of binding residues) in each cluster. We also update the deposition date of the representative chain to the earliest release time among all chains in the same cluster. We split all resulting representative chains into training and test datasets by the deposition dates. We observe that the data sets used by the considered existing predictors of DNA- and RNA binding residues were collected before November 2010. Correspondingly, the binding proteins released before November 2010 are assigned into the training dataset, and proteins released after November 2010 which are less likely to be used to train the published methods are assigned into the test dataset. To reduce the sequence similarity between training set and test dataset, we filter the test set by removing sequences that share >30% sequence similarity with any training sequence; this is based on pairwise sequence similarity between a given test sequence and each training sequence that we computed with the bl2seq program. We further remove 5 and 3 proteins with length ≥1000 residues in the training and test dataset, respectively, since some of the input methods utilized to develop our model and some of the existing predictors of DNA and RNA binding residues we compare with cannot complete their predictions for such long proteins. Finally, our training dataset contains 488 DNA- and/or RNA-binding proteins, 7823 DNA-binding residues, 95161 nonDNA-binding residues, 4699 RNA-binding residues, and 98241 nonRNA-binding residues. The independent test dataset (i.e., sharing low sequence similarity with the training dataset) includes 82 DNA-

and/or RNA-binding chains, 968 DNA-binding residues, 17926 nonDNA-binding residues, 808 RNA-binding residues, and 18074 nonRNA-binding residues. Residues with missing coordinates in training and test dataset (disordered residues for which we cannot compute annotation of binding) are excluded from our evaluation. Since there are substantially more negative samples (nonbinding residues) than the positive samples (binding residues), we balance the training dataset by under-sampling the nonbinding residues. Among the nonbinding residues, a small fraction (similar to the number of binding residues) binds to the other type of nucleic acid. i.e., these are DNA binding residues when the positive binding residues are RNA binding and vice versa. These residues are important for the predictive model to learn how to discriminate between DNA and RNA binding residues. Therefore, we keep all of these non-binding residues in the training dataset. We under-sample 25% (15%) of the remaining nonbinding residues that do not bind to either DNA or RNA molecule. As a result, the number of the non-binding residues is about twice larger than the number of the DNA-binding (RNA-binding) residues.

We also develop a negative set of proteins that are unlikely to bind either DNA or RNA. Similar to the way the negative dataset was developed in [77], we consider human proteins from the complete human proteome collected from the UniProt database. We include proteins that satisfy the following stricter, compared to [77], seven conditions: (1) their subcellular location is not in nucleus, chromosome, or nucleoplasm; (2) their functional annotations expressed with the gene ontology (GO) terms do not include DNA, RNA, nucleotide, nucleic acid, DNA binding, RNA binding, or nucleotide binding; (3) protein names do not contain DNA, RNA, nucleic acid, nucleotide, or ribosomal; (4) their function annotated in UniProt does not include DNA binding, RNA binding, nucleic acid binding, or nucleotide binding function; (5) their UniProt records do not have the following keywords: DNA, RNA, nucleic acid, nucleotide, ribosomal, ribosome, ribosomal protein, or chromosome; (6) they are not annotated as interacting with DNA, RNA, or nucleotide; and (7) they were reviewed in UniProt (i.e., these proteins underwent manual evaluation that assures higher quality of the annotations compared to the un-reviewed proteins). Using these criteria we collect a set of 5996 human proteins that are unlikely to bind either DNA or RNA. Based on the protocol in [77], we further

filter these proteins by removing the sequences that share ≥30% sequence similarity with each other or with any sequence in the training dataset. This reduces the redundancy among the dataset and also reduces the possibility of these proteins to bind to DNA or RNA given that proteins in the training dataset bind to these nucleic acids. To reduce computational cost of evaluation on the negative dataset, particularly given the high runtime of some of the existing predictors, we selected at random 82 proteins from the resulting dataset. These 82 proteins form the negative dataset of the nonbinding proteins.

5.2 Development of the DRNApred predictor

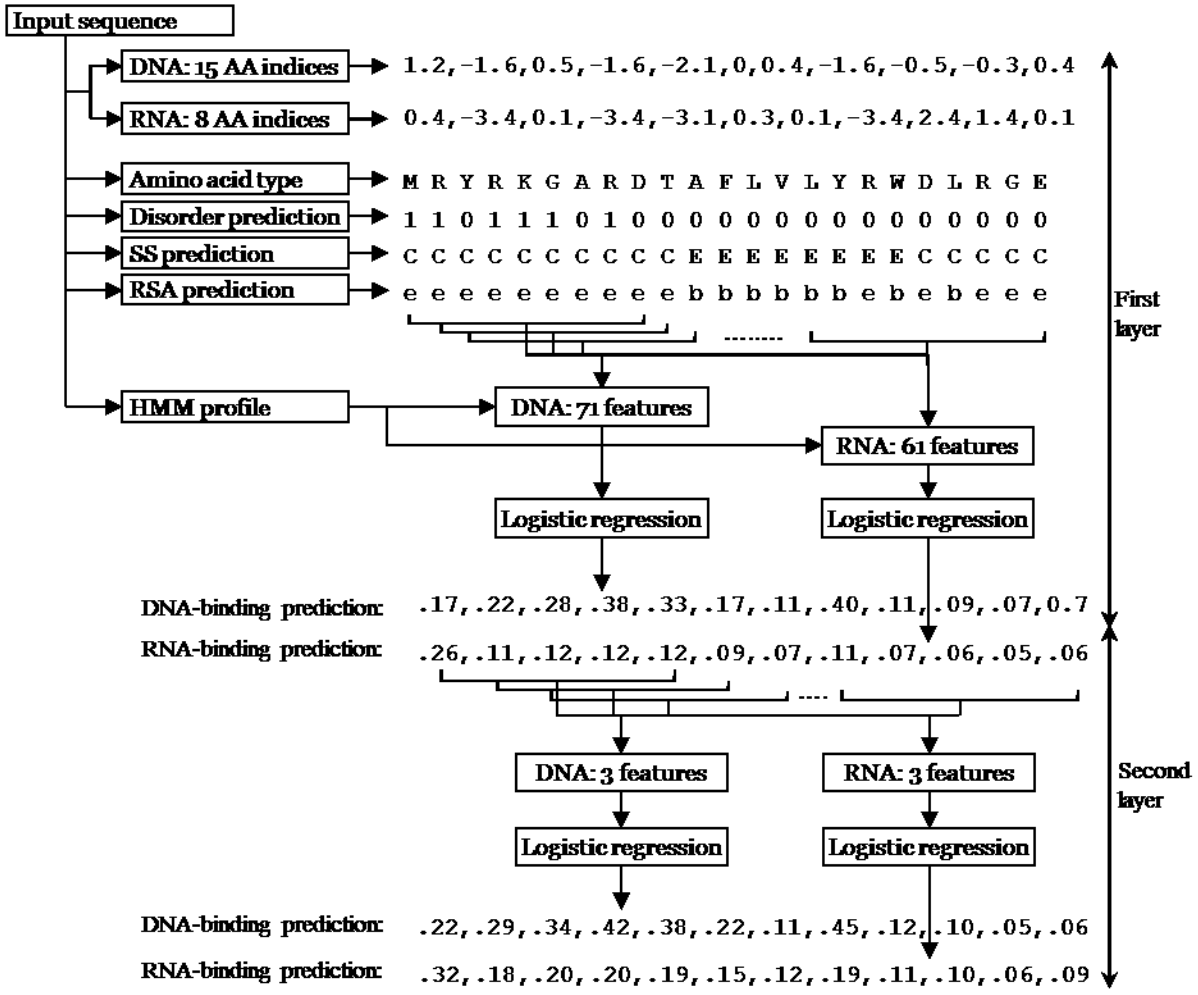


Figure 5.1. Architecture of DRNApred predictor

DRNApred generates DNA (RNA)-binding residue predictions using a 2-layer design, see Figure 5.1. The first predictive layer includes three steps. **First**, a variety of physicochemical and biochemical properties (15 and 8 for the prediction of DNA and RNA binding, respectively) together with the putative intrinsic disorder, secondary structure and solvent accessibility are calculated and predicted from the input protein sequence. **Second**, these inputs are processed using a sliding window to generate a set of 41 (31) numeral features which are combined with the 30 features based on the evolutionary profile HMM generated with the HHblits method to encode the input sequence. This small set of features was empirically selected using the training dataset from a large set of considered numerical features that were generated from inputs collected in the first step of the first layer. **Third**, the combined set of 71 (61) features is input into a logistic regression model that generates predictions of the DNA (RNA)-binding residues. Selection of logistic regression is motivated by several factors: (1) this model has been already successfully used in a related study [72]; (2) it is a simple linear models, which reduces likelihood of overfitting the training dataset compared to more complex models that use a larger number of parameters; (3) it is fast to generate on the training dataset, which is an important advantage given a relatively large size of our dataset; (4) it is fast to make prediction using this model, which is crucial given our goal to develop a runtime efficient (high throughput) predictor; (5) it provides good predictive performance when used to implement consensus-based prediction of RNA and DNA binding residues [82] (in that application we have utilized five different popular types of machine learning algorithms including logistic regression, SVM, C4.5 decision tree, k-nearest neighbor and Naïve Bayes and the consensus implemented using the logistic regression has secured the best predictive quality); and (6) we initially also considered SVM, compared it with logistic regression on the training dataset and these preliminary results favored logistic regression model. The second predictive layer re-predicts the predictions generated in the first layer to improve predictive performance. In this layer, we explore information about the putative binding of adjacent residues, including putative annotations of both RNA and DNA binding. Intuitively, residues surrounded by a large number of putative DNA binding residues are more likely to bind DNA compared to residues surrounded by fewer or no DNA binding residues. Also, residues surrounded

by many putative DNA binding residues are less likely to be predicted as RNA binding even if they are also surrounded by a modest number of putative RNA binding residues, compared to the residues that are not surrounded by the DNA binding residues. We use the predictions of the DNA-binding and RNA-binding residues generated in the first layer as input to compute 3 (3) numerical features using a sliding window; these features were empirically selected from a larger set of features using the training dataset. These 3 (3) features are then inputted into two logistic regression models that re-predict the DNA-binding and RNA-binding residues respectively.

Feature-based encoding of the input protein sequence

We apply a shotgun approach by generating a large variety of structural and physiochemical properties of the input sequence and encoding them into a large number of numerical features utilizing sliding windows of different sizes. Next, a smaller subset of predictive and non-redundant features is empirically selected from this large set of considered features.

In the first step of the first layer, we consider a comprehensive set of properties of the input sequence including amino acid (AA) type, information derived from putative intrinsic disorder, secondary structure (SS) and solvent accessibility (SA), AA indices that quantify physicochemical properties of residues in the input protein sequence, and evolutionary profile of that sequence. These properties have already been successfully used in the previous studies that focused on the prediction of DNA or RNA binding, which were reviewed in ref. [77]. Specifically, intrinsic disorder is predicted by the IUPred [83] and Espritz [84] methods. SS is predicted with the fast version of PSIPRED that does not use sequence alignment [85]. SA is predicted by the fast version of PROFphd [86], NETASA [87] and RVP-net [88] methods. We note that the above mentioned predictions were performed using runtime-efficient methods to ensure that our predictor is also computationally efficient. AA indices are collected from the AAindex database [89]. Some of these indices are redundant with each other (they quantify similar properties) and some may not be relevant to the prediction of the RNA and DNA binding. Therefore, we empirically select a subset of non-redundant and predictive AA indices. Specifically, we remove the indices that are incomplete (with missing values) and those

that are not predictive (lack correlations with the prediction outcomes) or redundant (have high mutual correlation with other indices). We compute the point-biserial correlation (PBC) of each index with the DNA-binding annotations and RNA-binding annotations in the training dataset to quantify whether these indices are predictive. Indices with PBC <0.05, which indicates that they offer low predictive power, were removed. Next, we remove the redundant indices among the remaining indices. The indices were sorted based on PBC values in the descending order. We start from the top ranked index, and the next ranked index is added into the pool of retained indices only if its Pearson correlation (PCC) with each of the indices in the pool is ≤ 0.9. As a result, we selected 164 indices that are predictive and non-redundant for the prediction of the DNA-binding residues and 105 indices for the prediction of the RNA-binding residues. Finally, the evolutionary profile HMM is generated using HHblits method [90] with the default parameter settings on the nr database. The profile is in the form of $N * 30$ matrix, where N is the length of the input protein sequence. For each position $n_i, i = 1, \dots N$, in the input sequence, this profile provides 30 scores including 20 scores representing the frequencies to observe each of the 20 amino acids at this particular position n_i in homologous proteins, 7 transition frequency scores indicating the probabilities to observe a match, insertion or deletion after this position, and the 3 local diversity values that quantify the diversity of the aligned sequences in a region around this position.

In the second step of the first layer, the input properties are further processed using sliding windows to generate a large set of numerical features. Sliding window is centered on the residue that we want to predict to accommodate for the bias introduced by adjacent residues. For each type of input property, we consider two types of features: features computed for each residue in the window (per residue features) and features computed by aggregation of information coming from multiple residues in the window (aggregated features):

- We calculate the per residue features by considering each residue in a window individually. We apply a sliding window of size 3 to include the information about the residue a_i that we want to predict and its left and right neighbors. Thus, the feature vector for the residue a_i is represented by $[V_{i-1}, V_i, V_{i+1}]$, where V is the input

information that includes the AA type and predicted disorder, SS, and SA. For the residues at the either termini of the sequence (C-terminus and N-terminus) where there is no neighbor on either left or right side, we fill the corresponding information with default values.

- Motivated by the recent work in [72], we also aggregate the input information over the sliding window. The considered information includes AA types, values of selected AA indices, and predicted disorder, SS, and SA. We aggregate their values over the whole sliding window. Moreover, we filter the positions in the window using the SA predictions, and calculate the aggregated values only for the solvent exposed residues in the window. We vary the window size from 9 to 21 with a step of 2. We also compute the same aggregated values for the entire protein chain.

Detailed description of the calculation of the per-residue features and the aggregated features is given in Table 5.1. In total, we generate 4580 features for the prediction of the DNA-binding residues, and 3990 features for the prediction of RNA-binding residues.

Table 5.1. Description of features that were considered in the design of the DRNApred method.

Exposed residues are determined using the prediction from the PROFphd method. For the aggregated feature that were computed for the exposed residues, we calculate the average value in two ways: sum of the information for the exposed residues divided by the number of the exposed residues in the window, and sum of the information for the exposed residues divided by the size of the window. The standard deviation is only calculated in the first case. Since the AA indices utilized in the DNA-binding and RNA-binding predictions are different, the corresponding aggregated features are different. Consequently, the number of features for the DNA-binding prediction is shown first and is followed by the number of features for the prediction of RNA-binding that is given inside brackets.

