DNA Binding Protein Identification Using HMM Profile



Rianon Zaman

Department of Computer Science and Engineering

United International University

A thesis submitted for the degree of

Master of Science in Computer Science and Engineering

February 2018

Contents

List of Figures **iv**

List of Tables **vi**

1 Introduction **1**

1.1 Problem Description 1
1.2 Motivation 2
1.3 Aims & Objectives 2
1.4 Methodology 3
1.5 Contribution 3
1.6 Organization 3

2 Background **4**

2.1 Biological Preliminaries 4
2.1.1 DNA 4
2.1.2 Protein 5
2.1.3 Protein Structure 6
2.1.4 Protein Synthesis 8
2.1.5 DNA-binding Protein 9
2.1.6 Hidden Markov Model 10
2.2 Literature Review 11

3 Materials & Methods **13**

3.1 Datasets 13
3.1.0.1 Dataset benchmark1075 13
3.1.0.2 Dataset independent186 14

3.2 System Overview 14
3.3 Feature Extraction 16
3.3.0.1 Monogram Features 17
3.3.0.2 Bigram Features 17
3.4 StandardScaler 18
3.5 Cross Validation 18
3.6 Description of the Classifier 19
3.6.1 Support Vector Machine 19
3.6.2 Random Forest: 20
3.6.3 AdaBoost: 20
3.7 Performance Evaluation 20

4 Result & Discussion **22**
4.0.1 Experimental Design 22
4.0.2 Comparison with Other Methods 23
4.0.3 HMMBinder Web Server Implementation 27
4.1 Data and Material Availability 27

5 Conclusion and Future Work **33**
5.1 Conclusion 33
5.2 Limitations 33
5.3 Future Work 33

Bibliography **35**

List of Figures

2.1 DNA Structure [1] 5
2.2 Levels of Protein Structure [2] 7
2.3 The Central Dogma of molecular biology : replication, transcription and translation.[3] 10
2.4 DNA Binding Protein.[4] 11

3.1 System diagram of HMMBinder 15

4.1 **Using Monogram Features.** Receiver operating characteristic curves for (a) SVM Linear kernel classifier using HMM Monogram features, (b) SVM Linear kernel classifier using PSSM Monoram features, (c) SVM RBF kernel classifier using HMM Monogram features and(d) SVM RBF kernel classifier using PSSM Monogram features. 24
4.2 **Using Bigram Features.** Receiver operating characteristic curves for (a) SVM Linear kernel classifier using HMM Bigram features, (b) SVM Linear kernel classifier using PSSM Bigram features, (c) SVM RBF kernel classifier using HMM Bigram features and (d) SVM RBF kernel classifier using PSSM Bigram features. 25
4.3 **Using (Mono+Bi)gram Features.** Receiver operating characteristic curves for (a) SVM Linear kernel classifier using HMM Mono+Bigram features, (b) SVM Linear kernel classifier using PSSM Mono+Bigram features, (c) SVM RBF kernel classifier using HMM Mono+Bigram features and (d) SVM RBF kernel classifier using PSSM Mono+Bigram features. 26
4.4 Screen shot of Web-Server homepage 28

4.5 Screen shot of Web-Server Result 28
4.6 Figure 4.10: Screen shot of Web-Server Datasets. 31
4.7 Screen shot of Web-Server Guide or Readme 31
4.8 Screen shot of Journal citation 32

List of Tables

3.1 Summary of Benchmark Datasets explored in this research: 14

4.1 Comparison of performance of different features and SVM kernels on the benchmark dataset using 10-fold cross validation. 29
4.2 Comparison of performance of the proposed method with other state-of-the-art predictors using jack knife test on the benchmark dataset. . . . 30
4.3 Comparison of performance of the proposed method with other state-of-the-art predictors on the independent dataset. 30

Abstract

DNA-binding proteins play important role in various processes within the cell.Various machine learning classification algorithms and feature extraction techniques have been used to solve this computational problem of identification of DNA-binding proteins in the last decade. It has many important uses in transcription, DNA replication, compacting the chromosomal DNA, study of antibiotics, drugs, steroids. In this paper, we propose a novel DNA-binding protein prediction method named HMMBinder. HMMBinder uses monogram and bigram features extracted from the HMM profiles of the protein sequences. We have proposed an approach to predict whether a given protein sequence is DNA Binding or not. To the best of our knowledge, this is the first application of HMM profile based features for DNA-binding protein prediction problem. Support Vector Machines(SVM) is used as a classification algorithm. Our method was tested on standard benchmark dataset and was able to improve significantly over the performance of the state-of-the-art methods. We get an accuracy of 86%.

Published Papers

Work relating to the research presented in this thesis has been published by the author in the following peer-reviewed journals:

1. Rianon Zaman, Shahana Yasmin Chowdhury, Mahmood A Rashid, Alok Sharma, Abdollah Dehzangi and Swakkhar Shatabda, HMM-Binder: DNA Binding Protein Prediction Using HMM Profile Based Features, BioMed Research International , Vol. 2017, 14 November 2017.

Acknowledgements

Here I want to express my gratitude and sincere appreciation to all those who have helped me to complete this MCSE Thesis.

I want to specially thank to my supervisor Dr. Swakkhar Shatabda, Associate Professor, Dept of CSE, UIU, for his numerous corrections, advises and support during the actual work process and writing the thesis. Without his guidance it's impossible for me to complete the thesis.

Again, I want to thank to Shahana Yeasmin Chowdhury for her constant support to accomplish the experiment successfully.

My heartfelt gratitude and appreciation to the panel of Examiners Dr. Mohammad Nurul Huda, Professor, Dept of CSE, UIU, for his constructive comments suggestions and critiquing.

I would like to express my deepest thanks to Dr. Dewan Md. Farid, Associate Professor, Dept of CSE, UIU, and Dr. Khondaker A. Mamun, Associate Professor, Dept of CSE, UIU for their favorable responses regarding the study.

Last but not least, I want to thank all the people who have willingly helped me out with their abilities.

Chapter 1

Introduction

This chapter will describe the importance and application of Deoxyribonucleic acid(DNA) binding protein and why we need to identify those, why we need automated predictor to identify DNA-binding protein. The motivation, Aims & Objectives, Methodology and our contribution are also stated here.

1.1 Problem Description

Proteins which make bond with DNA in both eukaryotes and prokaryotes while performing like activators or repressors are DNA Binding protein. DNA binding proteins play a vital roles in various cellular processes.They are essential in transcriptional regulation, recombination, genome rearrangements, replication, repair and DNA modification [26]. It has been observed that the percentage of prokaryotes and eukaryotes protein that can bind to DNA are only 2-3% and 4-5% [5, 6]. Prediction of DNA binding proteins are expensive and time consuming. Wide varieties of in vitro methods [7, 8] like filter binding assays, chromatin immunoprecipitation on microarrays (ChIP- chip) genetic analysis and X-ray crystallography are used to predict DNA binding proteins. Particularly, the number of newly discovered protein sequences has been increasing extremely fast. According to the Swiss-Prot [9] database contained only 3,939 protein sequence entries whereas the current number is 553941 which is more than 140 times from the 25 years ago. Due to the rapid growth of protein sequence, there is a urgent need of selecting a automated computational approach to identify DNA binding protein. Most of the computational methods used in the literature to predict DNA binding proteins

formulated the problem as a supervised learning problem. Practically, the number of known DNA-binding proteins is very small compared to the large non-binding proteins and unknown proteins. However, DNA-binding protein prediction is often modeled as a binary classification problem where given a protein sequence as input the task is to predict whether the protein is DNA-binding or not. Here we can notice that the challenge here is to select a proper dataset for training and testing incorporating the imbalanced situation.

1.2 Motivation

DNA Binding Proteins play important functions in regulation of chromatin structure by binding to short stretches of DNA. It has been noted during the sequencing of eukaryote genome that 10% of gene encode potential DNA Binding Protein.The interactions can be formed by different domains, such as the zinc finger or the helix-turn-helix. These interactions are involved in a variety of biological processes including DNA replication, DNA repair, viral infection, DNA packing, and DNA modifications. DNA Binding proteins have a fundamental role in the maintenance of genetic material and in the regulation of gene expression. It has many important applications in antibiotics, drugs, steroids for various biological effects and in bio-physical, bio-chemical and biological studies of DNA. So we expect that our research result will be helpful in this area.

1.3 Aims & Objectives

We get the followings objective from above aims:

- Finding out and preparing data set.

- Extracting the monogram and bigram features from Hidden Markov Model(HMM) profile.

- Using Support Vector Machine, AdaBoost and Random Forest as a classifier.

- Using the standard evaluation metrics to improve the feature used.

- Analyzing the data and the result.

- Developing a web server.

1.4 Methodology

According to Liu's [10] mentioned methodology, we followed the below procedures to establish a predictor for DNA-binding Protein identification with novel feature set.

- Selecting a benchmark dataset in order to train and test the predictor.

- Introducing or develop a powerful procedure to operate the prediction.

- Performing properly cross-validation tests to objectively evaluate the anticipated accuracy of the predictor.

1.5 Contribution

Our contribution in this study is listed below. In the first phase, a Benchmark dataset was selected. After that by our feature extraction method, we generate monogram and bigram features. Monogram and bigram features concatenated together are then used as training feature set to train the classifier. We used Support Vector Machine(SVM) with linear kernel as the classification algorithm and a trained model was stored for the testing phase.