Feature type	**Input type**		**Description**	**Window size**	**number of features**
Per residue	Amino acid type		20 dimensional binary vector to encode the amino acid type	w=3	60
	Disorder, SS, RSA		We include probability and binary values from 9 methods (5 disorder + 3 RSA + SS).	w=3	90
	HMM profile		20 amino acid emission frequencies + 7 transition frequencies +3 local diversities	NA	30
Aggregated	All residues	Amino acid type	amino acid composition (20 values for each window size)	w={9,11,13,15,17,19, 21, protein length}	160
			Amino acids are divided into 3 groups based on their properties (e.g. charge, hydrophobicity, etc.) [91]. We calculate the composition/transition/distribution of the amino acids in each group.	w={9,11,13,15,17,19, 21, protein length}	1176
		Disorder, SS, RSA	content of binary predictions over the window of size w	w={9,11,13,15,17,19, 21, protein length}	328
			average value and standard deviation of the probability predictions over the window of size w		
		AA indices	average value and standard deviation of AA indices over the window of size w	w={9,21}	656(420)
	Exposed residues	Amino acid type	amino acid composition of the exposed residues	w={9,11,13,15,17,19, 21, protein length}	320
			composition of the exposed amino acids in each group	w={9,11,13,15,17,19, 21, protein length}	336
		Disorder, SS, RSA	content of binary predictions of the exposed residues over the window of size w	w={9,11,13,15,17,19, 21, protein length}	440
			average value and standard deviation of the probability predictions of the exposed residues over the window of size w		
		AA indices	average value and standard deviation of AA indices of the exposed residues over the window of size w	w={9,21}	984(630)

Feature selection and parameterization of the predictive models in the first layer

To implement the second step of the first layer, we need to select a subset of non-redundant and predictive features which are useful to discriminate between binding residues and non-binding residues. There are two types of non-binding residues in our dataset: the non-binding residues that do not bind to either DNA or RNA and the non-binding residues that do not bind to the target type of nucleic acid but bind to the other type. For example, in the prediction of the DNA-binding residues, the nonDNA-binding residues can be further divided as non-binding residues that do not bind to either DNA or RNA, and the RNA-binding residues. The DNA-binding and RNA-binding residues share similar biochemical properties, thus they are likely to be confused by a predictive model. Hence, our aim is to select features that are not only useful to differentiate between binding and non-binding residues, but also to minimize the number of DNA-binding residues that are confused for RNA-binding and vice versa. To do this we assign weights to the residues in our training dataset. By default, the residues have a weight of 1. We assign weight >1 to the residues that could be cross predicted, e.g. the RNA-binding residues for the dataset we used to develop DNA-binding prediction method, and the DNA-binding residues for the dataset we used to develop RNA-binding prediction method. Next, the weight values are passed along with the value of the features to the logistic regression model. When building the model, the prediction errors for the instances (residues) with weight >1 are adjusted (increased) compared to the prediction errors for the instances with weight of 1. This way the regression will minimize the mis-prediction of residues with weights > 1. We select the best weight value by considering values ranging from 1 to 4 with step of 0.2. For each of the considered weight value, we empirically select a subset of predictive and non-redundant features from the original set of considered features using a two-step feature selection. We perform the selection exclusively using the training dataset with the 5-fold cross validation protocol. The training proteins were divided into 5 folds such that protein chains in a given test fold are dissimilar to the training sequences (sequences in the training folds). This simulates the tests on the test dataset. We cluster the chains in the training dataset using CD-HIT at 30% sequence identity, and assign the proteins that are clustered in the same group to the same cross-validation fold. In the first step of feature selection, we apply a wrapper-based

approach to rank the features. For each feature, we calculate its predictive performance (measured by AULC) when used as an input to univariate logistic regression model based on the 5-fold cross validation on the training dataset. In the second step, we execute the best first search-based feature selection using the wrapper with logistic regression model to select a subset of predictive and non-redundant features. Starting with the top ranked feature, we accept the next best-ranked feature into a selected set of features only if the addition of this feature improves AULC by at least 0.0001 based on the 5-fold cross validation on the training dataset when compared with the feature set before this addition; to compare, the AULC of a random predictor = 0.003. We go through the sorted list of features once to select the subset of features. Depending on the weight values (we repeat the selection for each considered value of weight), we select between 28 (23) and 41 (31) features for the prediction of DNA (RNA)-binding residues. Using the weight = 1.8 (3.6) as an example, Figure 5.2 shows the improvement of AULC by gradually (one by one) adding the 41 (31) selected features into the feature subset along the feature selection process. We observe a steady increase in the predictive performance as additional features are added into the set of selected features.

Figure 5.3 compares the predictive quality of the logistic regression models trained by using different weight values and the corresponding selected feature subsets. The predictive quality is measured by AULRC on the training dataset based on the 5-fold cross validation; this measure quantifies the amount of cross prediction between RNA and DNA binding residues. We select the weight value of 1.8 (3.6) with the corresponding subset of 41 (31) features that secures the best predictive quality (lowest AULRC). These parameters are utilized to implement the predictor of the DNA (RNA)-binding residues. We combine the selected features with the 30 features that compose the evolutionary profile. This leads to an additional improvement in predictive quality, as shown in Figure 5.2. We input the resulting 71 (61) features into logistic regression to build the two prediction models, one for the prediction of DNA-binding and one for RNA-binding.

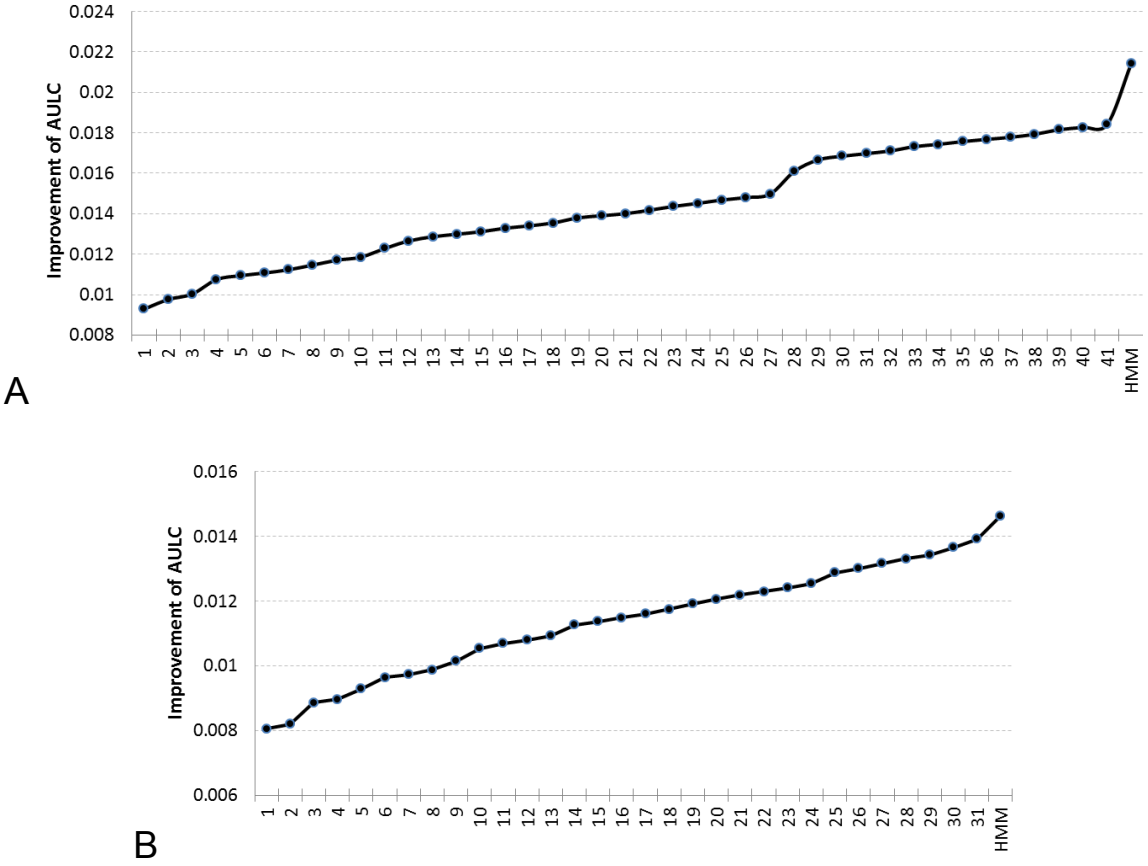


Figure 5.2. Improvement in the value of AULC through the feature selection based on 5-fold cross validation on the training dataset. Panel A is for the prediction of DNA-binding residues with the weight value = 1.8. Panel B is for the prediction of RNA-binding residues with the weight value = 3.6.

X-axis is the number of features added through the best first search in the feature selection. Last index on the x-axis 'HMM' represents addition of the entire HMM profile that includes 30 features.

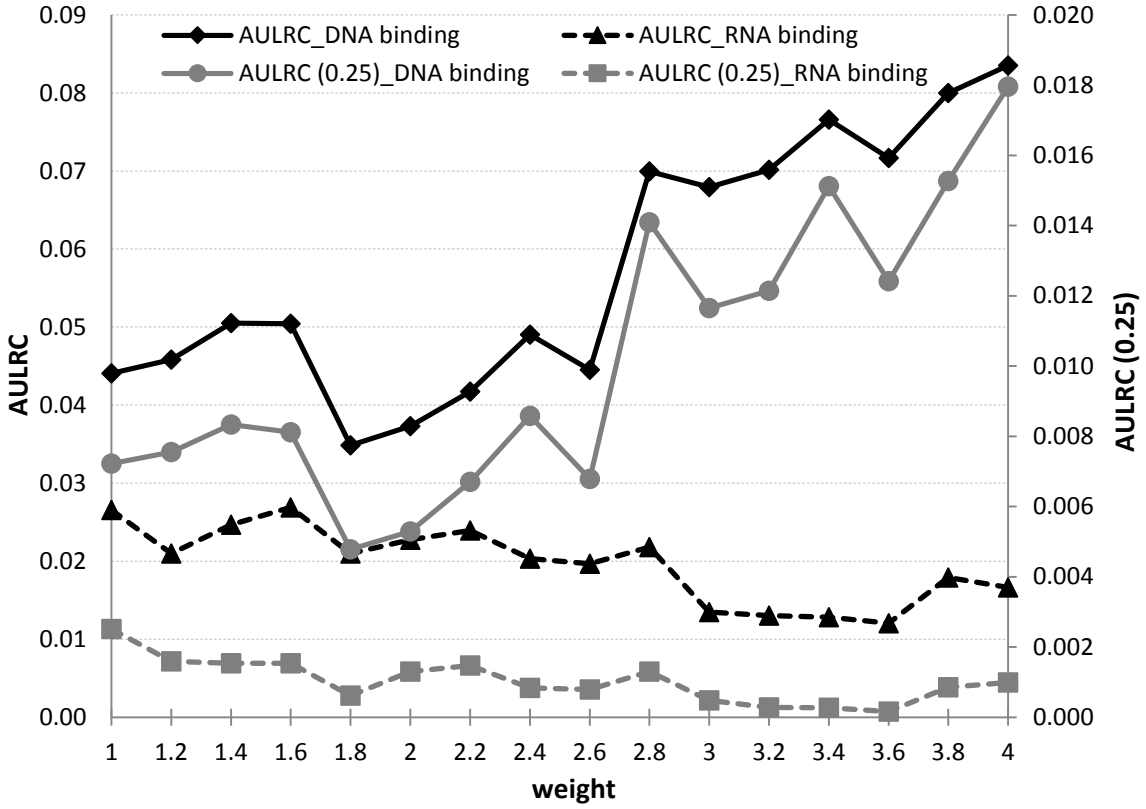


Figure 5.3. Predictive performance measured by AULRC on the training dataset based on 5-fold cross validation for the models that use different weights.

AULRC (0.25) is equivalent to AULRC but using a smaller cutoff on TPR at 0.25. Lines represent the results for the prediction of DNA-binding residues, and dotted lines indicate the results for the prediction of RNA-binding residues.

Design of the predictive models in the second layer

The binding residues predicted by the models from the first layer cluster together in the sequence. We observe that the predictions for the correct type of nucleic acid binding have higher density than the predictions for the other type of binding, i.e., if the currently predicted protein binds DNA, usually the number of predicted DNA binding residues is higher than the number of predicted RNA binding residues. Consequently, if we consider the predictions for the DNA and RNA binding residues in the same protein together, they can be re-predicted to improve predictive performance. In particular, we can reduce the amount of cross-predictions between DNA and RNA binding residues, i.e., we can reduce the number of predicted RNA binding residues based on the high number of predicted DNA binding residues and vice versa. Motivated by this, we design predictive models in

a second layer (meta-predictor) that re-predict the outputs of the two models from the first layer.

We first generate a set of per residue features and aggregated features using a sliding window from the predictions of the two models from the first layer. For the per residue features, we set the window size to 3 to include the predictions of the DNA-binding and RNA-binding for the current residue and its two adjacent neighbors. We also calculate the aggregated information over the window including the content of predicted DNA-binding and RNA-binding residues, average and standard deviation of the predicted propensities for DNA-binding and RNA-binding over the window. We use windows with varying sizes between 3 and 21 with a step of 2. We also calculate the same aggregated values for the whole sequence. This totals to 122 features.

We then select a subset of predictive and non-redundant features using the same feature selection procedure as for the first layer. We first rank features based on the predictive quality (measured by AULC) of the corresponding univariate logistic regression models on the 5-fold cross validation on the training set. Then starting from the top ranked features, we accept the next ranked feature into our feature set if the AULC value is not worse by more than 0.0001 compared to the prediction obtained with the model from the first layer, and the AULRC is better by (drops by) at least 0.001 when compared to the prediction using feature set before the addition. As a result, we select 3 (3) features for the prediction of DNA (RNA)-binding residues. Each of these two sets of three features is input into the corresponding logistic regression model to generate the final predictions.

To sum up, the design of our model implements three novel approaches to reduce the mis-prediction between the two types of nucleic acid binding residues: (1) we use the training dataset that includes both DNA-binding and RNA-binding proteins to train the model (the existing methods were developed using datasets that include only DNA binding or only RNA binding proteins); (2) we use weights >1 for the residues that could be cross predicted; and (3) we introduce the second predictive layer.

5.3 Results and discussion

We test and compare various designs of our predictive model on the test dataset to demonstrate that the novel strategies introduced to design DRNApred help to improve its predictive quality. Next, we comprehensively compare results generated by DRNApred with the results of the existing methods for the prediction of DNA (RNA)-binding residues. As part of this comparison, we compare predictive quality of DRNApred on the test dataset with the predictive quality of the existing methods. We also analyze the binding residues predicted by DRNApred and compare them with those generated by the existing methods. Moreover, we apply and compare DRNApred and the other considered predictors of DNA (RNA)-binding residues on the test dataset for the prediction of the DNA (RNA)-binding proteins. Lastly, we estimate and compare the runtime of our method and with the runtime of the other methods.

5.3.1 Improvement in predictive performance due to the use of novel design features

We have introduced three novel strategies in the design of our model to reduce the cross prediction. We compare the results obtained by our predictive model with the results obtained when designing the model without the use of these strategies to quantify their impact on the predictive performance. We consider the following four scenarios: (1) the model developed on the training dataset with just one target type of nucleic acid binding proteins (referred as *only DNA (RNA) binding data*); (2) the model trained on the combined dataset of both DNA-binding and RNA-binding proteins (referred as *combined data*); (3) the model designed on the combined dataset and using the weights to minimize the cross predictions (referred as *combined data with penalty*); and (4) the complete model implemented using 2 layers based on the combined dataset and weight (referred as *second layer*). The latter is utilized to implement our predictor of the DNA (RNA) binding residues.

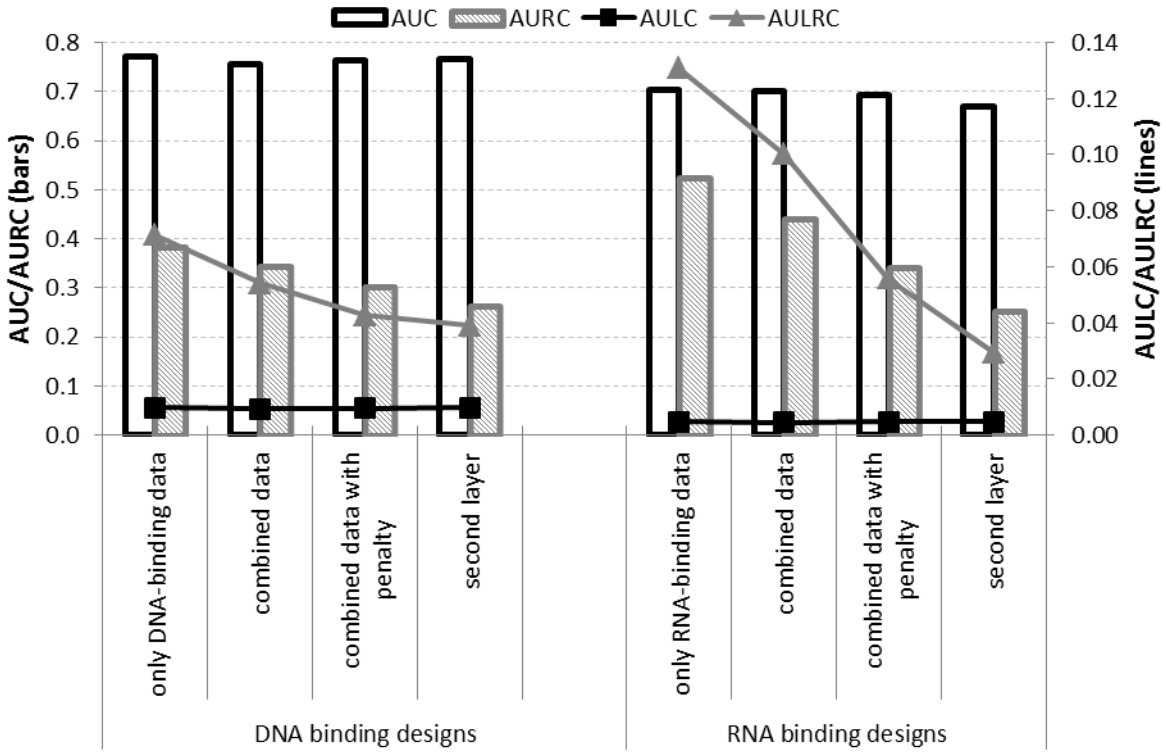


Figure 5.4. Comparison of predictive performance using different designs of the models for the prediction of DNA-binding (RNA-binding) residues on the test dataset.

Bars for the AUC and AURC are associated with the y-axis on the left side while lines for the AULC and AULRC are quantified with the scale on the y-axis on the right. The 'only DNA-binding data' ('only RNA-binding data') scenario is the model for the prediction of DNA-binding (RNA-binding) residues designed on the dataset with just DNA-binding (RNA-binding) proteins; the 'combined data' scenario is for the model built on the combined dataset with both DNA-binding and RNA-binding proteins; the 'combine data with penalty' scenario is for the model that uses combined dataset and weights that are used to penalize the cross predictions; the 'second layer' scenario considers model that extends the 'combine data with penalty' scenario with the second layer.

We evaluate the predictive performance for these four scenarios on the test dataset. The results are shown in Figure 5.4. For the DNA-binding models, the predictive quality measured by AUC and AULC is very similar across the four scenarios, while the cross predictions quantified by AURC and AULRC decrease dramatically as we improve our design by adding additional strategies. Learning from the dataset with both DNA-binding and RNA-binding proteins (the 'combined data' scenario) the model for prediction of DNA-binding residues substantially improves over the 'only DNA-binding data' scenario. This improvement involves reducing the cross prediction measured with AURC (AULRC) by 10% (25%), while maintaining similar overall predictive quality measured with AUC and AULC. The model based on the 'combined data with penalty' scenario

that uses weights further decreases AURC and AULRC by 12% and 20%, respectively, while maintaining similar AUC and AULC. The last 'second layer' scenario provides the best predictive performance by achieving similar AUC and AULC, and decreasing the cross prediction, as evidenced by lower values of AURC and AULRC, when compared to the 'combined data with penalty' scenario. The same observations are true for the models that predict RNA-binding residues. The model based on the 'second layer' scenario maintains the overall predictive quality measured with AUC and AULC and substantially reduces the cross prediction measured with AURC and AULRC when compared to the other three scenarios. The two models that use all four strategies are utilized to build the DRNApred method for the prediction of the DNA-binding and RNA-binding residues.

5.3.2 Predictive performance for the prediction of the DNA/RNA binding residues

We test the DRNApred method on the prediction of DNA-binding residues and the prediction of RNA-binding residues on the test dataset. We compare these results with the results generated by existing methods that predict DNA or RNA binding residues. We include the methods that were assessed in a recent relevant comparative review [77], where one of the main selection criteria was availability of webservers and short runtime. We include 5 methods for the prediction of DNA-binding residues and 3 methods for the prediction of RNA-binding residues. Results are shown in Table 5.2.

DRNApred evaluated on the prediction of DNA binding residues is shown to secure comparable overall predictive quality quantified with AUC and AULC values when compared to the other predictors of DNA binding residues. AULC is the area under the ROC curve where FPR has low values <5.4%. The 5.4% is the fraction of positives in the test dataset and the corresponding part of the curve covers predictions where the numbers of false positives (incorrectly predicted binding residues) is smaller than the number of positives (native binding residues). In other words, this is where the predictor does not over predict the binding residues (see detailed definition in section 2.4.4). The AUC value of DRNApred is lower than that of the best method BindN+ and comparable to the other considered methods. However, DRNApred's AULC value is the highest and significantly better than the AULC values of all other methods. The corresponding ROC

curves are shown in Figure 5.5A. An insert in the bottom right corner of this Figure focuses on a part of the curve that is used to compute AULC where the FPR<5.4%. The complete ROC curves of the six predictors of the DNA-binding residues are relatively similar. The curves of BindN+ and DBS-PSSM are better than the curve of DRNApred when FPR values are high and worse for the arguably more practical range with the lower values of FPR<5.4% (see the insert). Importantly, the TPR of DRNApred is about 6 times higher than its FPR at FPR = 5.4%, and close to 30% of the native DNA-binding residues can be found at this low FPR. This means that DRNApred correctly locates a large fraction of native binding residues when mis-predicting a relatively low fraction of the native non-binding residues. Specifically, at FPR = 5.4%, the number of TP = 290 and the number of FP = 968. Although the number of FPs is three times higher than the number of TPs, this rate is much lower compared to an expected rate that equals 19 (there are 19 time more non-binding residues than the number of binding residues). Moreover, although the number of FPs that we predicted at FPR = 5.4% is high, some of them could potentially correspond to binding residues. This is because the annotation of binding residues in our test dataset is incomplete and because some of the FPs are close (in the sequence) to the native binding residues. In the latter case they are likely to be TPs given the fact that we define binding residues using a somehow arbitrary threshold (more details in Section 5.3.3). We binarize the propensities generated by the considered methods to classify each residue as binding (propensity > threshold) and non-binding (propensity ≤ threshold). The threshold is determined to ensure that the number of predicted binding residues equals to the number of native binding residues in the test dataset. These binarized predictions are assessed with sensitivity and MCC; specificity is virtually identical for different methods given how the threshold was selected. We observe that DRNApred offers slightly higher sensitivity and comparable MCC when compared to the other considered predictors of DNA binding residues.

Although the overall predictive performance for the prediction of the DNA binding residues of DRNApred is similar to the other methods, our predictor significantly reduces the cross prediction between DNA and RNA binding residues. This is measured with AURC (area under the ratio curve) and AULRC (area under the lower range of the ratio curve where TPR<0.5) values. DRNApred obtains the lowest AURC and AULRC values

which are lower by (0.35-0.26)/0.35 = 25.7% and (0.069-0.039)/0.069 = 43.5%, respectively, compared to the second best BindN+ predictor. Figure 5.6A which plot of the values of ratio against TPR, further validates this conclusion. It shows that DRNApred is substantially better than the other methods (achieves the lowest ratio) over the entire range of TPR values. Comparison of the ratio based on the binary predictions also shows that our method significantly reduces the cross predictions. It obtains the lowest ratio value which is lower by (0.13-0.06)/0.13 = 53.8% compared to the second best BindN+.

Table 5.2. Comparison of the predictive performance of DRNApred with the other methods for the prediction of the DNA- (RNA-) binding residues on the test dataset.

Sensitivity, MCC and ratio are calculated from the binary predictions which are converted from the probability prediction using threshold that sets the number of predicted binding residues to be equal to the number of native binding residues in the test dataset. Significance of the difference in MCC, ratio, AULC and AULRC values between the best performing method and other methods was assessed based on 10 repetitions that utilize 70% of randomly chosen from the test dataset proteins; + (=) in the Sig column denotes that the difference was (was not) significant at *p*-value <0.05. Methods are sorted by their AULRC value.

	Methods	**Sensitivity**	**MCC**	**Sig**	**ratio**	**Sig**	**AUC**	**AULC**	**Sig**	**AURC**	**AULRC**	**Sig**
DNA binding	DRNApred	0.25	0.21		0.06		0.77	0.010		0.26	0.039	
	BindN+	0.22	0.18	+	0.13	+	0.79	0.008	+	0.35	0.069	+
	DP-Bind(svm)	0.24	0.20	=	0.14	+	0.75	0.009	+	0.43	0.087	+
	DP-Bind(klr)	0.24	0.20	=	0.15	+	0.76	0.009	+	0.43	0.087	+
	DP-Bind(plr)	0.22	0.18	+	0.16	+	0.74	0.008	+	0.44	0.093	+
	DBS-PSSM	0.21	0.17	+	0.18	+	0.77	0.008	+	0.41	0.095	+
RNA binding	DRNApred	0.16	0.12		0.02		0.67	0.005		0.25	0.029	
	Pprint	0.15	0.11	=	0.1	+	0.66	0.005	=	0.51	0.121	+
	RNABindR	0.14	0.10	+	0.16	+	0.73	0.004	+	0.51	0.135	+
	BindN+	0.12	0.08	+	0.2	+	0.67	0.003	+	0.63	0.195	+

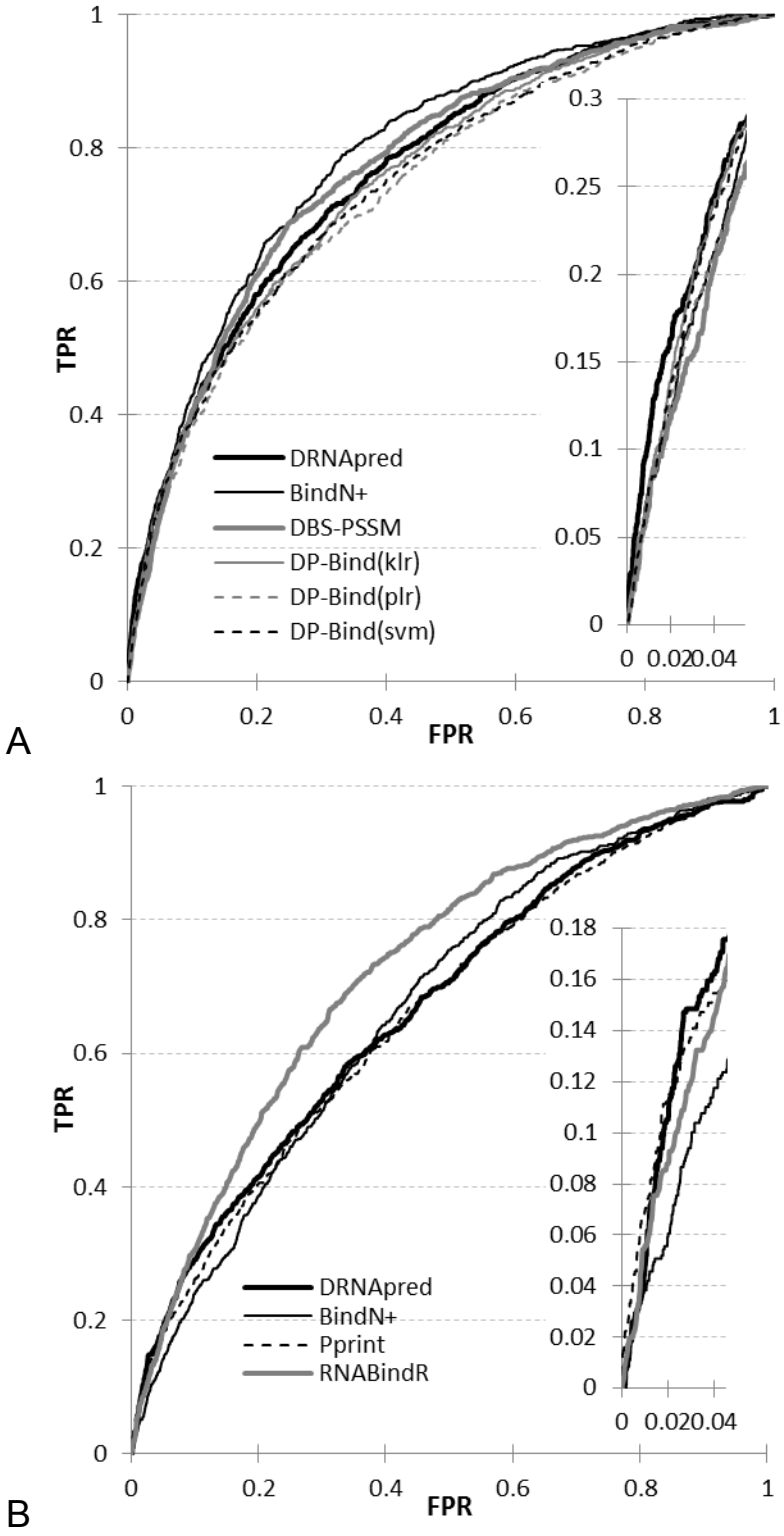


Figure 5.5. Comparison of ROCs of DRNApred and the other considered predictors of the DNA and RNA binding residues on the test dataset.

The insert in the bottom right corner focuses on the ROC curve where FPR<5.4% for the DNA binding (<4.5% for the RNA binding). AULC is calculated as the area under that part of the ROC curve. Panel A is for the prediction of DNA-binding residues, and panel B for the prediction of RNA-binding residues.

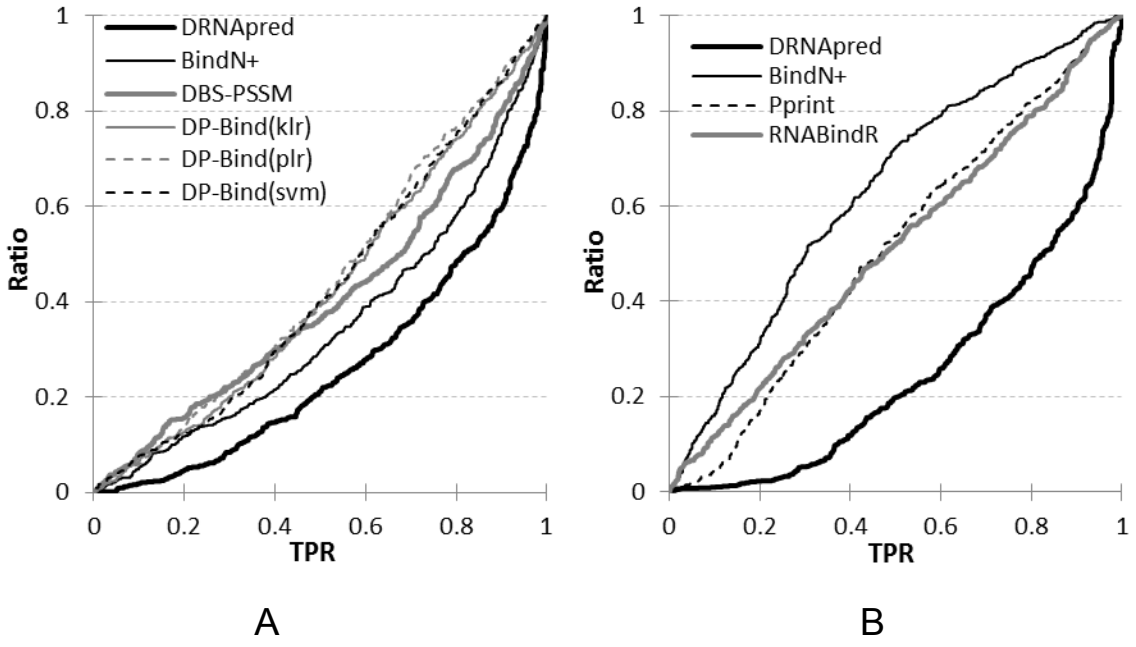


Figure 5.6. Comparison of the ratio curves for DRNApred and the considered predictors of the DNA and RNA binding residues on the test dataset.

The ratio curve is the plot of Ratio against TPR values. The curve that is closer to the *x*-axis for the same TPR corresponds to better predictions, i.e., lower amount of cross-predictions between RNA and DNA binding residues. Panel A is for the prediction of DNA-binding residues, and Panel B for the prediction of RNA-binding residues.