1.6 Organization

The organization of this document is described below. In the first chapter, we discussed about problem description, Motivations, Aims & objectives, Methodology and Contributions. In the second chapter, we showed some biological preliminaries and literature review. In the third chapter we will focus on dataset, System Overview, feature extraction, feature selection method, StandardScaler, Cross Validation, description of the classifier and performance evaluation process. In the fourth chapter, we will present and discuss about all kinds of experiments and results which we got in this study. In the fifth chapter, we will discuss about summary of the study, limitations and future works.

Chapter 2

Background

In this chapter, we will discuss about some basic terminologies like Protein, DNA Structure, Structure of the protein, DNA Binding Protein etc. Then we will discuss about literature review of DNA Binding Protein.

2.1 Biological Preliminaries

2.1.1 DNA

The full abbreviation of DNA is deoxyribonucleic acid. A organic chemical of complex molecular structure that is available in all prokaryotic and eukaryotic cells and in many viruses. Genetic information are coded in DNA which is important for the transmission of inherited traits. Almost every person has same DNA of which most of them is located in the cell nucleus but a little amount of DNA can also be found in the microchondia. In DNA, information is stored in coded form which is made up of four chemical bases: adenine (A), guanine (G), cytosine (C), and thymine (T).

There are about 3 billion bases inhuman DNA among which 99 percent of those are same in all people. This sequence is helpful in determining the information for building the organism. Base pairs are formed like A with T and C with G. Each of the base is fixed to a sugar molecules and a phosphate molecule which together is called a nucleotide. Double helix are arranged in two long strands which is like a ladder. DNA can replicate itself which is important during cell division where each cell have same copy of the DNA presented in the old cell.

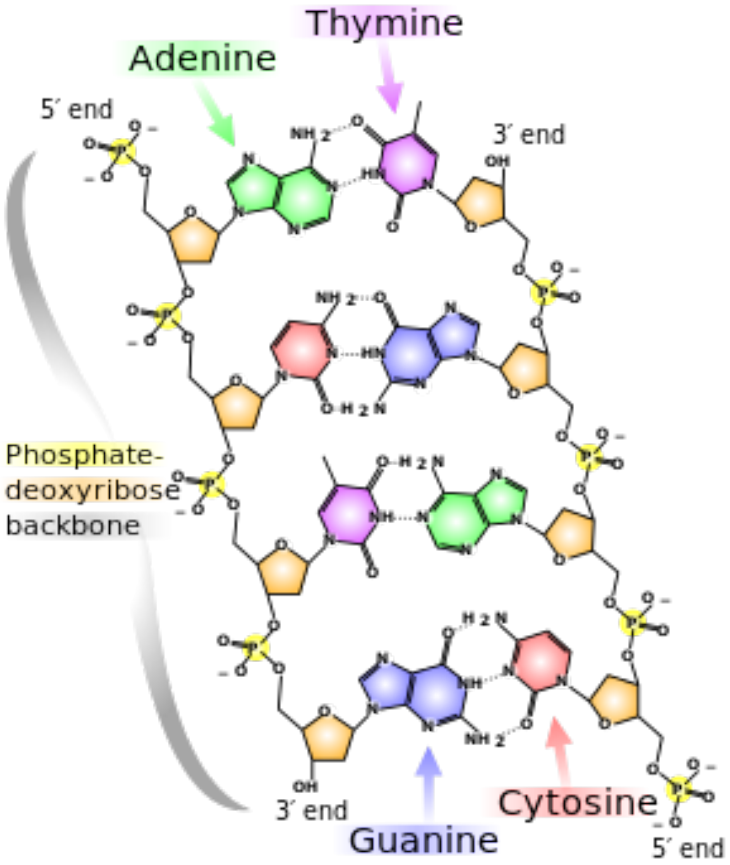


Figure 2.1: DNA Structure [1]

2.1.2 Protein

Exerting biological function in the nanometer scale is the task of proteins. They build the building blocks of all cell. Many important dynamic process of life like maintenance, replication, defense and reproduction are carried out by proteins. Proteins are made up of amino acids which joined together by chemical bonds and then folded in different ways to create three-dimensional structures for our body. There are two main type of amino acids in the body. First, the essential amino acids which our body can not manufacture and we must consume in our diet. Then other kind which our body can produce for itself. There are twenty natural amino acids which group together to form polypeptide chain or protein. The following task is carried out by protein our body:

- Because of Protein we can move.

- Protein can be the cause of killing us. Botox(Botulinum Toxin) is a protein which makes our muscle relax. The muscles which control our lungs stop working after taking food contaminated with botulism.

- Protein speed things up using enzymes that makes reactions go more quickly.

- Hemoglobin transport oxygen in our body.

2.1.3 Protein Structure

To understand the protein structure we have to know about the basic structure of amino acid. There is a carboxylic acid group(-COOH) on one side and an amino group on the other side. They may have the side chains which are unique to each indivdual amino acid. There are 20 different standard L-amino acids used by cells for protein construction. Biochemists often refer to four distinct aspects of a protein's structure. For levels of protein structure is shown in

i. Primary structure: The primary structure of a protein refers to covalent structure and sequence of amino acid of polypeptide chain. There are 20 amino acids. This are α amino acid. Amino acid is formed of amino group, carboxylic acid group and hydrogen bound to the central carbon atom. The R-group differs from amino acid to amino acid. The α carbon is tetrahedral and is bound to four different chemical moieties. For this reason, it is asymmetric (a chiral center) and there are two different enantiomers (D and L). Among them, glycine is somehow different whose R-group is a hydrogen atom which makes α carbon symmetric. Other amino acids are L-amino acids. A series of condensation reactions between the carboxyl group of one amino acid and the amino group of the next which are forming a covalent bond and a polypeptide chain. This is called an amide bond.

ii. Secondary structure: The secondary structures in proteins are the regions in which chains are arranged into regular structure which is known as alpha-helices and beta-pleated sheets. Hydrogen bonds hold them together.The diagram show that the form between lone pairs on an oxygen atom and the hydrogen which is attached to a nitrogen atom. The exact pattern of hydrogen bond is different in an alpha-helix and beta-pleated sheet. In an alpha-helix, the protein chain is coiled like a loosely-coiled spring. The meaning of "alpha" is if you look down the

length of the spring, the coiling is happening in a clockwise direction as it goes away from you. In a beta-pleated sheet, in order to lying alongside each other the chains are folded.

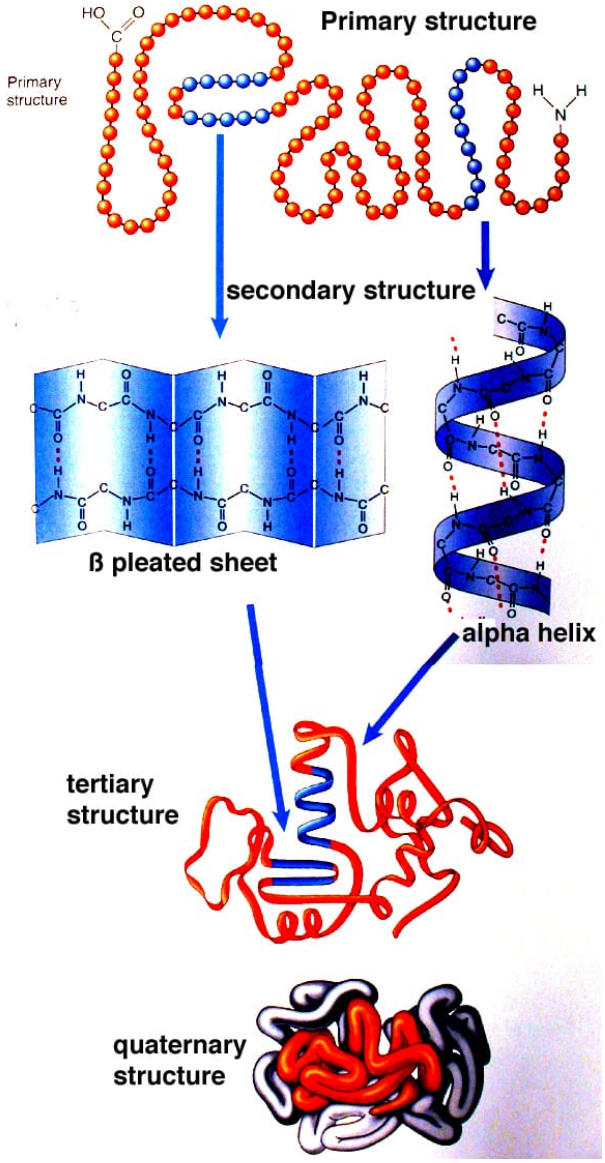


Figure 2.2: Levels of Protein Structure [2]

iii. Tertiary structure: Two or more Alpha Helix and/or Beta sheet further folded to form Tertiary Structure which is known as Globular protein. The whole chain

folds itself in its final three dimensional shape.The model shows the alpha-helices in the secondary structure as coils of "ribbon". The beta-pleated sheets end in an arrow head. The protein chain are random coils. This particular model has two molecules locked into it with whose reaction this enzyme catalyzes. The interactions between the side chains show of the "R" groups holds together the tertiary structure of a protein.

iv. Quaternary structure: Most of the proteins are homo- or hetero-oligomers or macromolecular assemblies or polymers. When two subunits (polypeptide chains) interact to form an interface, there are two possibilities like interface may be either isologous or heterologous. In the case of isologous, the association leads to the formation of a symmetric dimer possessing a two-fold (diad) axis. On the other hand in the case of heterologous, they may be symmetric.