Similar observations are true for DRNApred for the prediction of RNA-binding residues (Table 5.2). When compared with the other predictors, DRNApred offers comparable overall predictive quality measured with AUC, AULC values and more importantly, significantly lower amounts of cross predictions that are quantified with AURC and AULRC values. The DRNApred's AURC and AULRC are substantially lower by (0.51-0.25)/0.51 = 51% and (0.121-0.029)/0.121 = 76%, respectively, compared to the second best Pprint methods. ROC curves in Figure 5.5B shows that RNABindR is better than other predictors when FPR is relatively high, but it is outperformed by DRNApred when FPR<4.5%, i.e., when the number of false positives (incorrectly predicted non-binding residues) is lower than the number of native positives (native RNA binding residues). Ratio curve in Figure 5.6B further confirms the conclusion that our method significantly reduces the cross-prediction and demonstrates that this is true over the whole range of the TPR values. Comparison of the binary predictions when setting all methods to generate similar FPRs (number of predicted positives is set to equal the number of positives) reveals that DRNApred provides slightly higher sensitivity and

MCC, and much smaller ratio. The ratio of our predictor is lower by (0.1-0.02)/0.1 = 80% when compared to the second best Pprint method.

To sum up, DRNApred substantially reduces the cross predictions between DNA-binding and RNA-binding residues while maintaining similar overall predictive quality when compared to the existing methods. We explain significance of this result in Section 5.3.3. Moreover, the new predictor correctly predicts the largest number of DNA-binding or RNA-binding residues when the number of predicted binding residues is reasonably low and no larger than the number of native binding residues.

We also assess the considered methods on the negative dataset composed on proteins that are unlikely to bind nucleic acids. Since there are no positive data (binding residues) in this negative dataset, we quantify the predictive quality with the FPR. The results reveal that all methods obtain comparable and low FPR values that range between 2 to 5% (2 to 4%) for the prediction of DNA (RNA)-binding residues. Among the predictors of the DNA-binding residues BindN+, DP-Bind_S, DP-Bind_K, and DBS-PSSM secure FPR = 3%, DP-Bind_P has FPR = 4%, and DRNApred obtains FPR = 5%. Considering the predictors of the RNA-binding residues, DRNApred and BindN+ generate predictions characterized by FPR = 2%, while Pprint has FPR = 4%.

5.3.3 Analysis of the predicted binding residues

We observe that native RNA and DNA binding residues tend to cluster together in the protein sequence. This is because close proximity in the sequence implies proximity in the corresponding structure and regions on the protein surface that interact with the nucleic acids tend to be relatively large given the large size of the RNA and DNA molecules. Moreover, the annotation of the binding residues suffers inaccuracies given how they are defined. The use of a distance between atoms in protein and nucleic acids results in somehow arbitrary inclusion or exclusion of binding residues that are close to the cut-off value used to define binding. This means that some of the non-binding residues adjacent to the annotated binding residues could be in fact involved in binding. Altogether, these observation points to a conjecture that residues that are in close proximity in the sequence to the annotated binding residues are more likely to in fact bind

DNA/RNA compared to residues that are far away. In other words, false positives localized close to the native binding residues are more desirable (more likely to be true positives) compared to the false positives that are far away from the binding residues.

We analyze the binding residues predicted by different methods to compare how close they are from the native binding residues. For each method, we count the number of correctly predicted binding residues (there residues have distance of 0 from the native binding residue), and incorrectly predicted binding residues that are ≤1, ≤2, … residues away from the nearest native binding residue. The corresponding fractions of these predicted binding residues out of the total number of the predicted binding residues are plotted in Figure 5.7. The total number of the predicted binding residues for each method is set to be the same and equal to the number of the native binding residues. The fraction of the putative binding residues predicted in the incorrect type of binding proteins can be read from the gap between the value of 1 and the value of the fraction at the end of a given curve. For example, Figure 5.7A shows that about 20% and 35% of the DNA binding residues identified by DRNApred and BindN+ were predicted in the RNA binding proteins, respectively. We argue that DRNApred predicts higher quality false positives compared to the other considered methods since they are localized closer to native binding residues. This is true for the prediction of both DNA and RNA binding residues. Figure 5.7 reveals that 31% (16%) of the putative DNA (RNA)-binding residues generated by DRNApred are correctly predicted and 51% (46%) are close to the nearest native binding residues (≤5 residues away). To compare, 24% (15%) of the predicted DNA (RNA)-binding residues are correctly identified by the best existing predictor and 44% (33%) are close to the native binding residues. The observation that DRNApred correctly predicts more binding residues is in consistent with its higher MCC and sensitivity (Table 5.2). Moreover, as the distance increases the DRNApred's curve saturates faster and reaches a much higher value compared to the curves from the other methods. This means that our model cross predicts much less than the other methods. Specifically, analysis of the far right end of the curves in Figure 5.7 demonstrates that DRNApred mis-predicts 20% (18%) of DNA (RNA)-binding residues in the RNA (DNA)-binding proteins, compared to the 35% (44%) by the second best BindN+ (Pprint) methods. Overall, DRNApred correctly finds more binding residues and captures more

putative binding residues that are likely to bind to DNA (RNA) although they lack such annotation in the test dataset. Importantly, our model generates much fewer heavy mis-predictions that are defined as the putative RNA binding residues identified in the DNA binding proteins and the putative DNA binding residues found in the RNA binding proteins.

We also evaluate how predictive quality measured with MCC and TPR would change if the predicted binding residues which are 0, ≤1, ≤2, and ≤3 residues away from the nearest native binding residue would be considered as correctly predicted, see Figure 5.8. We argue that the corresponding false positives that we re-consider as true positives could be in fact interacting with the nucleic acids or be useful to identify the nearby binding residues. As expected, both MCC and TPR for all considered methods improve as we include additional true positives. Interestingly, inclusion of just the adjacent positions (distance = 1 on the *x*-axis) results in a substantial increase in TPR of DRNApred by about 8% for both DNA and RNA binding, given the TPR is 25% for DNA and 16% for RNA at the distance = 0 (only the native binding residues are considered). The MCC also registers a very large increase from 0.21 to 0.31 for the DNA binding and from 0.12 to 0.22 for RNA binding. At distance = 3, our method achieves the TPR = 0.38 (0.31) and MCC = 0.39 (0.31) for the prediction of the DNA (RNA)-binding residues. Moreover, DRNApred secures the largest increases in both MCC and TPR when compared to the other methods. This again demonstrates that our predictor is better at finding desirable, high quality false positives that could be in fact relevant to the nucleic acid binding.

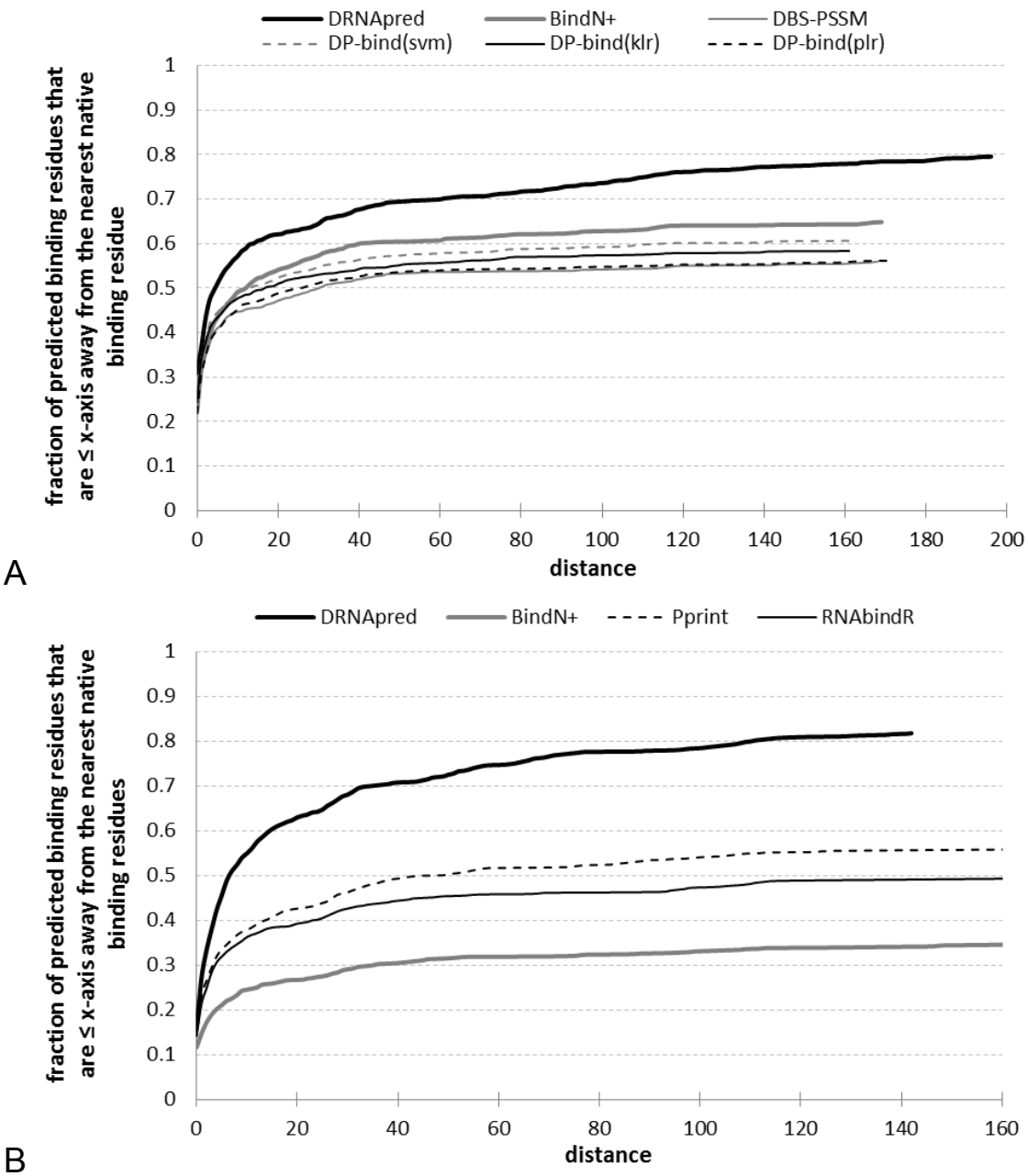


Figure 5.7. Summary of the distance measured by the number of residues in the sequence between the predicted binding residues and the nearest native binding residues.

The summary is quantified with fractions of binding residues that are ≤ a given distance, shown on the x-axis, away from the nearest native binding residue. The fraction is defined as the count of residues up to a given distance away divided by the total number of the putative binding residues. The curves do not reach the fraction of 1 because the remaining residues are predicted in proteins that do not have the corresponding native binding residues (the distance to the nearest native binging residue is undefined). These are putative RNA binding residues that are predicted in the DNA binding proteins and vice versa. Panel A summarizes results for the prediction of the DNA-binding residues and panel B for the prediction of the RNA-binding residues.

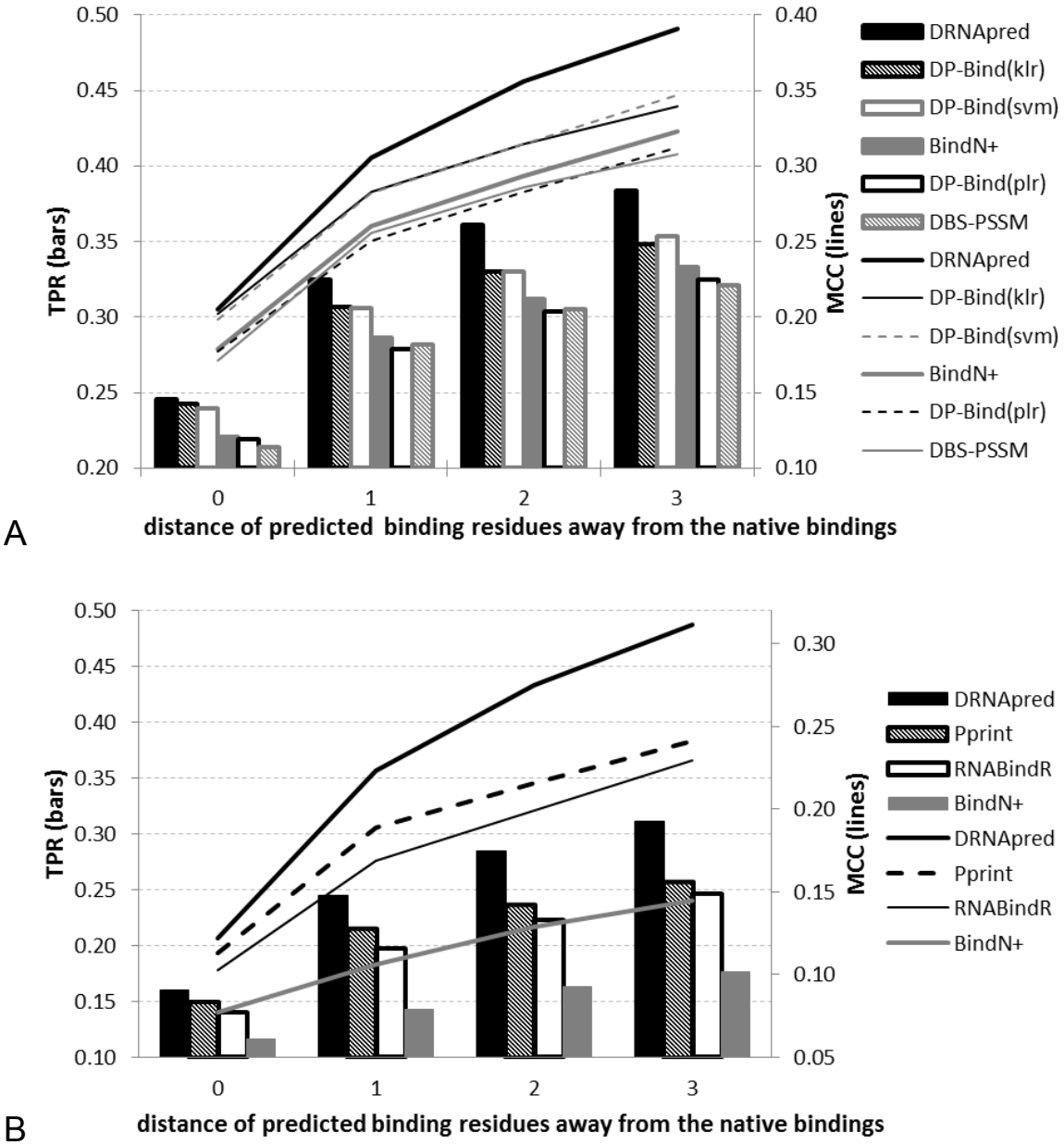


Figure 5.8. Comparison of MCC and TPR values for DRNApred and other considered predictors of DNA and RNA binding residues when reconsidering putative binding residues that are close to native binding residues as true positives. The predicted binding residues that are no farther than 0, 1, 2, and 3 positions (*x*-axis) in the sequence from the closest native binding residue are considered as correct predictions.

TPR values are shown using bars and the y-axis on the left. MCC values are shown using lines and the y-axis on the right. Panel A is for the predictors of the DNA-binding residues while Panel B is for the predictors of the RNA-binding residues.

5.3.4 Predictive performance for the prediction of the DNA/RNA-binding proteins

We test performance of DRNApred and the other predictors of the DNA and RNA-binding residues for the predictions of the DNA and RNA-binding proteins on the test dataset. In contrast to the residue-level predictions, in this case we assess whether a given protein is correctly identified as binding to a specific nucleic acid. A protein is annotated as binding to DNA (RNA) if at least one residue in this protein is annotated as binding to DNA (RNA). A protein is assumed to be predicted as binding to DNA (RNA) if the number of predicted DNA (RNA)-binding residues in this protein is larger than a small threshold. This is to accommodate for the predicted false positives. The threshold is set so the FPR of a given method on the test set equals 5%. Table 5.3 summarizes the results. DRNApred outperforms other methods for the prediction of both DNA and RNA-binding proteins by a wide margin. DRNApred's MCC is statistically significantly better than MCCs of the other methods. The TPR of our predictor is 5 and 6 times higher than FPR for the DNA and RNA binding, respectively, and is also much higher than the TPR values of the other predictors.

Table 5.3. Comparison of predictive performance of DRNApred and the other considered methods for the prediction of DNA and RNA-binding proteins on the test dataset.

Binding type	Methods	TPR	MCC	Sig	AUC	Sig
DNA	DRNApred	0.27	0.26		0.68	
	DP-Bind(svm)	0.06	0.00	+	0.54	+
	DP-Bind(klr)	0.04	-0.05	+	0.53	+
	BindN+	0.12	0.10	+	0.52	+
	DP-Bind(plr)	0.02	-0.10	+	0.45	+
	DBS-PSSM	0.00	-0.19	+	0.44	+
RNA	DRNApred	0.30	0.32		0.65	
	Pprint	0.24	0.26	+	0.63	=
	RNABindR	0.12	0.11	+	0.59	+
	BindN+	0.00	-0.16	+	0.45	+

Besides evaluating the predictions at the low FPR, we vary the threshold (the minimal number of the predicted binding residues that corresponds to prediction of a binding protein) using the complete range. We plot relation between the corresponding TPR and

FPR values (ROC curve) in Figure 5.9. The plot shows that DRNApred improves over the other methods for small and modest values of FPR. Predictions when TPR values are high are arguably less interesting since they would lead to a substantial overprediction of the DNA or RNA binding proteins. These results are in agreement with the fact that our predictor secures the highest AUC values (Table 5.3). By tuning the threshold, DRNApred achieves maximal MCC = 0.31 and 0.36 for the prediction of the DNA and RNA-binding proteins, respectively, compared to the second best method DP-Bind_S with MCC = 0.23 and RNABindR with MCC = 0.28. The main reason why the other methods lack in predictive quality is that they cross-predict between DNA and RNA binding residues. In other words, their correct predictions of DNA binding proteins are coupled with the incorrect predictions of RNA binding proteins as DNA binding, resulting in high FPRs and low AUC and MCC values.

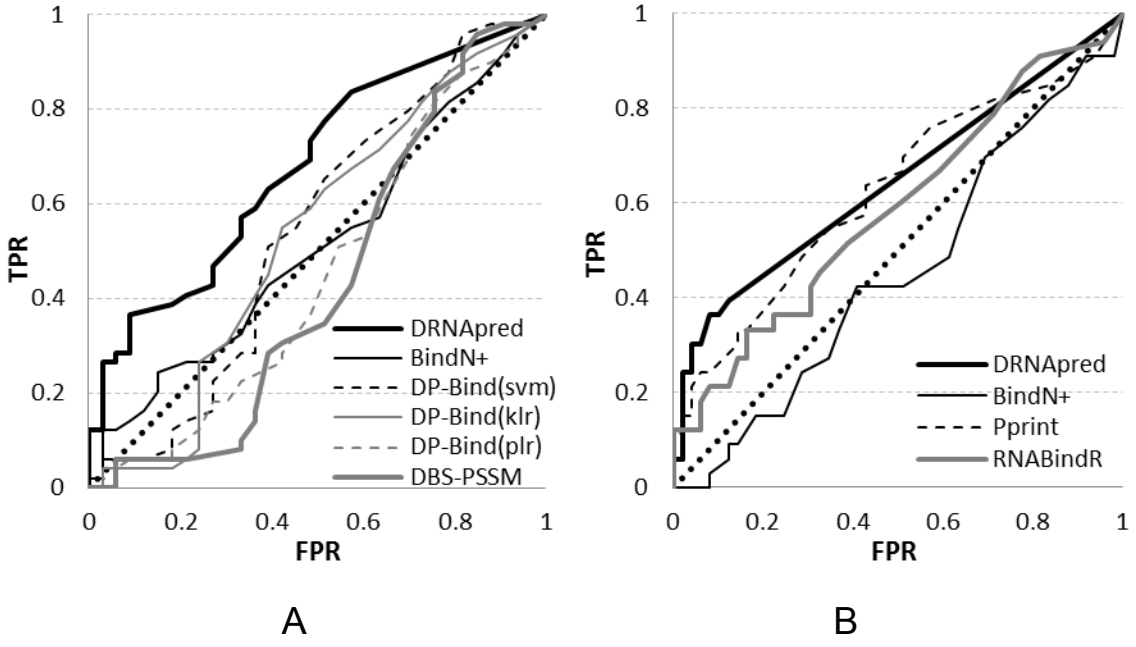


Figure 5.9. Comparison of ROCs for DRNApred and the other predictors for the prediction of DNA and RNA-binding proteins on the test dataset.

Panel A is for the prediction of the DNA-binding proteins and Panel B is for the prediction of the RNA-binding proteins. The dotted black diagonal line represents a random prediction.

5.3.5 Comparative evaluation of runtime

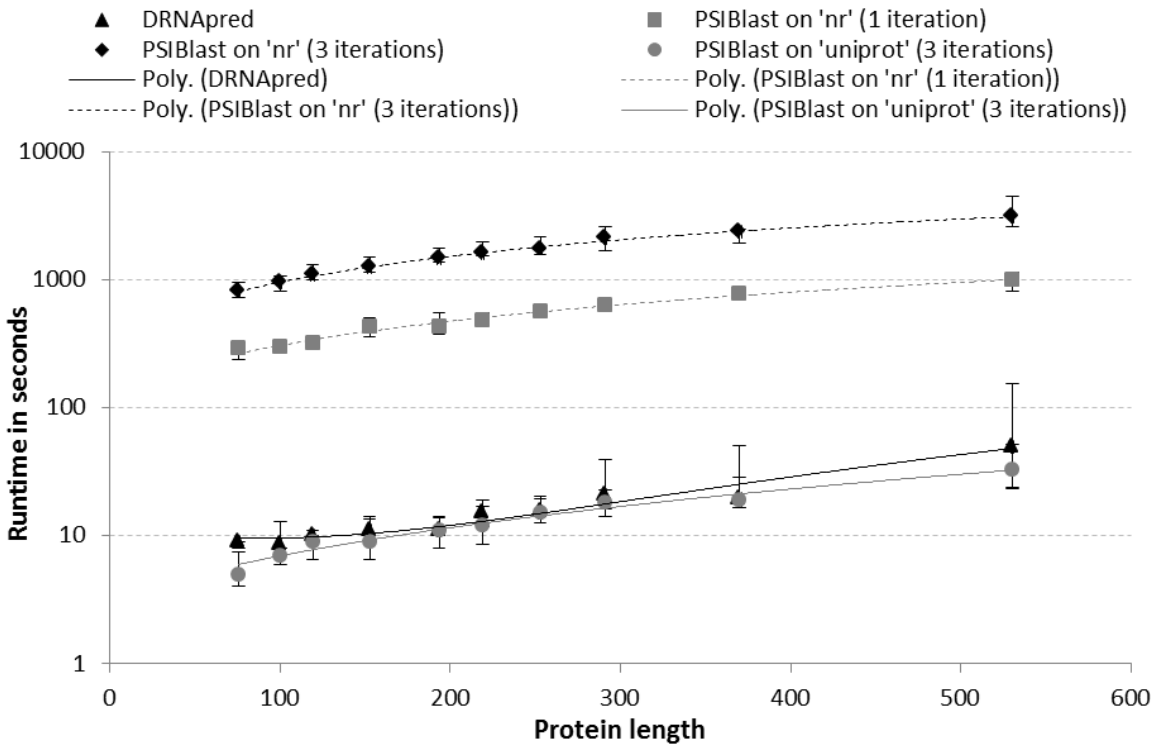


Figure 5.10. Comparison of runtime in the function of protein length for DRNApred and the other predictors of the DNA and RNA binding residues on the test dataset.

The y-axis is the runtime in seconds shown using base 10 logarithmic scale. The x-axis is the protein length. We sorted proteins by their sequence length and divided them into 10 equally sized sets that include proteins with increasing size. The plot reports the median runtime (markers) and the 25th and 75th centiles (error bars) against the median protein length for each of the 10 protein sets. The measurements were made using a modern desktop computer with i7-CPU and 23GB RAM. Lines show polynomial fit into the measured data.

Besides predictive quality runtime is a key factor that determines whether a given predictor can be applied in a high-throughput manner to annotate a large collection of proteins, such as a complete proteome. The considered predictors, except for DRNApred, utilize PSI-BLAST to derive evolutionary profile that they use as one of their inputs. The calculation of the profile is the main computational cost of these methods. We approximate lower bound of their runtime by the time to run PSI-BLAST. Based on the database and the number of iterations that each of these methods used to run PSI-BLAST, we divide them into three groups: methods that use the 'nr' database with 1 iteration (referred to as *PSIBlast on 'nr' (1 iteraction)*), the 'nr' database with 3 iterations (referred to as *PSIBlast on 'nr' (3 iterations)*), and 'uniprot' database with 3 iterations (referred to

as *PSIBlast on 'uniprot' (3 iterations)*). Figure 5.10 compares the runtime of DRNApred and the other methods based on predictions on the test dataset. Although the absolute value of the runtime depends on computer hardware used, we focus on relative differences which are hardware independent. DRNApred is at least 3 orders of magnitude faster than the other methods that utilize PSI_BLAST run with the 'nr' database. Our runtime is comparable to the runtime of BindN+ which utilizes PSI-BLAST on a much smaller UniProt database. Both methods predict an average size protein in about 10 seconds using a modern desktop computer.

We fit the measured runtime using polynomials for DRNApred, BindN+ and the other methods that use PSI_BLAST with the 'nr' database (see lines in Figure 5.10). We use these polynomials to estimate the total runtime to predict the complete human proteome that has 69178 proteins; this is the largest proteome among all species. The DRNApred is estimated to take about 48 days, BindN+ 21 days, while the other methods take substantially more time, with 531 to 1475 days which translates into 1.5 to 4 years. These estimates are based on computations with a single processor (i7-CPU and 23GB RAM). We note that our estimate is similar to the actual runtime of 55 days that we measured when using DRNApred to predict the human proteome. We performed calculations using 8 processors by processing a different subset of proteins on each processor, which cut down the time to 7 days. These results suggest that DRNApred is sufficiently fast to perform genome-wide predictions using a desktop computer.

In the nutshell, the runtime of DRNApred is relatively low and comparable to the fastest current method, allowing for genome-wide predictions, while our predictor offers substantially better predictive performance.

5.4 Conclusions

Although many methods for the prediction of the DNA and RNA-binding residues from the protein sequence have been published, their weakness is that they cross-predict a substantial number of the nucleic acid binding residues (DNA-binding residues are predicted as RNA-binding as vice versa) and require relatively high runtime. Motivated by this, we developed a new high-throughput (runtime-efficient) method that accurately

and specifically predicts only DNA-binding and only RNA-binding residues. We designed the DRNApred method by considering a comprehensive set of features extracted from diverse sequence-derived information including amino acid type, physiochemical properties of amino acids, evolutionary profiles, and putative intrinsic disorder, secondary structure, solvent accessibilities using a dataset with both DNA-binding and RNA-binding proteins. We implemented a weight-based mechanism to penalize cross-predictions, performed empirical selection of a subset of predictive and non-redundant features, and used logistic regression algorithm to produce the predictive model. To further reduce the cross-predictions, our method uses a two-layer design where initial predictions generated by the regression in the first layer are fed into another regression-based model in the second layer that re-predicts the DNA-binding and RNA-binding residues. We empirically demonstrate that the three novel design ideas (use of the combined dataset with RNA and DNA binding proteins, use of penalties, and use of the second layer) contribute to improving the predictive quality by reducing the amount of cross predictions. We comparatively tested DRNApred on the test dataset for the prediction of the DNA and RNA-binding residues and proteins. We show that DRNApred substantially reduces the cross predictions (measured with AURC and AULRC) for the prediction of the binding residues while maintaining similar overall predictive quality (measured with AUC and AULC) when compared to the existing methods. Importantly, empirical analysis reveals that our predictor finds arguably higher quality false positives that are located nearby the native binding residues. It also predicts substantially fewer DNA binding residues in the RNA binding proteins and vice versa when compared with the other considered predictors. We also compared predictive performance for the prediction of the DNA and RNA-binding proteins. We show that DRNApred secures the highest AUCs and outperforms the other methods by correctly predicting more DNA and RNA-binding proteins at the same false positive rate. Moreover, empirical tests demonstrate that DRNApred is computationally efficient, at least 3 orders of magnitude faster than majority of the other methods, excluding BindN+. We show that DRNApred and BindN+ have similar runtime profiles, that both can be used to perform genome-wide predictions on a desktop computer, while DRNApred provides better predictive performance and lowest levels of cross predictions.

Chapter 6

Goal 4: Identification of known and novel DNA- and RNA-binding residues/proteins on proteomic-scale

A significant amount of effort has been made to annotate the DNA and RNA binding proteins in the human proteome. Several databases with the annotated DNA and RNA binding proteins have been developed, such as the RBPDB database of RNA-binding proteins [92], animalTFDB of DNA-binding transcription factors [93, 94], and the UniProt [95] database that includes annotations of nucleic acids binding via the gene ontology (GO) terms. These databases annotate 4.7% and 1.8% of the human proteins as DNA-binding and RNA-binding, respectively. These fractions are low compared to the estimated number of the nucleic acid binding proteins. For instance, the fraction of transcription factors alone (a subset of the DNA-binding proteins that transcribe DNA into RNA) in human was estimated to be 7.9% [94]. Similarly, the number of RNA-binding proteins was recently estimated to be at least 7.5% [96]. We take advantage of the runtime efficiency of our method and apply it to perform prediction of the nucleic acid binding proteins and residues on the entire human proteome to facilitate finding of the still missing DNA and RNA binding proteins and to find out how well our method predicts the already known DNA and RNA binding proteins.

We apply the time-efficient DRNApred method to perform large-scale predictions on the complete human proteome that includes about 70000 proteins. We assess its predictive performance by measuring whether it specifically predicts only the target type of binding proteins/residues in the known binding proteins from the human proteome,

e.g. whether its predictions of the DNA-binding residues target primarily the known DNA-binding proteins and how many of these predictions are in the known RNA-binding proteins, and vice versa. We compare these results with the predictions generated with BindN+ to validate whether DRNApred outperforms BindN+ (as we have shown that using a test dataset in Chapter 5) by reducing the cross predictions. We also assess whether novel binding proteins that are predicted with DRNApred (predicted binding proteins that not overlap with the known binding proteins) are likely to be correctly predicted. We use indirect evidence by comparing subcellular localization of these novel binders with the subcellular localization of the known binding proteins and by estimating the charge of the residues in the novel binders that putatively interact with the nucleic acids. A significant overlap in the localization would suggest that the novel nucleic acids binding proteins are correctly predicted. Knowing that DNA/RNA binding residues are positively charged to bind to the negatively charged DNA/RNA molecule, we compare the fraction of positively charged residues among the predicted binding and nonbinding residues in the known and novel binding proteins. Again, similar levels of charge between the known and putative binders, and similar levels of differences from the charge of the nonbinding residues would suggest that the novel putative nucleic acid binders are likely correctly predicted.

6.1 Material and methods

Datasets of native DNA and RNA binding proteins in human

We collect 69178 human proteins that constitute the complete human proteome published in the UniProt database. We annotate the known RNA and DNA binding proteins in the human proteome using the databases utilized in [72] including the gene ontology (GO) terms in UniProt [95], RBPDB database of RNA-binding proteins [92], and animalTFDB of DNA-binding transcription factors [93, 94]. A protein is annotated as a known RNA binding and/or DNA binding protein if it is included in any of the corresponding databases. As a result, 3229 proteins (4.7% of the human proteome) are annotated as the DNA-binding proteins and 1276 proteins (1.8% of the human proteome) are annotated as the RNA-binding proteins.

Prediction of the DNA and RNA binding residues and proteins in human

We apply DRNApred and BindN+ to predict the RNA and DNA binding residues and the RNA and DNA binding proteins in the human proteome. We compare with the results generated by BindN+ because this is the only runtime efficient method that could provide predictions on such large scale (Figure 5.10) and since it offers good levels of predictive performance (Table 5.2). We use the predicted DNA and RNA binding residues generated by both methods to define predicted DNA and RNA binding proteins, respectively. The predicted binding proteins have the number of the corresponding predicted binding residues higher than a threshold that corresponds to the FPR of the binding protein predictions on the test dataset equals 5% (detail are given in section 5.3.4). This is to accommodate for spurious predictions that are associated with the false positive predictions inherent in the outputs of the predictive models.