2.1.4 Protein Synthesis

The process by which individual cells build their specific proteins is called Protein Synthesis. Within this, the process are involved in both DNA (deoxyribonucleic acid) and ribonucleic acids (RNA). The process is initiated in the cells nucleus, where specific enzymes unwind the needed section of DNA, which makes the DNA in this region accessible and a RNA copy can be made. This RNA molecule then moves from the nucleus to the cell cytoplasm, where the actual process of protein synthesis take place.There are two steps of protein synthesis: transcription and translation. The information which is encoded in DNA are taken in transcription step and encodes it into mRNA. In translation, the mRNA works with a ribosome and tRNA to synthesize proteins.

$$DNA \Rightarrow Transcription \Rightarrow Protein$$

In the first step of transcription, the DNA molecule partially unwinds in order to transcribe the DNA that codes for the needed protein. When the DNA molecule is unwound at correct location, an enzyme called RNA polymerase helps line up nucleotides to create a complementary strand of mRNA. As mRNA is a single-stranded molecule, only one of the two strands of DNA is used as a template for the new RNA strand.The new strand of RNA is made according to the rules of base pairing:

- DNA cytosine pairs with RNA guanine

- DNA guanine pairs with RNA cytosine

- DNA thymine pairs with RNA adenine

- DNA adenine pairs with RNA uracil

For example, the mRNA complement to the DNA sequence TTGCAC is AACGUG. After transcription, the new RNA strand is released so that two unzipped DNA strands bind together to form the double helix. Because the DNA template remains unchanged after transcription, it is possible to transcribe identical molecule of RNA when the first one is complete.

In translation process, mRNA is sent to the cytoplasm where make bonds with ribosomes.There are three binding sites in Ribosomes: one for mRNA and two for tRNA. The two tRNA sites are labeled the A site and P site. Translation begins when the mRNA chain makes binding with the ribosome. The first codon fills the P site and the second codon fills the A site. The tRNA molecule forms a temporary base pair with the mRNA in the A site. A peptide bond is formed between the amino acid attached to the tRNA in the A site and the methionine in the P site. The tRNA in the A site moves over to the P site so that a new codon fills the A site. The exact tRNA which carries the appropriate amino acid pairs bases with this new codon. Between the two adjacent amino acids held by tRNA molecules a peptide bond is formed. The tRNA which was in the P side will binds with amino acid.When one of the three codons enters the A site, the process stops. At that point, the protein chain which was connected to the tRNA in the P site is released. Then the translation process is complete.

2.1.5 DNA-binding Protein

The proteins that bind to single or double -stranded DNA is called DNA-binding proteins. Information that are coded for all cont traits are contained in DNA. There are thousands of proteins in our cells that help modulate DNA's functions. These proteins are called DNA-binding proteins, because they are attached to DNA.Proteins which make bond with DNA in both eukaryotes and prokaryotes while performing like activators or repressors are DNA Binding. It has been observed that the percentage of prokaryotes and eukaryotes protein that can bind to DNA are only 2-3% and 4-5% [5, 6].

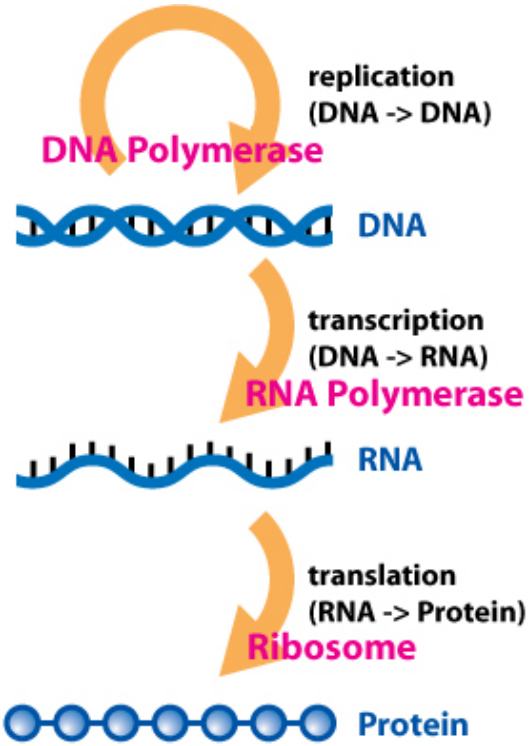


Figure 2.3: The Central Dogma of molecular biology : replication, transcription and translation.[3]

DNA binding proteins do many important task like controlling protein function, regulating cell growth and division and storing DNA inside the nucleus. DNA-binding proteins are the parts of the cell that read and interpret the instructions to create and maintain the body. Let's look at the different types of DNA-binding proteins and examples of each. DNA binding proteins play vital roles in various cellular processes. They are essential in transcriptional regulation, recombination, genome rearrangements, replication,repair and DNA modification [11].

2.1.6 Hidden Markov Model

Hidden Markov Model (HMM) is a statistical Markov model in which the system being modeled is taken as a Markov process with unobserved states. We call the observed event a 'symbol' and the invisible factor underlying the observation a 'state'. It can be represented as the simplest dynamic Bayesian network. L. E. Baum and his coworkers developed the mathematics behind the HMM. The hidden Markov model can be

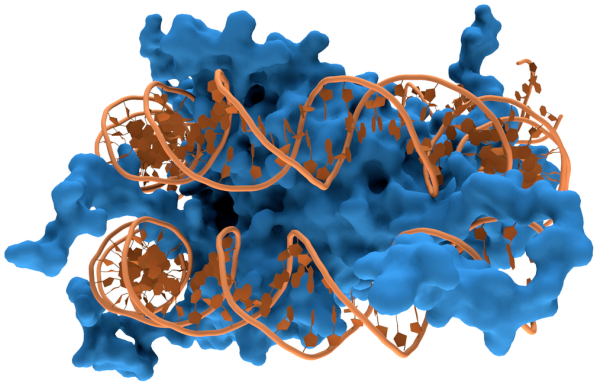


Figure 2.4: DNA Binding Protein.[4]

represented as the simplest dynamic Bayesian network. HMM consists of important parameters, such as,

-$\sum$: Set of emission characters.

- Q: Set of hidden states, each emitting symbols from $\sum$.

- a_{kl} :Q $\times$ Q matrix containing the probabilities of changing from state k to state l.

- e_{kb} :Q $\times$ $\sum$ matrix of probability of emitting symbol b while being in state k.

Hidden Markov models are especially known for their application in reinforcement learning and temporal pattern recognition such as speech, handwriting, gesture recognition, part-of-speech tagging, musical score following,[8] partial discharges[9] and bioinformatics.

2.2 Literature Review

Experimental techniques to predict DNA binding proteins are expensive and time consuming. Wide varieties of *in vitro* methods [7, 8] like filter binding arrays, chromatin immunoprecipitation on microarrays (ChIP- chip) genetic analysis and X-ray crystallography are used to predict DNA binding proteins.

Most of the computational methods used in the literature to predict DNA binding proteins formulated the problem as a supervised learning problem. Practically,

the number of known DNA-binding proteins is very small compared to the large non-binding proteins and unknown proteins. However, DNA-binding protein prediction is often modeled as a binary classification problem where given a protein sequence as input the task is to predict whether the protein is DNA-binding or not. Note that the challenge here is to select a proper dataset for training and testing incorporating the imbalanced situation. Many supervised learning algorithms have been used in the literature to solve the problem which are Artificial Neural Networks(ANN) [12], Support Vector Machines(SVM) [10, 13], ensemble methods [14], Naive Bayes classifier [15], Random Forest [16], Convolutional Neural Networks [17], Logistic Regression [18], AdaBoost Classifier [8] etc. Support Vector Machines (SVM) is one of the best performing classifiers used for DNA-binding protein identification than other algorithms because of robustness and correctness than other classifier [10, 13, 19, 20].

A great number of web based tools and methods are developed for DNA-binding protein prediction and are available for use. In this paper, we would like to mention several of them: DNABinder [13], DNA-Prot [21], iDNA-Prot [16], iDNA-Prot|dis [19], DBPPred [22], iDNAPro-PseAAC [10], PseDNA-Pro [23], Kmer1+ACC [24], Local-DPP [25], SVM-PSSM-DT [26], PNImodeler [27], CNNsite [17] and BindUP [28]. Most of these methods have used sequence, profile or structure based features. In structural feature based methods [18, 29, 30] in the literature, features used were structural motifs, electrostatic potential, the dipole moment, α-carbon only models. Sequence based methods [10, 19, 22, 25, 31–33] are often depended on the PSSM profile based information or pseudo-amino acid compositions. In [34], HMM based profiles were used for generating features for prediction of protein fold recognition and it got better accuracy than the traditional approach PSSM.

In this paper, we propose HMMBinder, a novel DNA-binding protein prediction tool using HMM profile based features of a protein sequence. Our method uses monogram and bigram features derived from the HMM profile which shows effectiveness compared to the PSSM or sequence based features. We used SVM as the classifier and standard benchmark datasets were used to test our method. Using the standard evaluation metrics our method significantly improves over the state-of-the-art methods and the features used in the literatures.

Chapter 3

Materials & Methods

In this chapter we will discuss about the techniques we used in our thesis which will be helpful to give a better understanding. We will describe about the datasets, system overview, feature extraction method, feature selection method, description of the classifier and performance evaluation.

3.1 Datasets

At first, we need a set of reliable benchmark datasets in order to develop an effective predictor. Any dataset S consists of positive $\mathbb{S}^+$ and negative $\mathbb{S}^-$ samples and can be formally denotes as following:

$$\mathbb{S} = \mathbb{S}^+ \cup \mathbb{S}^- \tag{3.1}$$

Here $\mathbb{S}^+$ represents the set of positive samples which is DNA Binding protein and $\mathbb{S}^-$ denotes the negative samples which is non DNA-binding proteins [19]. In this paper we use a popular benchmark dataset called benchmark1075 to train our model and to test the model using cross validation. Later we test the performance using cross validation on a separate independent test set known as independent186 dataset too. This section provides a brief overview of these two datasets [10, 19, 23, 25, 35].

3.1.0.1 Dataset benchmark1075

This dataset was first introduced in [19]. This dataset consists of 1075 protein sequences. Among them, 525 are DNA-binding and 550 are non DNA-binding protein sequences. All the protein sequences were taken from PDB [36]. This dataset is one

of the largest DNA-binding protein prediction datasets and thus suitable for training purpose. There are two criteria to construct a benchmark dataset: (1) Proteins which has a length less than 50 AA are removed (2) The proteins which have sequence similarity less than 25(%) were cut off in order to reduce redundancy. Using these criteria, 525 DNA binding protein and 550 non DNA binding protein were selected from PDB.

3.1.0.2 Dataset independent186

Lou et al. [22] constructed this independent dataset consisting 93 DNA-binding and 93 non DNA-binding protein sequences. They used BLASTCLUST [37] on the benchmark dataset to remove the sequences that have more that 25% of similarity to avoid homology bias.

Table 3.1: Summary of Benchmark Datasets explored in this research:

Benchmark	**ref**	**+ve protein**	**-ve protein**	**Total**
Independent dataset	[22]	93	93	186
Benchmark dataset	[36]	520	550	1070

3.2 System Overview

In this section, we provide the details of the materials and methods of this paper. Figure 3.1 provides a system diagram of our proposed method. For the training phase, all the protein sequences are fed to HHblits [38] which is a sequence alignment software using the latest UniProt database. HHblits produces HMM file as output which is then used by our feature extraction method to generate monogram and bigram features. Monogram and bigram features are concatenated together and then used as training feature set for the classifier. We use SVM with linear kernel and RBF kernel as the classification algorithm and a trained model is stored for the testing phase. This stored model is also used for the web-server implementation of HMMBinder. The pseudo code of our experiment is given below:

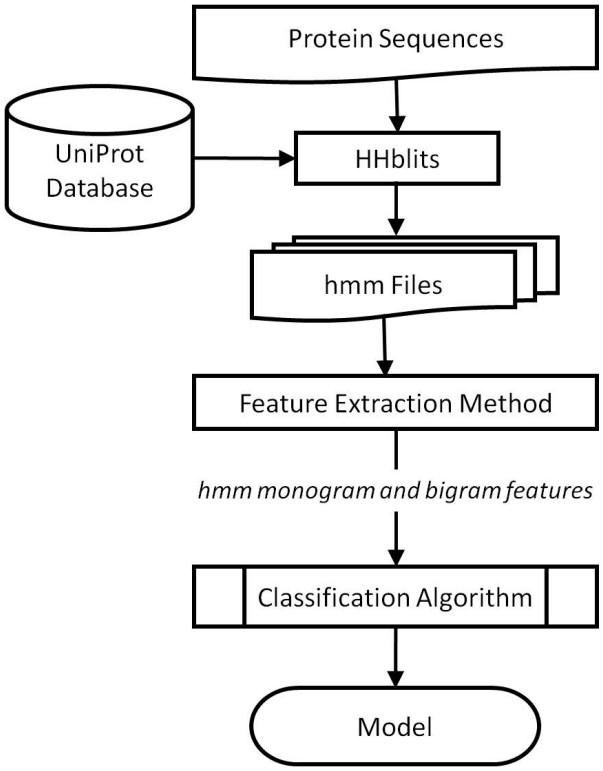


Figure 3.1: System diagram of HMMBinder

Step 1: Produce the HMM files from given protein sequence using HHblits.

Step 2: Extract the monogram and bigram features from the HMM files.

Step 3: Scale the feature using StandardScaler.

Step 4: Apply K-fold cross validation on the scaled data.

Step 4.1: For cross validation, at first we partition the data randomly into 10 equal segments.

Step 4.2: We use one segment as validation set and remaining 9 sets as learning the training data.

Step 4.3: Learning data is used to find the model parameters.

Step 4.4: Validation data is used to compute classification accuracy and preserve the the classification accuracy.

Step 4.5: Repeat the steps 1-4 for n times.

Step 4.6: Calculate the average classification accuracy for n times.

Step 5: Use the SVM linear kernel and SVM RBF kernel for classification purpose.

3.3 Feature Extraction

The training dataset $\mathbb{S}$ used for a binary classification problem consists of two types of instances: positive and negative. Formally,

$$\mathbb{S} = \mathbb{S}^+ \cup \mathbb{S}^- \tag{3.2}$$

Next, task is to represent each protein instance as feature vectors suitable for training. The idea is to represent each of the protein instances as a vector of features.

$$\mathbb{P} = [f_1, f_2, \cdots, f_n] \tag{3.3}$$

Here, a protein, $P \in \mathbb{S}$ is shown as a feature vector with dimension n=30. Most of the methods in the literature of DNA-binding protein prediction use either sequence and PSSM profile based features or structure based features. To the best of our knowledge there have been no application of features using HMM profiles. In this paper, we have used HHBlits [38] to generate HMM profiles. HMM profiles are comparatively more effective [38, 39] for remote homology detection. HMM profiles were generated using four iterations of HHBlits with a cutoff value set to 0.001 using the latest UniProt database [9]. HMM profiles are $L \times 20$ matrix produced by HHBlits where L is the length of protein sequence. These 20 values are the substitution probability of each type of amino-acid residue along the protein sequence at each position. These values are first converted to linear probabilities using the following formula:

$$p = 2^{\frac{-N}{1000}} \tag{3.4}$$

We generated two types of features–monogram and bigram–using the generated HMM profile matrix noted here as H. Previously this features were extracted from protein sequence and obtained satisfactory results. For protein sequence, monogram features were named after the occurrence of the amino acids feature group while bigram used to refer to as dipeptides.

Monogram and bigram features were extracted from HMM profile. For a protein sequence, the HMM profile produces a $L \times 30$ matrix where L is the length of protein sequence. Among the 30 columns, based on the position the first 20 columns shows

the substitution probability with the sequence. The last 10 columns denotes the three states namely insertion,deletion and match. The insertion denotes when an amino acid is inserted in a particular position and deletion refers when an amino acid is deleted from a particular position whereas match denotes when an exact match is found. To compare the effectiveness of PSSM and HMM for identifying DNA binding protein, we extract the features from the first 20 columns of the HMM profile. We provide a brief description of monogram and bigram features extracted from the HMM profile matrix.

3.3.0.1 Monogram Features

For a particular amino acid with all the other amino acids with the protein sequence, monogram feature group is calculated by global summation of the substitution probabilities for the first 20 columns of the HMM-profile. Monogram features [40] are calculated taking the normalized sum of the column wise substitution probability values. Size of these feature group is 20 because of 20 different amino acids. The feature can be defined formally as below:

$$HMM - Monogram(j) = \frac{1}{L}\sum_{i=1}^{L} H_{i,j} \tag{3.5}$$

Note that values of j is dependent on the columns, i.e., $1 \leq j \leq 20$. Here, H_{ij} is the values in the i^{th} row and j^{th} column of the matrix. We denote monogram features as M which is a vector of the form $M = [m(1), m(2), \cdots, m(20)]$

3.3.0.2 Bigram Features

The interaction among the neighboring amino acid is represented by bigram which represent the possibility of appearing of two consecutive amino acid along the protein sequence. Bigram features have been successfully used in the literature for protein attribute prediction [41]. Bigram features are normalized bigrams taken for all pair of columns. Hence the total number of features generated from this group is 400. Bigram features are generated using the following formula:

$$HMM - Bigram(j,k) = \frac{1}{L-1}\sum_{i=1}^{L} H_{i,j}H_{i+1,k} \tag{3.6}$$

Here j and k denote the column pairs for which the bigram is calculated and are in the range, $1 \leq j \leq 20$ and $1 \leq k \leq 20$. We denote this feature vector as B, where B has the form of $B = [b(1,1), b(1,2), \cdots, b(1,20), b(2,1), \cdots, b(20,20)]$

We also generate Positive Specific Scoring Matrix (PSSM) profiles for each of the protein sequences using PSI-BLAST [42]. PSSMs were generated using three iterations of PSI-BLAST using the nr database with a cutoff value of 0.001. PSSM profiles also have the similar form to HMM profiles which is a matrix of the same dimension and each of the matrix values denote substitution probabilities. We generate monogram and bigram features from PSSM files as well. These PSSM based monogram and bigram features are well used in the literature [40, 41, 43–46]. Note that all the monogram features are 20 dimensinal vector and bigram features are 400 dimensional vectors. We have also used a combination of the monogram and bigram features which is a 420 dimensional vector.

3.4 StandardScaler

In standardScaler we try scale the mean to 0 and variance to 1. In this way, we remove the mean and scale the data to the unit variance. Thus we standardize the features. It is the common requirement for any machine learning tool. If the data is not normally distributed, its behaviour may be unusual. The elements like RBF of support vector machine which are used in the objective function of a machine learning, usually consider that all the features have center in 0 and variance in the same order. If it gets other than that, the estimator may become unable to learn from other features.