Assessment of predictive performance

We evaluate the extent of the cross predictions of the native DNA and RNA binding proteins for DRNApred and BindN+ on the human proteome. In other words, we evaluate whether these methods specifically predict only the desired one target type of binding proteins without confusing DNA-binding and the RNA-binding proteins. We calculate the ratio of the fraction of correct predictions to the fraction of the incorrect cross predictions among the known binding proteins. For example, for the prediction of the DNA-binding proteins, we calculate the ratio of the fraction of the correctly predicted known DNA-binding proteins to the fraction of the predicted DNA-binding proteins among the known RNA-binding proteins. This ratio quantifies the ability of a given method to predict the correct type of binding proteins while maintaining low rate of mis-prediction of the incorrect type of binding proteins. A random predictor would attain the ratio = 1, i.e., its fraction of correct predictions in the correct type of nucleic acid is equal to the fraction of incorrect predictions in the other type of the nucleic acid. The ratio > 1 indicates better than random prediction, with the higher number corresponding to a more accurate method.

Moreover, we also compare the cross predictions at the residue level. That is, we assess whether DRNApred and BindN+ specifically predict the target type of binding residues without confusing the DNA-binding and RNA-binding residues. The proteins in the human proteome are annotated per sequence. That means that we have the information whether a given protein binds to DNA or RNA, but not which amino acids in that protein bind to DNA or RNA. Thus, we perform the tests at the residue level indirectly by investigating whether the predicted binding residues are located in the target type of binding proteins. We calculate the ratio of the fraction of predicted binding residues in the correct type of binding proteins to the fraction of the predicted binding residues in the incorrect type of binding proteins using the set of known binding proteins. For example, for the prediction of the DNA-binding residues, we calculate the ratio of the fraction of the predicted DNA-binding residues in the known DNA-binding proteins to the fraction of the predicted DNA-binding residues in the known RNA-binding proteins. A random predictor would secure ratio = 1. The ratio > 1 indicates that a given method is better than random and higher values correspond to stronger predictive performance.

Validation of novel DNA and RNA binding proteins and residues predicted by DRNApred

We analyze the novel DNA and RNA-binding proteins and residues predicted by DRNApred method. These residues and proteins do not overlap with the known DNA and RNA-binding proteins.

First, we investigate and compare the subcellular localization between the novel and known binding proteins. A pattern of similar localization would indicate a high likelihood that the novel binding proteins are in fact correctly predicted. The subcellular location is annotated based on the (GO Cellular Components (CC) terms collected from the UniProt resource for the human proteome. We use all proteins for which this information is complete. Our goal is to find out whether the GO-CC terms associated with the known binding proteins are similar to the GO-CC terms of the novel putative binding proteins. First, for each GO-CC term we calculate its fraction of occurrence (defined as number of occurrences divided by the number of proteins) among the known binding proteins. We also calculate the fraction of occurrence of this GO-CC term in the whole human

proteome to establish a point of reference. The GO-CC term is assumed to be enriched in the known binding proteins if the fraction of occurrence in these proteins is much higher (at least 100% increase) than the fraction in the whole proteome. Next, we investigate whether each enriched GO-CC term is also enriched in the novel putative binding proteins. We calculate the fraction of their occurrence among the novel predicted binding proteins, and compare them with the corresponding points of reference (fractions in the whole proteome excluding the known binding proteins). We consider a given GO-CC term as enriched in the novel putative binding proteins if its fraction of occurrence in these proteins is much higher (by at least 100%) compared to the reference. We hypothesize that the putative novel DNA and RNA binding proteins are correctly predicted if their enriched GO-CC terms cover most of the GO-CC terms enriched in the known binding proteins.

Second, we analyze and compare the residue level predictions between the novel and known binding proteins. We compare the levels of the positively charged residues (Arginine and Lysine) between the binding and nonbinding residues. This is motivated by the observation that DNA and RNA binding residues are positively charged in order to bind to the negatively charged phosphate backbone of the DNA or RNA molecule. We expect and empirically confirm that the fractions of the putative positively charged binding residues (number of positively charged putative binding residues divided by the number of putative binding residues) among the known DNA and RNA binding proteins are substantially higher than the fractions among the putative non-binding residues in these proteins and among the residues in the human proteins. We also compute the same fractions among the novel putative DNA and RNA binding proteins. We hypothesize that the putative novel DNA and RNA binding residues are likely predicted correctly if their fractions are much higher than the fractions for the non-binding residues in these proteins and in the non-binding human proteins, while being comparable to the corresponding fractions for the known binding proteins.

6.2 Results and discussion

6.2.1 Assessment of predictive performance on the known DNA and RNA binding proteins in the human proteome

We assess predictive quality of DRNApred and BindN+ on the human proteome by comparing their cross predictions of the DNA and RNA binding proteins in the sets of known DNA-binding and RNA-binding proteins. In particular, we calculate and compare their ratio of the fraction of the correctly predicted known binding proteins to incorrectly cross predicted known binding proteins. The results are shown using black bars in Figure 6.1. Both DRNApred and BindN+ generate better than random results for the prediction of the DNA-binding proteins; their ratio values are above 1. BindN+ obtains ratio =1.5, which means that it predicts 1.5 times higher fraction of DNA-binding proteins in the known DNA-binding proteins than in the known RNA-binding proteins. DRNApred outperforms BindN+ by securing the ratio = 2, a 50% improvement. Moreover, BindN+ secures ratio ≈ 1 for the prediction of the RNA-binding proteins. This reveals that this method substantially cross-predicts the DNA-binding proteins as RNA binding. The predictions of the RNA-binding proteins by DRNApred are substantially better, with the ratio = 3. This means that DRNApred predicts three time more correct RNA binding proteins compared to the incorrectly cross predicted DNA binding proteins. Overall, these results demonstrate that DRNApred provides specific predictions of the DNA binding and the RNA binding proteins.

We also assess the predictive quality of DRNApred and BindN+ by comparing their cross predictions of the predicted binding residues in the sets of known DNA-binding and RNA-binding proteins. We calculate the ratio of the fraction of the predicted binding residues among the correct type of known binding proteins to the fraction of the putative binding residues in the cross predicted type of known binding proteins. The results are shown using grey bars in Figure 6.1. DRNApred achieves ratio of 2.1 for the prediction of the DNA-binding residues, which means that it predicts over two times higher fraction of DNA-binding residues in the known DNA-binding proteins than in the known RNA-binding proteins. To compare, BindN+ obtains ratio = 1.3, which suggests that it cross

predicts a more substantial number of DNA binding residues. DRNApred also outperforms BindN+ when considering the prediction of RNA-binding residues. BindN+ obtains a ratio at about 1 indicating that it predicts similar fraction of RNA binding residues in both known DNA-binding and RNA-binding proteins. DRNApred secures a high ratio = 6.2. The observation that DRNApred accurately and specifically predicts each type of the nucleic acid binding on the human proteome is consistent with the conclusions that we have reached on the test dataset (see Section 5.3.2).

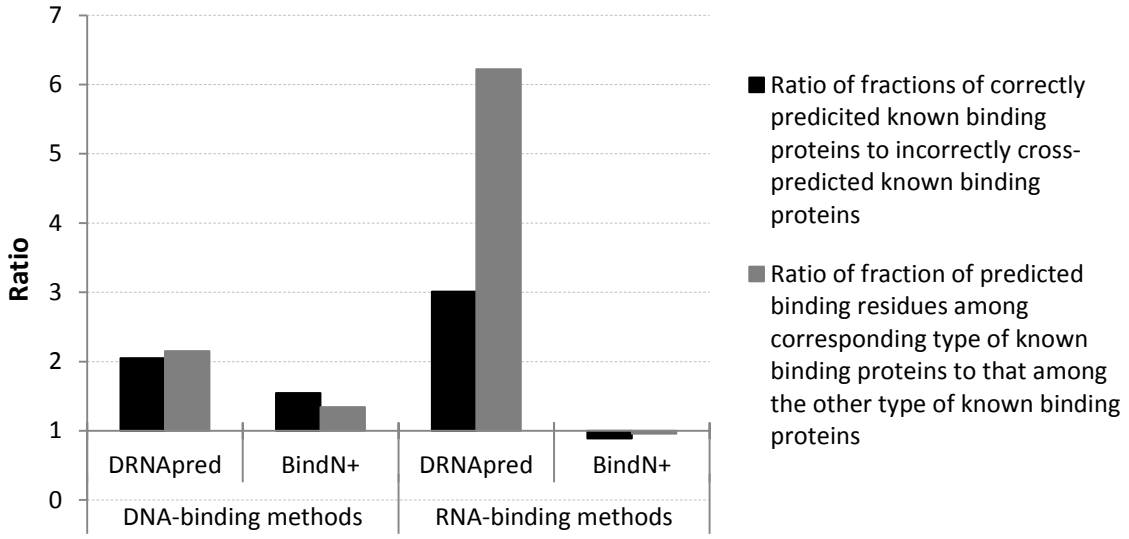


Figure 6.1. Predictive performance of DRNApred and BindN+ for the prediction of binding proteins and residues in the known binding proteins from the human proteome.

The y-axis shows ratio between the fraction of predictions on the correct type of known binding proteins and the fraction of predictions on the cross predicted type of known binding proteins. Random predictor would return ratio =1 and higher ratio indicates a smaller amount of cross predictions. Black (gray) bars summarize comparison for the prediction of the binding proteins (residues).

6.2.2 Evaluation of novel putative RNA and DNA binding proteins

We investigate the degree of an overlap in subcellular localizations, which are annotated based on the gene ontology cellular component (GO-CC) terms, between the novel binding proteins and the known binding proteins. We create a list of the GO-CC terms that are substantially enriched in the known binding proteins, by at least three folds, when compared to their abundance in the whole proteome. These terms are significantly associated with the localization of the DNA binding and the RNA binding proteins. Next,

we calculate the fraction of these terms that are substantially enriched, by at least 100%, in the novel binding proteins. A high fraction indicates that both known and novel putative RNA and DNA binding proteins share similar subcellular localization. Results are shown in Figure 6.2. The *x*-axis shows the minimal level of enrichments of the considered GO-CC terms in the known binding proteins. The numbers of these terms, which are shown above the bars, are fairly high indicating that they can be used to pinpoint the subcellular localization of the native binders. As the required enrichment of the GO-CC in the known binders grows from at least 3 to 10 folds so does the fraction of these terms that are also significantly enriched in the novel putative binders. These fractions start at 65% and 78% for the DNA and RNA binding proteins, respectively, when considering the over 100 terms that are enriched by at least 3 folds in the native binders. Given that we use 42 and 69 terms that are enriched by at least 10 folds in the DNA and RNA binding proteins, respectively, 100% and 90% of them are also enriched in the novel putative binders. This revels that virtually all of the subcellular localizations that are significantly associated with the native RNA and DNA binding proteins are also significantly enriched in the novel RNA and DNA binding proteins that were predicted by DRNApred. In other words, the localizations of the putative and native RNA and DNA proteins are in agreement, suggesting that the novel binding proteins are likely predicted correctly.

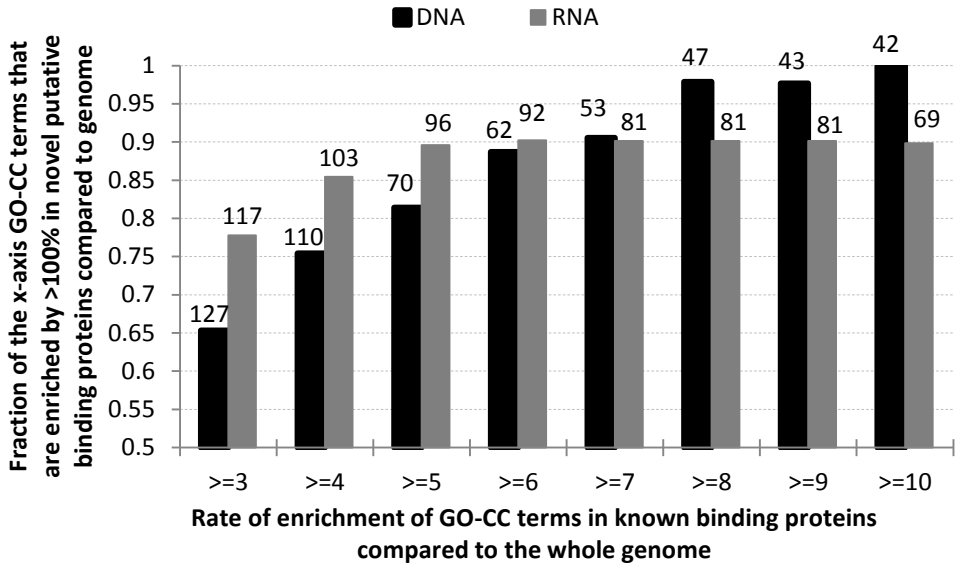


Figure 6.2. Fraction of the gene ontology cellular component (GO-CC) terms associated with the known binding proteins that are also enriched by at least 100% in novel putative binding proteins.

The enrichment in the GO-CC terms is computed against their abundance in the proteome. The *x*-axis shows the minimal level of enrichments of the GO-CC terms in the known binding proteins, the corresponding numbers of significantly enriched terms are shown above the bars. Grey (black) bars summarize results for the RNA (DNA) binding proteins.

We analyze whether the predicted binding residues in the novel binding proteins are similar to the binding residues in the native binding proteins. Since one of the hallmarks of the DNA and RNA binding is inclusion of charged residues, we compare the fractions of the positively charged residues among the predicted binding and nonbinding residues in these proteins with the fractions in the known binding proteins and in the whole proteome. Results are summarized in Figure 6.3. Overall, about 11% of residues in the human proteome are positively charged. There are 3.4 and 2.7 (1.9 and 1.8) times more positively charged residues among the predicted DNA-binding (RNA-binding) residues in the known and novel putative DNA-binding (RNA-binding) proteins, respectively, when compared to the proteome. This is expected for the native binders while the similar levels of the enrichment in the novel putative binders suggest that they are likely correctly identified by DRNApred. Moreover, the fraction of the positively charged residues among the putative nonbinding residues in both known and putative DNA and RNA binding proteins is similar to the level of the positively charged residues in the proteome. The differences in the levels of the positively charged residues between the

putative binding and nonbinding residues support our claim that the putative binding residues generated by DRNApred are likely to bind the two nucleic acids. This observation is consistent for both native and novel putative DNA and RNA binding proteins.

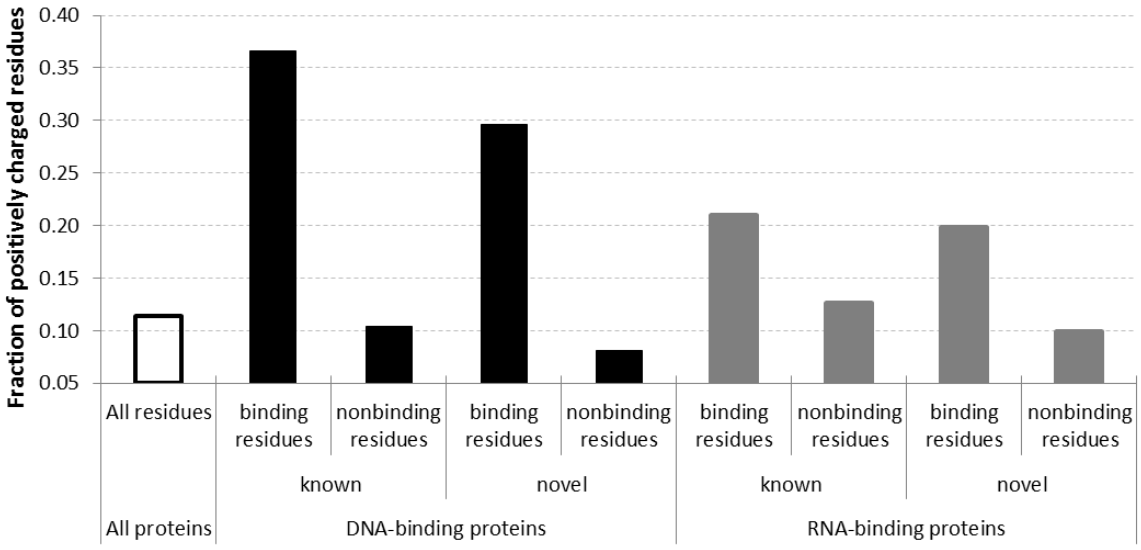


Figure 6.3. Fraction of the positively charged residues among the binding and nonbinding residues in the known and novel binding proteins and among the residues in the entire human proteome.