3.5 Cross Validation

In residuals evaluations, the performance of a learner for a new prediction is totally unknown. The idea of the cross validation is not to use total dataset to train a learner. Some of the data are removed while training the learner. Then after the training, this removed data can be used to test the performance of the learned model. There are various cross validation like K-fold cross validation, leave one out cross validation, leave P Out, shuffle and split cross validation. Among them we used, Kfold cross validation with random shuffle. In K-fold cross validation, the samples are divided into k samples. Among them k-1 number of samples are used to train the model and last left fold is used to test the model. We have used 10 fold cross validation with random shuffle one 9 folds are used to train the model and the last sample are used to test the model.

Then in the next iteration, another 9 folds are used for training and remaining samples are used for testing. In this way, it is working.

3.6 Description of the Classifier

To investigate the performance of our proposed approaches, several classification techniques such as, Random Forest, AdaBoost and SVM are used in this study.

3.6.1 Support Vector Machine

We have used Support Vector Machines (SVM) as our classification technique. SVM is successfully used in protein attribute prediction in general [34, 43, 47] and particularly in DNA-binding protein prediction [10, 13]. SVM is maximum margin classifier that attempts to learn a hyperplane from the training samples that separates the positive and negative data points in a binary classification problem. The hyperplane that is selected is the one for which the separation width or the margin is maximum and the nature of the hyperplane depends on the kernel functions used. SVM generally tries to optimize a multiplier function that goes as the following:

$$L = \alpha argmax \sum_{j} \alpha_j - \frac{1}{2} \sum_{j,k} \alpha_j \alpha_k y_) j y_k \phi(\vec{x}_j . \vec{x}_k) \tag{3.7}$$

The prediction of a SVM classifier is defined as below:

$$h(\vec{x}) = sign(\sum_{j} \alpha_j y_j (\vec{x} . \vec{x}_j) - b) \tag{3.8}$$

Here the transformation of the data points by the function ϕ could be linear, polynomial or any other kernel functions. In this paper, we explored linear and Radial Basis Function(RBF) kernels. Linear kernel is in the following form:

$$K(\vec{x}_i, \vec{x}_j) = (\vec{x}_i . \vec{x}_j + c)^d \tag{3.9}$$

Here $d = 1$ for the linear kernels. RBF kernels follow the following definition:

$$K(\vec{x}_i, \vec{x}_j) = e^{(-)\frac{||\vec{x}_i - \vec{x}_j||^2}{2\sigma^2}} \tag{3.10}$$

3.6.2 Random Forest:

Random Forest is a classification technique by ensemble learning method. Random forests [48] is the combination of bootstrap aggregating and random feature selection techniques so that it can produce several tree predictors.The decision is a combination of the predictions by the tree classifiers.

3.6.3 AdaBoost:

AdaBoost technique is introduced by Freund and Schapire[49] for multi class classification research purpose.It's create strong classifier from the ensemble of different weak classifier[50]. It is a machine learning meta-algorithm. For the reason of enhancing the accuracy, it may be used combined with other learning algorithm.

3.7 Performance Evaluation

A good number of effective evaluation metrics have been suggested for use in single valued and multi-valued classification and prediction [51, 52]. In the literature of DNA-binding protein prediction, we have found that the most widely used metrics are: accuracy, sensitivity, specificity, MCC, auROC, and auPR values. In this section, we first provide a description of these evaluation metrics used in this paper.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3.11}$$

Accuracy is the ratio or percentage of correctly classified negative or positive instances from a given number of protein instances. Here, TP is the total number of true positives or correctly classified positive samples and TN is the correctly classified negative samples. FP and FN are incorrectly classified positive and negative instances respectively. Sensitivity is the true positive rate or the ratio of true positives to the total number of positive examples. Sensitivity is defined in the following equation:

$$Sensitivity = \frac{TP}{TP + FN} \tag{3.12}$$

Specificity on the other hand is the true negative rate and can be defined as the following equation:

$$Specificity = \frac{TN}{TN + FP} \tag{3.13}$$

All these three measures have a maximum value of 1 which is the best classifier and a minimum value of 0 meaning the worst classifier. Mathew's Correlation Coefficient (MCC) denotes how good a binary classification is working. The value of MCC is in the range $[-1, +1]$. A perfect classifier should have a maximum MCC value of $+1$. MCC is defined as the following equation:

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{3.14}$$

Note that all these metrics for probabilistic outputs depends on the threshold set for the classifiers. Two other metrics not dependent on thresholds are area under Receiver Operating characteristic curve (auROC) and area under precision recall curve (auPR). The value of auROC and auPR has maximum value of 1 for the perfect classifier. ROC curve plots true positive rate against false positive rate at different threshold values and Precision-recall curve plots precision against recall.

To reduce the training bias several sampling methods are proposed in the literature [53] and widely used for protein attribute prediction [52]. In this paper, we have used 10-fold cross validation and jack-knife tests which are widely used in the literature of DNA-binding protein prediction [10, 16, 19, 22].

Chapter 4

Result & Discussion

In this chapter, we present the results of the experiments that were carried in this study. All the methods were implemented in Python3.4 programming language. The Scikit-learn library [54] of Python was used for implementing the machine learning algorithms. All experiments were conducted on computing services provided by CITS, United International University.

4.0.1 Experimental Design

We have run a number of experiments to test the effectiveness of the HMM profile based features on the benchmark dataset. We have six group of features extracted for this experiment: PSSM-Monogram, PSSM-Bigram, PSSM-Mono+Bigram, HMM-Monogram, HMM-Bigram and HMM-Mono+Bigram. Each of these feature sets are tested with SVM classifiers using linear and RBF kernels. We further tested the performances of these features using two ensemble classifiers: Random Forest and AdaBoost Classifiers. For these experiments we have performed 10-fold cross validation. The results in terms of accuracy, sensitivity, specificity, auPR, auROC and MCC are reported. Only the average of these values are reported in Table 4.1. Note that the best results were found using the combination of HMM-Monogram and Bigram features and 82.87% accuracy was achieved using SVM linear kernels. In each cases of the SVM linear kernel, HMM based features achieved better accuracy compared to PSSM based features. Similar results could be noticed for auROC, MCC and sensitivity analysis. Specificity, auROc and auPR is slightly improved in the experiments with SVM with RBF kernels. We

also show the ROC curves for each of these experiments in Figure 4.1, Figure 4.2, and Figure 4.3.

4.0.2 Comparison with Other Methods

We have compared the performance of HMMBinder with several previous methods and tools used for DNA-binding protein prediction on the benchmark dataset *benchmark1075*. They are: DNABinder [13], DNA-Prot [21], iDNA-Prot [16], iDNA-ProtDNA [19], DBPPred [22], iDNAPro-PseAAC [10], PseDNA-Pro [23], Kmer1+ACC [24] and Local-DPP [25]. The results reported in this study for these methods are taken from [10, 25]. The comparisons were made in terms of accuracy, sensitivity, specificity, MCC and auROC. To make a fair comparison with the other methods, we performed jack knife test as done in earlier studies and the results are reported in Table 4.2.

The best values in Table 4.2 are shown in bold faced fonts. The results shows a clear margin of more than 7% improvement of accuracy over the previous best method Local-DPP [25]. Similar improvements were found in other metrics too. Specially, MCC is increased by 22% compared to the previous best method.

We further experimented to test the effectiveness of HMMBinder on the independent test set also. These results are shown in Table 4.3. Here the results are not the best but among the best. In terms of accuracy, our results are almost similar to iDNAPro-PseAAC [10]. Their results were significant in the benchmark dataset and performed similar to ours in the independent dataset. Specificity value of HMMBinder was among the best and only second to DNA-threader which failed miserably in terms of accuracy. Considering the difficulty level of the independent dataset, we believe that our method has not over trained on the benchmark dataset and the performance is promising and can be claimed as a generalized method after training and testing. Based on these results, we decided to build the web application based on the model trained on the benchmark dataset.

Note that the results on the independent dataset is comparative but not improved in comparison to the state-of-the-art methods. The main focus of this research was to build a classifier based on HMM profiles instead of the PSSM profile based features and we experimentally showed the effectiveness of the HMM profile based features over PSSM. In future, we aim to focus on the independent dataset to perform better.

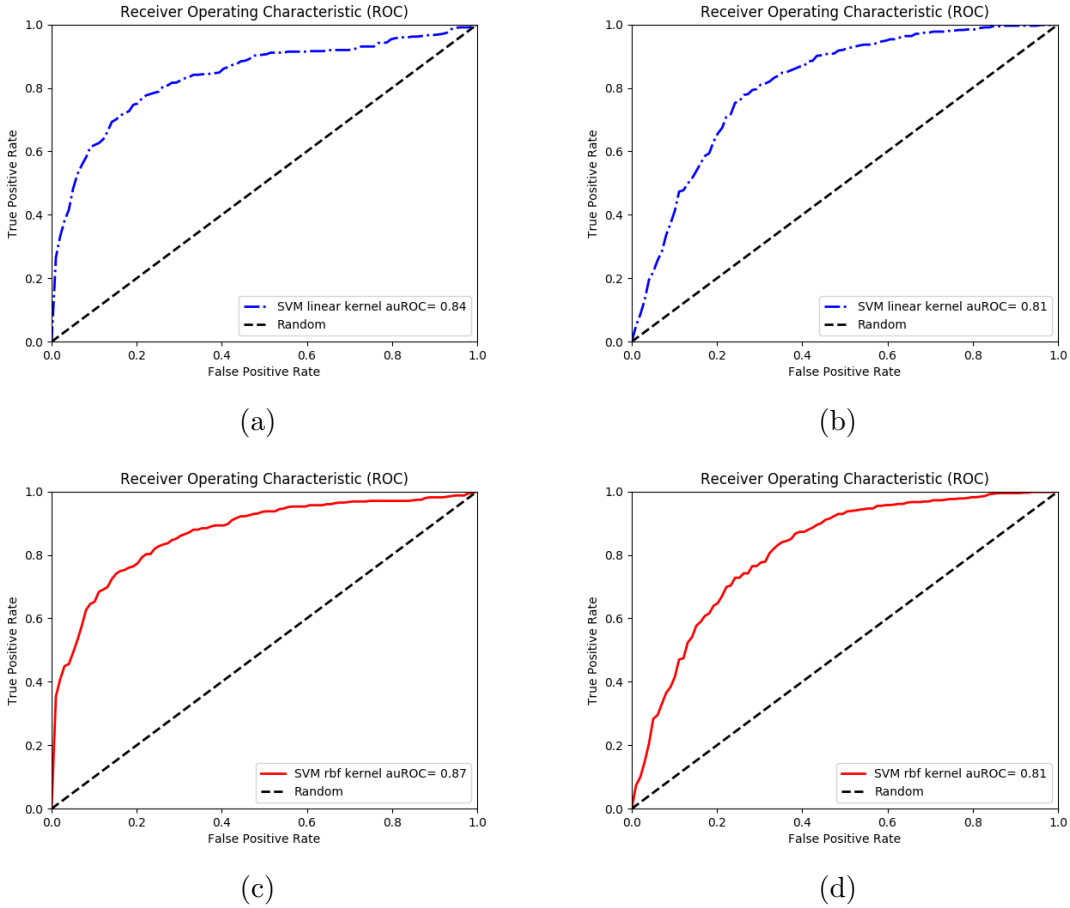


Figure 4.1: Using Monogram Features. Receiver operating characteristic curves for (a) SVM Linear kernel classifier using HMM Monogram features, (b) SVM Linear kernel classifier using PSSM Monoram features, (c) SVM RBF kernel classifier using HMM Monogram features and(d) SVM RBF kernel classifier using PSSM Monogram features.

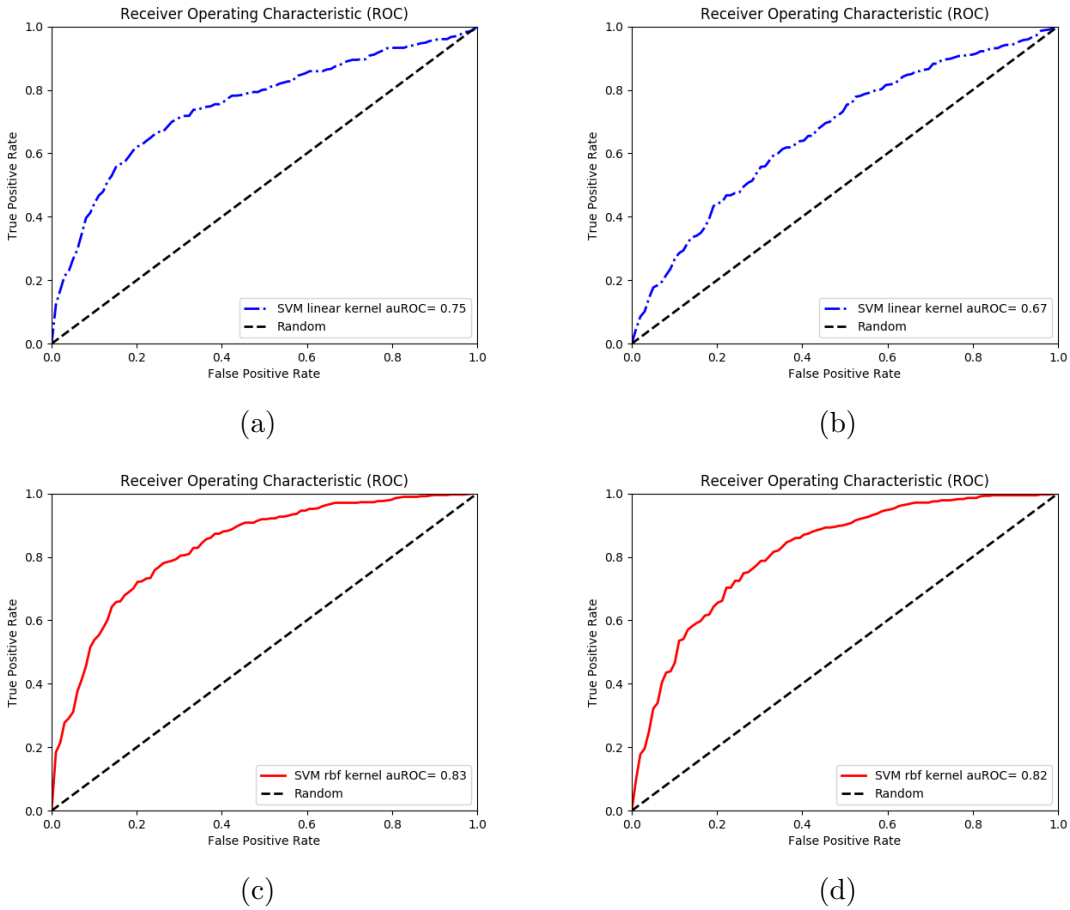


Figure 4.2: Using Bigram Features. Receiver operating characteristic curves for (a) SVM Linear kernel classifier using HMM Bigram features, (b) SVM Linear kernel classifier using PSSM Bigram features, (c) SVM RBF kernel classifier using HMM Bigram features and (d) SVM RBF kernel classifier using PSSM Bigram features.

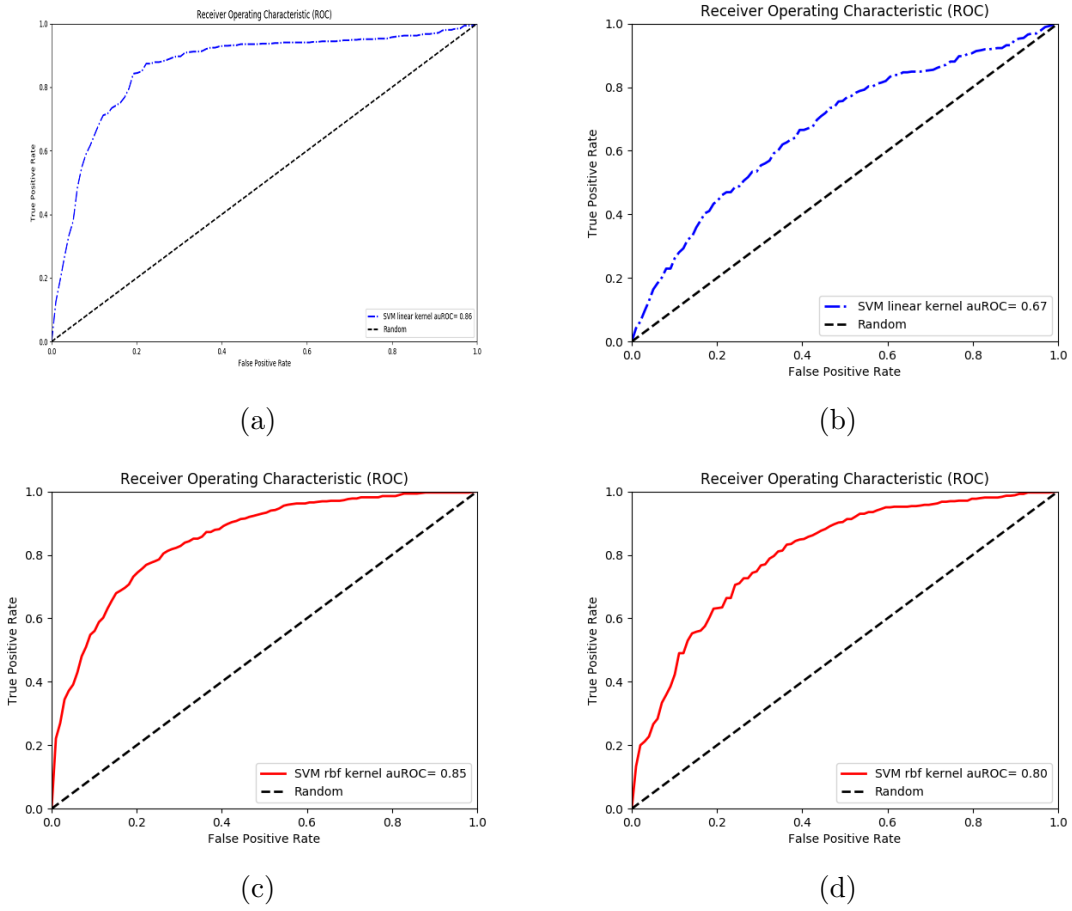


Figure 4.3: Using (Mono+Bi)gram Features. Receiver operating characteristic curves for (a) SVM Linear kernel classifier using HMM Mono+Bigram features, (b) SVM Linear kernel classifier using PSSM Mono+Bigram features, (c) SVM RBF kernel classifier using HMM Mono+Bigram features and (d) SVM RBF kernel classifier using PSSM Mono+Bigram features.