Grey (black) bars summarize results for the RNA (DNA) binding proteins. The hollow bar shows the results for the human proteome.

6.3 Conclusions

A substantial number of the DNA and RNA binding proteins are yet to be discovered in the human proteome. To this end, we apply our time-efficient DRNApred method to perform large-scale prediction and assessment of the DNA and RNA binding proteins and binding residues in the human proteome. We compare predictive quality, in particular focusing on the cross prediction between RNA and DNA binding in the known binders, between DRNApred and BindN+. We show that DRNApred substantially reduces the cross predictions at both residue and protein levels when compared to BindN+. It obtains a higher ratio of the fraction of correctly predicted known binding proteins and residues to the incorrectly cross predicted known binding proteins and residues, respectively. These observations are consistent with our conclusions from Section 5.3.2 that are based

on the test dataset. We also analyze whether the novel binding proteins share similarities with the native binding proteins. We show that their subcellular localizations and content of positively charged residues among their binding residues are similar. This provides support to the claim that DRNApred can be used to accurately discover novel DNA and RNA binding proteins in human.

Chapter 7

Summary, major contributions, conclusions and future work

This thesis is focused on the development and validation of novel predictive models which accurately and specifically predict DNA-binding and RNA-binding residues from protein sequences in the high-throughput fashion. Our focus is primarily on the ability of these models to differentiate between DNA- and RNA-binding residues while offering state-of-the-art overall predictive performance.

In goal 1 we comprehensively reviewed 14 sequence-based methods for the prediction of DNA-binding residues and 16 methods for the prediction of RNA-binding residues. We summarized how they define binding residues and discussed the variety of their designs, defined in terms of how they encode the input amino acids, predictive model that they apply, and format of their outputs. Although these methods vary in their design, many of them share certain design elements like the use evolutionary information and sliding windows to encode inputs and the use of SVM as the predictive model. The input features used to predict DNA-binding residues are similar to the inputs used by the predictors of RNA-binding residues, which is not surprising given the chemical similarity between DNA and RNA. We also empirically assessed a selected set of conveniently available to the end user predictors of the DNA-binding and RNA-binding residues on a test dataset with the DNA-binding and RNA-binding proteins, respectively. Our results demonstrate that these predictors provide good predictive quality when separating RNA binding residues from nonbinding residues and DNA binding residues from nonbinding residues. Their AUCs range between around 0.7 and 0.8 and MCCs between around 0.1 and 0.3; these results are based on a new and challenging test dataset characterized by low sequence similarity to the datasets used to design these methods. We also assessed how these predictors differentiate between different types of nucleic acid binding

residues by testing them on the test dataset that includes both DNA and RNA binding proteins; this was never done before. We found that these methods substantially cross predict the binding residues, which means that they mis-predict RNA-binding residues as DNA-binding and vice versa. This is likely the results of use of similar input features and the fact that these methods were trained based on data sets that use either only DNA-binding or only RNA-binding proteins. These results that were published online in May 2015 [77] were very recently confirmed by a more recent study that was released in December 2015 [35]. The high amounts of cross predictions prompted us to investigate the development of methods that would reduce these amounts, including consensus-based approaches and new predictors.

In goal 2 we investigated a comprehensive set of designs of consensus-based methods with the underlying goal to improve predictive performance and reduce the cross predictions. These consensuses include simple logic-based and majority vote-based consensuses and a more sophisticated machine learning (ML) consensus. Our empirical evaluation have shown that the logic-based consensus that combines several predictors with logic AND and logic OR operators, and a simple majority vote consensus do not offer improvements when compared with the individual input methods on a challenging test dataset. Moreover, while we demonstrate that the ML consensus offers improved predictive quality; neither consensus type solves the cross prediction problem. To this end, we attempted to address the cross prediction by conducting a first-of-its-kind study in which we designed novel consensuses for the combined prediction of DNA- and RNA-binding residues. We designed three types of such consensuses including a ML based approach. We empirically show that this ML consensus offers strong predictive performance in the combined prediction and also for the prediction of DNA-binding or RNA-binding residues individually. It provides higher values of MCC compared with the best-performing individual predictors. Most importantly, it also substantially reduces the cross-prediction. However, this consensus is hard to use, given that the end user would have to collect predictions from 8 methods, and is not runtime efficient since these predictors are relatively slow.

In goal 3 we address the weaknesses of the consensus-based solution by developing a novel high throughput (low runtime) method DRNApred that accurately and specifically predict only DNA-binding and only RNA-binding residues from protein sequences. The three main novel features of our design are the application of a new in this area architecture with two layers, and use of a new dataset with both DNA-binding and RNA-binding proteins and a weight-based mechanism to penalize cross predictions. In the first layer, we considered a comprehensive set of numerical features to encode the input sequence. They include the amino acid type, physiochemical properties of amino acids, evolutionary profiles, and putative intrinsic disorder, secondary structure, solvent accessibility. We introduced the weight-based mechanism into the model training process that includes empirical selection of a subset of predictive and non-redundant features and computation of a logistic regression-based predictive model. The second layer includes regression-based predictive model which takes DNA-binding and RNA-binding predictions from the first layer as input and redoes predictions to further reduce the cross predictions. Our empirical results show that the three novel design ideas results in substantial reduction of the cross predictions. We compared DRNApred with selected state-of-the-art existing methods on a challenging test dataset that shares low sequence similarity with proteins used to build these predictive tools. This empirical comparison demonstrates that DRNApred secures similar overall predictive quality (measured with AUC and AULC) when compared to the other methods and it also dramatically reduces the cross predictions (measured with AURC and AULRC) for the prediction of both DNA-binding and RNA-binding residues. DRNApred also finds arguably higher quality false positives (novel putative binding residues) that are located close to the native binding residues. We also compared DRNApred and the other considered methods for the prediction of DNA-binding and RNA-binding proteins. The results show that DRNApred outperforms these methods. It secures the highest AUC value and correctly predicts more binding proteins at low false positive rates. We also demonstrate that DRNApred is computationally efficient and could be applied on the proteomic scale.

In goal 4 we applied the runtime efficient DRNApred and BindN+ methods to perform a large scale prediction on the entire human proteome. We compared the predictive quality of these two methods for the prediction of RNA and DNA binding

proteins and residues among the known DNA and RNA binding proteins. Results show that DRNApred outperforms BindN+ by substantially reducing the cross predictions. More importantly, we analyze the novel binding proteins that are predicted by DRNApred. We compare the subcellular localizations between these novel binding proteins and the known binding proteins. Results show that these two sets of proteins share similar localizations which suggest that our novel binding proteins are likely to be correctly predicted. We also analyze the binding residues predicted with DRNApred. Using one of the hallmarks of the protein-nucleic acids binding, we compare the fraction of the positively charged residues among the predicted binding and nonbinding residues in the novel and known binding proteins. The results show that the predicted binding residues have higher fraction of positively charged residues compared to the predicted nonbinding residues in both known and novel binding proteins. This further validates our claim that the predicted binding residues are likely to be correct.

7.1 Major contributions

- **Goal 1: Assessment of predictive performance of existing sequence-based DNA- and RNA- binding residue predictors.**
    - Comprehensive review of 30 sequence-based predictors of DNA-binding and RNA-binding residues. This review covered aspects of their design, outputs and availability. Compared to the previous reviews that consider only methods for prediction of one type of nucleic acid binding, our review analyzes both the DNA-binding and RNA-binding residue predictors, and includes several recently published methods.
    - Development of a new benchmark dataset characterized by a more complete annotation of RNA and DNA binding residues. Our dataset was published and is the first to contain both DNA-binding proteins and RNA-binding proteins for which the binding annotation is improved by transferring annotation from the same or similar proteins.

- Empirical assessment of predictors of the DNA-binding (RNA-binding) residues on RNA-binding (DNA-binding) proteins to quantify the extent of cross predictions.
- Introduction of a new measure, ratio, that quantifies the amount of cross predictions for the binary predictions.

- **Goal 2: Development of novel consensus-based predictors to improve accuracy of the prediction of DNA- and RNA- binding residues.**
  - Comprehensive study of different types of consensuses including simple consensuses and machine learning consensus for the prediction of DNA/RNA-binding residues. We demonstrate that the machine learning based consensus provides improved predictive performance when compared with the individual predictors of DNA/RNA-binding residues.
  - First-of-its-kind study to design a method for the combined prediction of DNA- and RNA- binding residues to solve the cross prediction between DNA-binding and RNA-binding residues. We show that our approach substantially reduces the cross prediction problem.

- **Goal 3: Development of DRNApred, a new high-throughput method that accurately and specifically predicts only DNA-binding and only RNA-binding residues.**
  - Development of a novel high-throughput method that accurately and specifically predicts only the DNA-binding and RNA-binding residues from protein sequence. Our method is developed using three novel ideas including use of a combined dataset of both DNA and RNA binding proteins, use of a penalty for the cross predictions, and using a second predictive layer.
  - Introduction of a new measure, AURC, that quantifies the amount of cross predictions for the propensity predictions.

- First-of-its kind analysis of the predicted binding residues. We analyze how close they are from the native binding residues, and how the predictive quality would change if the predictions in a close proximity of the native binding residues were considered as correctly predicted.
- First-of-its kind analysis of the predictive performance of DRNApred and the considered predictors of the DNA/RNA-binding residues for the prediction of the DNA/RNA-binding proteins.
- First-of-its kind comparison of the runtime of DRNApred and the considered predictors of the DNA and RNA binding residues.

- **Goal 4: Identification of known and novel DNA- and RNA-binding residues/proteins on proteomic-scale.**
  - Prediction of the DNA and RNA binding proteins in human proteome using DRNApred and BindN+.
  - Assessment of how specifically DRNApred and BindN+ predict target type of binding proteins/residues in the human proteome.
  - Validation of the novel binding proteins/residues predicted by DRNApred.

7.2 Conclusions

The major conclusions from this thesis as are follows. First, the current methods for the prediction of the DNA and RNA binding residues offer good predictive performance when tested on the corresponding type of the nucleic acid. However, these methods substantially cross predict between DNA and RNA binding residues. Second, simple consensus methods do not offer improvements compared to individual predictors of the DNA and RNA binding residues. Third, machine learning-based consensuses that address prediction of DNA or RNA residues offer improved predictive performance but they also suffer high rates of cross predictions. Four, a novel type of consensus combining predictions of DNA and RNA binding residues offers strong predictive performance and reduces the cross predictions. However, these consensuses are difficult to implement and

are characterized by long runtime. Five, accurate prediction combined with low cross prediction, low runtime and convenient to run architecture is possible with the help of novel predictor DRNApred. Six, DRNApred generates accurate predictions on proteomic scale that can be used to accurately find novel putative DNA and RNA binding proteins.

7.3 Future work

Given the contents of the datasets that were used to develop our and other methods in this area, these methods are designed to predict the DNA and RNA binding residues in the structured protein and protein regions. This is because these datasets are derived from structures of the protein-RNA and protein-DNA complexes. However, research shows that protein-DNA and protein-RNA interactions frequently occur also in the intrinsically disordered regions and intrinsically disordered proteins (Section 2.1 discusses intrinsic disorder) [97-100]. A time efficient method disoRDPbind [72] that predicts nucleic acid binding residues in the disordered regions was recently published. Thus, one interesting extension of this work would be to combine the predictions of DNA and RNA binding residues in the structured regions (by DRNApred) with the predictions in the disordered regions (by disoRDPbind). This would lead to the development of a more complete set of putative DNA and RNA binding proteins and residues.

Given the arguably high predictive quality of the putative results generated by DRNApred and DisoRDPbind, another practical extension would be to predict DNA and RNA binding proteins and residues on large scale of multiple proteomes (species) and make these results accessible to the end users as a convenient web-based database. Similar efforts have been already made to provide access to putative annotations of intrinsic disorder on the scale of thousands of proteomes: the MobiDB [101, 102] or D^2P^2 [103] databases. These databases are widely used, which is evident based on their high citations counts, relative to when they were published and the venue where they were published. D^2P^2 and MobiDB that were published in 2013 and 2012, respectively, were already cited 90 times each (source: Google Scholar as of May 2016; the 90 citations for MobiDB include 56 citations for version 1 and 36 for version 2). A large-scale database of putative DNA and RNA binding would ease access to this information for the less

computer-savvy biologists who would be able to retrieve pre-computed results using a web browser.

Both of these future advances would potentially have substantial impact. Analysis of the DNA and RNA binding predictions across different proteomes/species might provide novel insights into the evolution and cellular functions of the corresponding proteins. It may also help us to better understand the molecular-level mechanisms underlying the protein DNA/RNA interactions.

Bibliography

1. Luscombe, N.M., et al., *An overview of the structures of protein-DNA complexes.* Genome Biol, 2000. **1**(1): p. Reviews001.
2. Charoensawan, V., D. Wilson, and S.A. Teichmann, *Genomic repertoires of DNA-binding transcription factors across the tree of life.* Nucleic acids research, 2010. **38**(21): p. 7364-7377.
3. Re, A., et al., *RNA–Protein Interactions: An Overview.* Rna Sequence, Structure, and Function: Computational and Bioinformatic Methods, 2014: p. 491-521.
4. Noller, H.F., *RNA structure: reading the ribosome.* Science, 2005. **309**(5740): p. 1508-1514.
5. Glisovic, T., et al., *RNA-binding proteins and post-transcriptional gene regulation.* FEBS letters, 2008. **582**(14): p. 1977-1986.
6. Berman, H.M., et al., *The protein data bank.* Nucleic acids research, 2000. **28**(1): p. 235-242.
7. Pruitt, K.D., et al., *RefSeq: an update on mammalian reference sequences.* Nucleic acids research, 2014. **42**(D1): p. D756-D763.
8. Kuznetsov, I.B., et al., *Using evolutionary and structural information to predict DNA‐binding sites on DNA‐binding proteins.* PROTEINS: Structure, Function, and Bioinformatics, 2006. **64**(1): p. 19-27.
9. Terribilini, M., et al., *Prediction of RNA binding sites in proteins from amino acid sequence.* Rna, 2006. **12**(8): p. 1450-1462.
10. Nagarajan, R., S. Ahmad, and M.M. Gromiha, *Novel approach for selecting the best predictor for identifying the binding sites in DNA binding proteins.* Nucleic acids research, 2013. **41**(16): p. 7606-7614.
11. Zhao, H., Y. Yang, and Y. Zhou, *Prediction of RNA binding proteins comes of age from low resolution to high resolution.* Molecular bioSystems, 2013. **9**(10): p. 2417-2425.
12. Puton, T., et al., *Computational methods for prediction of protein–RNA interactions.* Journal of structural biology, 2012. **179**(3): p. 261-268.
13. Walia, R.R., et al., *Protein-RNA interface residue prediction using machine learning: an assessment of the state of the art.* BMC bioinformatics, 2012. **13**(1): p. 1.
14. Si, J., et al., *MetaDBSite: a meta approach to improve protein DNA-binding sites prediction.* BMC systems biology, 2011. **5**(Suppl 1): p. S7.
15. Yan, J., M. Marcus, and L. Kurgan, *Comprehensively designed consensus of standalone secondary structure predictors improves Q 3 by over 3%.* Journal of Biomolecular Structure and Dynamics, 2014. **32**(1): p. 36-51.
16. Zhang, H., et al., *Critical assessment of high-throughput standalone methods for secondary structure prediction.* Briefings in bioinformatics, 2011. **12**(6): p. 672-688.