Additionally, we would like to highlight two points. Firstly, the datasets that we used were filtered using BLASTCLUST. It is important to remove the sequences with similarity more than 25% from the dataset before applying the training and testing methods. We used the dataset proposed by Lou et al. [22], a widely accepted standard independent test dataset where the sequences with similarity 25% or more with other sequences had been removed. We believe it would be interesting to see the effects of the other heuristic CLUSTALW [55]. Secondly, feature selection methods are gaining much popularity in case of bioinformatics data and supervised machine learning. We believe that using sophisticated feature selection methods, such as maximum relevance minimum redundancy(mRMR) [56] and maximum relevance maximum distance (MRMD) [57] could improve the results further.

4.0.3 HMMBinder Web Server Implementation

We have implemented a web based application based on the proposed method. We call this HMMBinder. This is readily available to use from the url at: `brl.uiu.ac.bd/HMMBinder`. The server was implemented using PHP web programming language in the front end and python based prediction engine at the backend. The software requires an HMM profile as input to the tools that can be generated by HHBlits. The feature are extracted automatically by the python program and the predicted value from a trained model is shown in the web form. The web site contains a 'read me' guide and the necessary information required to run the application. Some screenshots from our webserver is given below:

4.1 Data and Material Availability

All the data and materials which were used in this paper are available at: `brl.uiu.ac.bd/HMMBinder/`

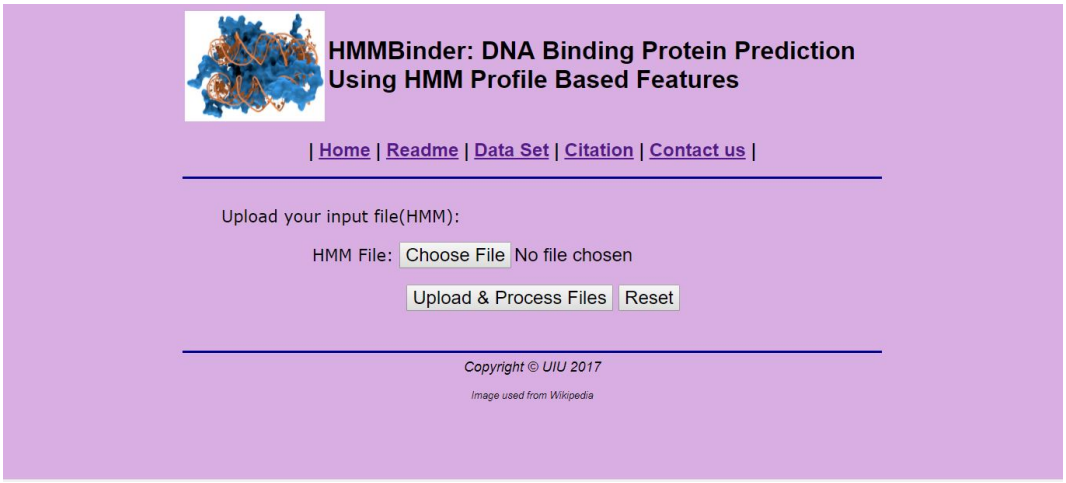


Figure 4.4: Screen shot of Web-Server homepage

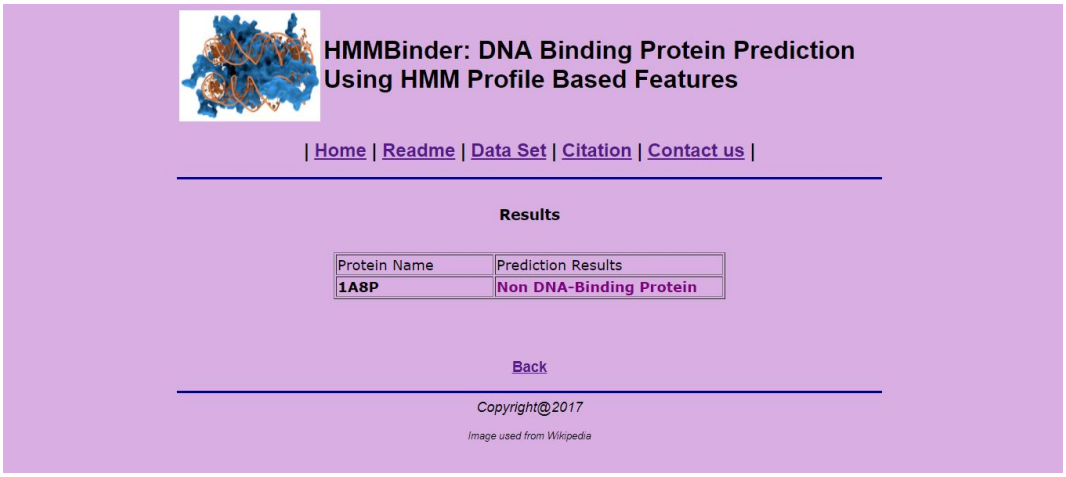


Figure 4.5: Screen shot of Web-Server Result

Features	Accuracy	Sensitivity	Specificity	auPR	MCC	auROC
SVM with linear kernel						
HMM-Monogram	76.77%	**0.8420**	0.6976	0.6931	0.5367	0.8358
PSSM-Monogram	74.74%	0.6636	0.8362	0.8368	0.5040	0.8105
HMM-Bigram	70.59%	0.7071	0.7049	0.7060	0.4095	0.7511
PSSM-Bigram	62.20%	0.6454	0.5973	0.6025	0.2502	0.6703
HMM(Mono+Bi)	**82.87%**	0.8150	**0.8415**	**0.8428**	**0.6538**	**0.8639**
PSSM(Mono+Bi)	72.40%	0.7364	0.7120	0.7136	0.4486	0.8028
SVM with rbf kernel						
Features	**Accuracy**	**Sensitivity**	**Specificity**	**auPR**	**MCC**	**auROC**
HMM-Monogram	**78.83%**	**0.8227**	0.7559	0.7535	**0.5761**	**0.8667**
PSSM-Monogram	73.71%	0.6890	0.7880	0.7903	0.4771	0.8121
HMM-Bigram	76.68%	0.7052	**0.8251**	0.8253	0.5283	0.8318
PSSM-Bigram	74.92%	0.7490	0.7495	0.7516	0.4966	0.8166
HMM(Mono+Bi)	77.43%	0.7129	0.8324	**0.8329**	0.5440	0.8496
PSSM(Mono+Bi)	72.40%	0.7363	0.7120	0.7136	0.4486	0.8028
Random Forest						
Features	**Accuracy**	**Sensitivity**	**Specificity**	**auPR**	**MCC**	**auROC**
HMM-Monogram	**74.44%**	**0.7938**	**0.6976**	**0.6936**	**0.4871**	0.8243
PSSM-Monogram	66.14%	0.7290	0.5895	0.5862	0.3173	0.7332
HMM-Bigram	72.19%	0.7553	0.6903	0.6880	0.4400	**0.8273**
PSSM-Bigram	71.00%	0.7854	0.6300	0.6305	0.4174	0.7833
HMM(Mono+Bi)	74.43%	**0.7938**	**0.6976**	0.6931	**0.4871**	0.8218
PSSM(Mono+Bi)	72.68%	0.7909	0.6589	0.6645	0.4557	0.7698
AdaBoost						
Features	**Accuracy**	**Sensitivity**	**Specificity**	**auPR**	**MCC**	**auROC**
HMM-Monogram	73.31%	0.7013	0.7632	0.7603	0.4579	0.8026
PSSM-Monogram	67.07%	0.7654	0.5703	0.5737	0.3448	0.7157
HMM-Bigram	73.97%	0.7360	0.7432	0.7396	0.4762	0.8063
PSSM-Bigram	70.53%	0.7436	0.6647	0.6708	0.4116	0.7710
HMM(Mono+Bi)	**78.00%**	**0.78.03**	**0.7795**	**0.7732**	**0.5532**	**0.8577**
PSSM(Mono+Bi)	70.07%	0.7327	0.6666	0.6687	0.4005	0.7887

Table 4.1: Comparison of performance of different features and SVM kernels on the benchmark dataset using 10-fold cross validation.

Method	Accuracy	Sensitivity	Specificity	MCC	auROC
iDNAPro-PseAAC	76.76%	0.7562	0.7745	0.53	0.8392
DNAbinder (dimension 21)	73.95%	0.6857	0.7909	0.48	0.8140
DNAbinder (dimension 400)	73.58%	0.6647	0.8036	0.47	0.8150
DNA-Prot	72.55%	0.8267	0.5976	0.44	0.7890
iDNA-Prot	75.40%	0.8381	0.6473	0.50	0.7610
iDNA-Prot|dis	77.30%	0.7940	0.7527	0.54	0.8310
PseDNA-Pro	76.55%	0.7961	0.7363	0.53	–
Kmer1+ACC	75.23%	0.7676	0.7376	0.50	0.8280
Local-DPP	79.20%	0.8400	0.7450	0.59	–
HMMBinder	**86.33%**	**0.8707**	**0.8555**	**0.72**	**0.9026**

Table 4.2: Comparison of performance of the proposed method with other state-of-the-art predictors using jack knife test on the benchmark dataset.