17. Fan, X. and L. Kurgan, *Accurate prediction of disorder in protein chains with a comprehensive and empirically designed consensus.* Journal of Biomolecular Structure and Dynamics, 2014. **32**(3): p. 448-464.
18. Kozlowski, L.P. and J.M. Bujnicki, *MetaDisorder: a meta-server for the prediction of intrinsic disorder in proteins.* BMC bioinformatics, 2012. **13**(1): p. 1.
19. Walsh, I., et al., *Comprehensive large-scale assessment of intrinsic protein disorder.* Bioinformatics, 2014: p. btu625.
20. Albrecht, M., et al., *Simple consensus procedures are effective and sufficient in secondary structure prediction.* Protein Engineering, 2003. **16**(7): p. 459-462.
21. Gong, Y., et al., *Crystal structures of aprataxin ortholog Hnt3 reveal the mechanism for reversal of 5′-adenylated DNA.* Nature structural & molecular biology, 2011. **18**(11): p. 1297-1299.
22. Hellman, L.M. and M.G. Fried, *Electrophoretic mobility shift assay (EMSA) for detecting protein–nucleic acid interactions.* Nature protocols, 2007. **2**(8): p. 1849-1861.
23. Kuo, M.-H. and C.D. Allis, *In vivo cross-linking and immunoprecipitation for studying dynamic protein: DNA associations in a chromatin environment.* Methods, 1999. **19**(3): p. 425-433.
24. Dahl, J.A. and P. Collas, *A rapid micro chromatin immunoprecipitation assay (ChIP).* Nature protocols, 2008. **3**(6): p. 1032-1045.
25. Dahl, J.A. and P. Collas, *μChIP: Chromatin Immunoprecipitation for Small Cell Numbers.* Chromatin Immunoprecipitation Assays: Methods and Protocols, 2009: p. 59-74.
26. Nelson, J., O. Denisenko, and K. Bomsztyk, *The fast chromatin immunoprecipitation method.* Chromatin Immunoprecipitation Assays: Methods and Protocols, 2009: p. 45-57.
27. Hollis, T., *Crystallization of protein-DNA complexes.* Macromolecular Crystallography Protocols: Volume 1, Preparation and Crystallization of Macromolecules, 2007: p. 225-237.
28. Varani, G., Y. Chen, and T.C. Leeper, *NMR studies of protein-nucleic acid interactions.* Protein NMR Techniques, 2004: p. 289-312.
29. Zhao, H., Y. Yang, and Y. Zhou, *Structure-based prediction of RNA-binding domains and RNA-binding sites and application to structural genomics targets.* Nucleic acids research, 2010: p. gkq1266.
30. Liu, R. and J. Hu, *DNABind: A hybrid algorithm for structure‐based prediction of DNA‐binding residues by combining machine learning‐and template‐based approaches.* Proteins: Structure, Function, and Bioinformatics, 2013. **81**(11): p. 1885-1899.
31. Bhardwaj, N. and H. Lu, *Residue-level prediction of DNA-binding sites and its application on DNA-binding protein predictions.* FEBS letters, 2007. **581**(5): p. 1058-1066.
32. Dey, S., et al., *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 research, 2012: p. gks405.

33. Tjong, H. and H.-X. Zhou, *DISPLAR: an accurate method for predicting DNA-binding sites on protein surfaces.* Nucleic Acids Research, 2007. **35**(5): p. 1465-1477.
34. Liu, Z.-P., et al., *Prediction of protein–RNA binding sites by a random forest method with combined features.* Bioinformatics, 2010. **26**(13): p. 1616-1622.
35. Miao, Z. and E. Westhof, *A Large-Scale Assessment of Nucleic Acids Binding Site Prediction Programs.* PLoS Comput Biol, 2015. **11**(12): p. e1004639.
36. Si, J., R. Zhao, and R. Wu, *An overview of the prediction of protein DNA-binding sites.* International journal of molecular sciences, 2015. **16**(3): p. 5194-5215.
37. Si, J., et al., *Computational Prediction of RNA-Binding Proteins and Binding Sites.* International journal of molecular sciences, 2015. **16**(11): p. 26303-26317.
38. Yan, C., et al., *Predicting DNA-binding sites of proteins from amino acid sequence.* BMC bioinformatics, 2006. **7**(1): p. 1.
39. Wu, J., et al., *Prediction of DNA-binding residues in proteins from amino acid sequences using a random forest model with a hybrid feature.* Bioinformatics, 2009. **25**(1): p. 30-35.
40. Cheng, C.-W., et al., *Predicting RNA-binding sites of proteins using support vector machines and evolutionary information.* BMC bioinformatics, 2008. **9**(Suppl 12): p. S6.
41. Kurgan, L. and F. Miri Disfani, *Structural protein descriptors in 1-dimension and their sequence-based predictions.* Current Protein and Peptide Science, 2011. **12**(6): p. 470-489.
42. Gao, M. and J. Skolnick, *A threading-based method for the prediction of DNA-binding proteins with application to the human genome.* PLoS Comput Biol, 2009. **5**(11): p. e1000567.
43. Zhao, H., Y. Yang, and Y. Zhou, *Highly accurate and high-resolution function prediction of RNA binding proteins by fold recognition and binding affinity prediction.* RNA biology, 2011. **8**(6): p. 988-996.
44. Ahmad, S., M.M. Gromiha, and A. Sarai, *Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information.* Bioinformatics, 2004. **20**(4): p. 477-486.
45. Ahmad, S. and A. Sarai, *PSSM-based prediction of DNA binding sites in proteins.* BMC bioinformatics, 2005. **6**(1): p. 33.
46. Wang, L. and S.J. Brown, *BindN: a web-based tool for efficient prediction of DNA and RNA binding sites in amino acid sequences.* Nucleic acids research, 2006. **34**(suppl 2): p. W243-W248.
47. Ho, S.-Y., et al., *Design of accurate predictors for DNA-binding sites in proteins using hybrid SVM–PSSM method.* Biosystems, 2007. **90**(1): p. 234-241.
48. Hwang, S., Z. Gou, and I.B. Kuznetsov, *DP-Bind: a web server for sequence-based prediction of DNA-binding residues in DNA-binding proteins.* Bioinformatics, 2007. **23**(5): p. 634-636.
49. Ofran, Y., V. Mysore, and B. Rost, *Prediction of DNA-binding residues from sequence.* Bioinformatics, 2007. **23**(13): p. i347-i353.
50. Lee, J.-h., et al. *Striking similarities in diverse telomerase proteins revealed by combining structure prediction and machine learning approaches*. in *Pacific*

Symposium on Biocomputing. Pacific Symposium on Biocomputing. 2008. NIH Public Access.

51. Wang, L., M.Q. Yang, and J.Y. Yang, *Prediction of DNA-binding residues from protein sequence information using random forests.* Bmc Genomics, 2009. **10**(1): p. 1.
52. Chu, W.-Y., et al., *ProteDNA: a sequence-based predictor of sequence-specific DNA-binding residues in transcription factors.* Nucleic acids research, 2009: p. gkp449.
53. Wang, L., et al., *BindN+ for accurate prediction of DNA and RNA-binding residues from protein sequence features.* BMC Systems Biology, 2010. **4**(1): p. 1.
54. Carson, M.B., R. Langlois, and H. Lu, *NAPS: a residue-level nucleic acid-binding prediction server.* Nucleic acids research, 2010. **38**(suppl 2): p. W431-W435.
55. Ma, X., et al., *Sequence-based prediction of DNA-binding residues in proteins with conservation and correlation information.* Computational Biology and Bioinformatics, IEEE/ACM Transactions on, 2012. **9**(6): p. 1766-1775.
56. Jeong, E., I.-F. Chung, and S. Miyano, *A neural network method for identification of RNA-interacting residues in protein.* Genome informatics, 2004. **15**(1): p. 105-116.
57. Jeong, E. and S. Miyano, *A weighted profile based method for protein-RNA interacting residue prediction*, in *Transactions on Computational Systems Biology IV*. 2006, Springer. p. 123-139.
58. Wang, Y., et al., *PRINTR: prediction of RNA binding sites in proteins using SVM and profiles.* Amino acids, 2008. **35**(2): p. 295-302.
59. Tong, J., P. Jiang, and Z.-h. Lu, *RISP: a web-based server for prediction of RNA-binding sites in proteins.* Computer methods and programs in biomedicine, 2008. **90**(2): p. 148-153.
60. Kumar, M., M.M. Gromiha, and G. Raghava, *Prediction of RNA binding sites in a protein using SVM and PSSM profile.* Proteins: Structure, Function, and Bioinformatics, 2008. **71**(1): p. 189-194.
61. Spriggs, R.V., et al., *Protein function annotation from sequence: prediction of residues interacting with RNA.* Bioinformatics, 2009. **25**(12): p. 1492-1497.
62. Murakami, Y., et al., *PiRaNhA: a server for the computational prediction of RNA-binding residues in protein sequences.* Nucleic acids research, 2010. **38**(suppl 2): p. W412-W416.
63. Huang, Y.-F., et al., *Predicting RNA-binding residues from evolutionary information and sequence conservation.* BMC genomics, 2010. **11**(4): p. 1.
64. Zhang, T., et al., *Analysis and prediction of RNA-binding residues using sequence, evolutionary conservation, and predicted secondary structure and solvent accessibility.* Current Protein and Peptide Science, 2010. **11**(7): p. 609-628.
65. Wang, C.-c., et al., *Identification of RNA-binding sites in proteins by integrating various sequence information.* Amino acids, 2011. **40**(1): p. 239-248.
66. Ma, X., et al., *Prediction of RNA‐binding residues in proteins from primary sequence using an enriched random forest model with a novel hybrid feature.* Proteins: Structure, Function, and Bioinformatics, 2011. **79**(4): p. 1230-1239.

67. Terribilini, M., et al., *RNABindR: a server for analyzing and predicting RNA-binding sites in proteins.* Nucleic acids research, 2007. **35**(suppl 2): p. W578-W584.
68. Schneider, R. and C. Sander, *The HSSP database of protein structure-sequence alignments.* Nucleic acids research, 1996. **24**(1): p. 201-205.
69. Hsu, C.-M., et al., *Efficient discovery of structural motifs from protein sequences with combination of flexible intra-and inter-block gap constraints*, in *Advances in Knowledge Discovery and Data Mining*. 2006, Springer. p. 530-539.
70. Lichtarge, O., H.R. Bourne, and F.E. Cohen, *An evolutionary trace method defines binding surfaces common to protein families.* Journal of molecular biology, 1996. **257**(2): p. 342-358.
71. Zvelebil, M.J., et al., *Prediction of protein secondary structure and active sites using the alignment of homologous sequences.* Journal of molecular biology, 1987. **195**(4): p. 957-961.
72. Peng, Z. and L. Kurgan, *High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder.* Nucleic acids research, 2015: p. gkv585.
73. Anderson, T.W. and D.A. Darling, *Asymptotic theory of certain" goodness of fit" criteria based on stochastic processes.* The annals of mathematical statistics, 1952: p. 193-212.
74. Fornes, O., et al., *On the Use of Knowledge-Based Potentials for the Evaluation of.* Advances in protein chemistry and structural biology, 2014. **94**: p. 77.
75. Kauffman, C. and G. Karypis, *Computational tools for protein–DNA interactions.* Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2012. **2**(1): p. 14-28.
76. Liu, L.A. and P. Bradley, *Atomistic modeling of protein–DNA interaction specificity: progress and applications.* Current opinion in structural biology, 2012. **22**(4): p. 397-405.
77. Yan, J., S. Friedrich, and L. Kurgan, *A comprehensive comparative review of sequence-based predictors of DNA-and RNA-binding residues.* Briefings in bioinformatics, 2015: p. bbv023.
78. Chen, K., et al., *A critical comparative assessment of predictions of protein-binding sites for biologically relevant organic compounds.* Structure, 2011. **19**(5): p. 613-621.
79. Zhang, Y. and J. Skolnick, *TM-align: a protein structure alignment algorithm based on the TM-score.* Nucleic acids research, 2005. **33**(7): p. 2302-2309.
80. Altschul, S.F., et al., *Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.* Nucleic acids research, 1997. **25**(17): p. 3389-3402.
81. Huang, Y., et al., *CD-HIT Suite: a web server for clustering and comparing biological sequences.* Bioinformatics, 2010. **26**(5): p. 680-682.
82. Yan, J. and L. Kurgan, *Consensus-Based Prediction of RNA and DNA Binding Residues from Protein Sequences*, in *Pattern Recognition and Machine Intelligence*. 2015, Springer. p. 501-511.
83. Dosztanyi, Z., et al., *The pairwise energy content estimated from amino acid composition discriminates between folded and intrinsically unstructured proteins.* Journal of molecular biology, 2005. **347**(4): p. 827-839.

84. Walsh, I., et al., *ESpritz: accurate and fast prediction of protein disorder.* Bioinformatics, 2012. **28**(4): p. 503-509.
85. Jones, D.T., *Protein secondary structure prediction based on position-specific scoring matrices.* Journal of molecular biology, 1999. **292**(2): p. 195-202.
86. Rost, B. and C. Sander, *Conservation and prediction of solvent accessibility in protein families.* Proteins: Structure, Function, and Bioinformatics, 1994. **20**(3): p. 216-226.
87. Ahmad, S. and M.M. Gromiha, *NETASA: neural network based prediction of solvent accessibility.* Bioinformatics, 2002. **18**(6): p. 819-824.
88. Ahmad, S., M.M. Gromiha, and A. Sarai, *RVP-net: online prediction of real valued accessible surface area of proteins from single sequences.* Bioinformatics, 2003. **19**(14): p. 1849-1851.
89. Kawashima, S., et al., *AAindex: amino acid index database, progress report 2008.* Nucleic acids research, 2008. **36**(suppl 1): p. D202-D205.
90. Remmert, M., et al., *HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment.* Nature methods, 2012. **9**(2): p. 173-175.
91. Rao, H., et al., *Update of PROFEAT: a web server for computing structural and physicochemical features of proteins and peptides from amino acid sequence.* Nucleic acids research, 2011. **39**(suppl 2): p. W385-W390.
92. Berglund, A.-C., et al., *InParanoid 6: eukaryotic ortholog clusters with inparalogs.* Nucleic acids research, 2008. **36**(suppl 1): p. D263-D266.
93. Zhang, H.-M., et al., *AnimalTFDB: a comprehensive animal transcription factor database.* Nucleic acids research, 2012. **40**(D1): p. D144-D149.
94. Zhang, H.-M., et al., *AnimalTFDB 2.0: a resource for expression, prediction and functional study of animal transcription factors.* Nucleic acids research, 2015. **43**(D1): p. D76-D81.
95. Consortium, U., *Activities at the universal protein resource (UniProt).* Nucleic acids research, 2014. **42**(11): p. 7486.
96. Gerstberger, S., M. Hafner, and T. Tuschl, *A census of human RNA-binding proteins.* Nature Reviews Genetics, 2014. **15**(12): p. 829-845.
97. Peng, Z., et al., *Exceptionally abundant exceptions: comprehensive characterization of intrinsic disorder in all domains of life.* Cellular and Molecular Life Sciences, 2015. **72**(1): p. 137-151.
98. Wang, C., V.N. Uversky, and L. Kurgan, *Disordered nucleiome: Abundance of intrinsic disorder in the DNA‐and RNA‐binding proteins in 1121 species from Eukaryota, Bacteria and Archaea.* Proteomics, 2016.
99. Ward, J.J., et al., *Prediction and functional analysis of native disorder in proteins from the three kingdoms of life.* Journal of molecular biology, 2004. **337**(3): p. 635-645.
100. Chen, J.W., et al., *Conservation of intrinsic disorder in protein domains and families: II. functions of conserved disorder.* Journal of proteome research, 2006. **5**(4): p. 888-898.
101. Potenza, E., et al., *MobiDB 2.0: an improved database of intrinsically disordered and mobile proteins.* Nucleic acids research, 2014: p. gku982.
102. Di Domenico, T., et al., *MobiDB: a comprehensive database of intrinsic protein disorder annotations.* Bioinformatics, 2012. **28**(15): p. 2080-2081.

103. Oates, M.E., et al., *D2P2: database of disordered protein predictions.* Nucleic acids research, 2012: p. gks1226.