Method	Accuracy	Sensitivity	Specificity	MCC	auROC
iDNAPro-PseAAC	69.89%	0.7741	0.6237	0.402	0.7754
iDNA-Prot	67.20%	0.6770	0.6670	0.344	–
DNA-Prot	61.80%	0.6990	0.5380	0.240	–
DNAbinder	60.80%	0.5700	0.6450	0.216	0.6070
DNABIND	67.70%	0.6670	0.6880	0.355	0.6940
DNA-Threader	59.70%	0.2370	**0.9570**	0.279	–
DBPPred	76.90%	0.7960	0.7420	0.538	0.7910
iDNA-Prot|dis	72.00%	0.7950	0.6450	0.445	**0.7860**
Kmer1+ACC	70.96%	0.8279	0.5913	0.431	0.7520
Local-DPP	**79.00%**	**0.9250**	0.6560	**0.625**	–
HMMBinder	69.02%	0.6153	0.7634	0.394	0.6324

Table 4.3: Comparison of performance of the proposed method with other state-of-the-art predictors on the independent dataset.

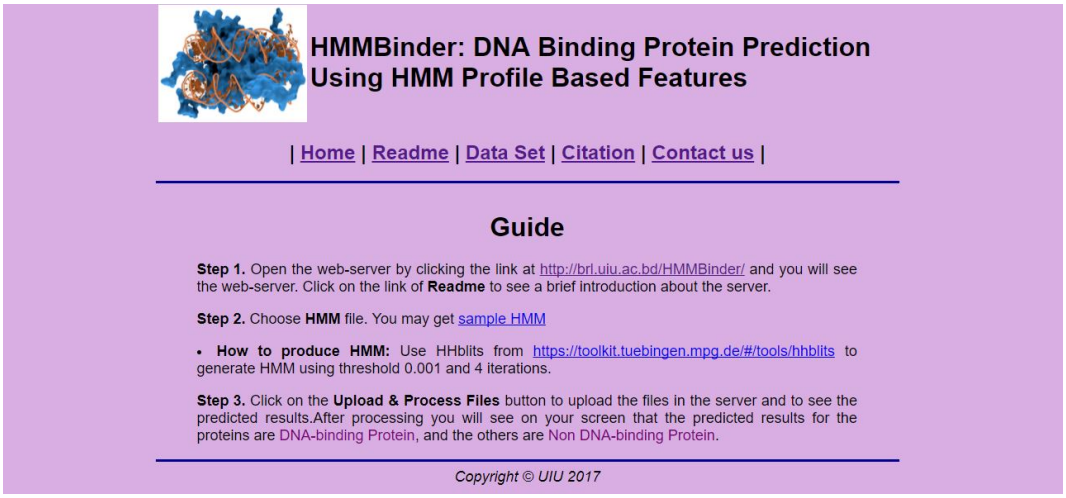


Figure 4.6: Figure 4.10: Screen shot of Web-Server Datasets.

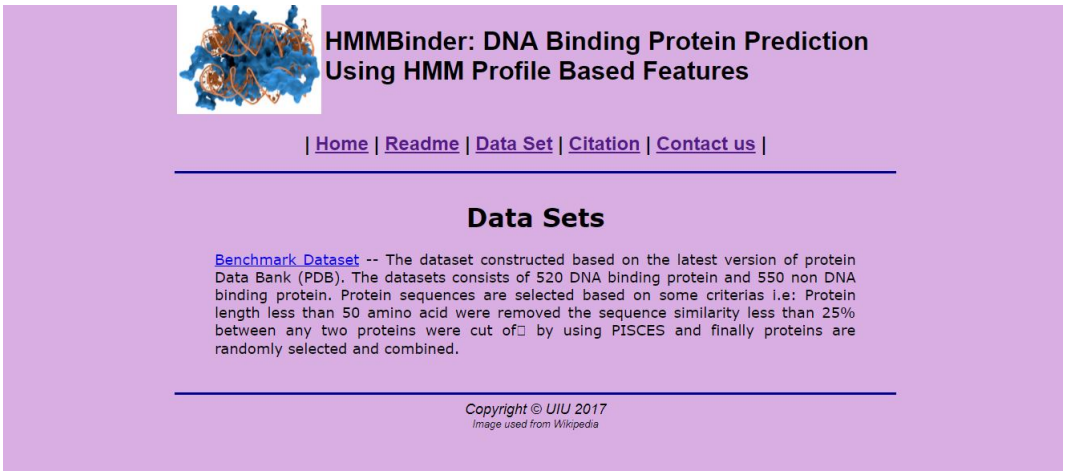


Figure 4.7: Screen shot of Web-Server Guide or Readme

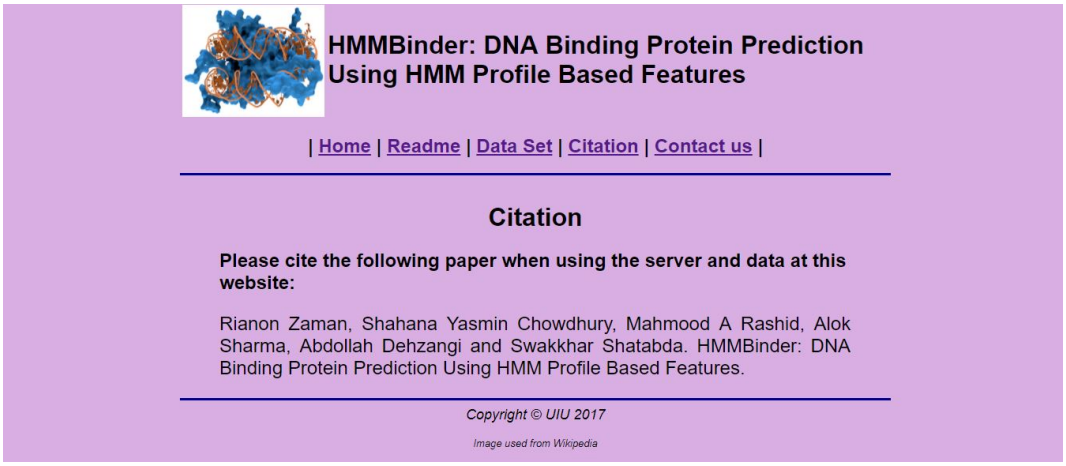


Figure 4.8: Screen shot of Journal citation

Chapter 5

Conclusion and Future Work

5.1 Conclusion

In our paper, we have introduced HMMBinder which is a HMM profile based method for the DNA-binding protein prediction problem. We have used monogram and bigram features extracted from the HMM profiles generated by HHblits and a SVM classification algorithm to train our data on a standard benchmark dataset. Our method is able to make considerable improvement over the other state-of-the-art methods on this dataset and performed comparably well in the independent dataset. We have also established a web based for our method that is trained on the benchmark dataset. As we used 10-fold cross validation, our sample data is randomly shuffled and averaged. So there is no chance of biased result.

5.2 Limitations

We didn't get expected results on independent dataset. Moreover, We didn't invent any feature selection method which will improve our performance. Our web server doesn't work properly during the predictions of all DNA Binding Protein and Non DNA Binding Protein.

5.3 Future Work

In future, we would like to extract more effective features and generate larger dataset to train our model. Besides, we will try to improve our results on the independent

dataset.

Bibliography

[1] L. J. R. M. R. K. W. P. Alberts B, Johnson A, "Dna," *https://en.wikipedia.org/wiki/DNA*, 2014. iv, 5

[2] "Protein structure," *https://schoolworkhelper.net/protein-structures-primary-secondary-tertiary-quaternary*, 2012. iv, 7

[3] F. Crick, "Central dogma of molecular biology," *Nature*, vol. 227, no. 5258, 1970. iv, 10

[4] S. R. Pabo CO, "Protein-dna recognition," *Biochem*, vol. 53, no. 1, 1984. iv, 11

[5] M.-G. Y. Stawiski EW, Gregoret LM, "Annotating nucleic acid-binding function based on protein structure," *J Mol Biol*, vol. 326, pp. 1065–1079, 2003. 1, 9

[6] J. S. T. J. Shanahan HP, Garcia MA, ""identifying dnabinding proteins using structural motifs and the electrostatic potential"," *Nucleic Acids Res*, vol. 32, pp. 4732–4741, 2004. 1, 9

[7] R. Jaiswal, S. K. Singh, D. Bastia, and C. R. Escalante, "Crystallization and preliminary x-ray characterization of the eukaryotic replication terminator reb1–ter dna complex," *Acta Crystallographica Section F: Structural Biology Communications*, vol. 71, no. 4, pp. 414–418, 2015. 1, 11

[8] R. E. Langlois and H. Lu, "Boosting the prediction and understanding of dna-binding domains from sequence," *Nucleic acids research*, p. gkq061, 2010. 1, 11, 12

[9] U. Consortium *et al.*, "Uniprot: the universal protein knowledgebase," *Nucleic acids research*, vol. 45, no. D1, pp. D158–D169, 2017. 1, 16

[10] B. Liu, S. Wang, and X. Wang, "Dna binding protein identification by combining pseudo amino acid composition and profile-based protein representation," *Scientific reports*, vol. 5, p. 15479, 2015. 3, 12, 13, 19, 21, 23

[11] B. H. T. J. Luscombe NM1, Austin SE, ""an overview of the structures of protein-dna complexes"," *Genome Biol*, vol. 1, 2000. 10

[12] S. Ahmad, 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*, vol. 20, no. 4, pp. 477–486, 2004. 12

[13] M. Kumar, M. M. Gromiha, and G. P. Raghava, "Identification of dna-binding proteins using support vector machines and evolutionary profiles," *BMC bioinformatics*, vol. 8, no. 1, p. 463, 2007. 12, 19, 23

[14] L. Song, D. Li, X. Zeng, Y. Wu, L. Guo, and Q. Zou, "ndna-prot: identification of dna-binding proteins based on unbalanced classification," *BMC bioinformatics*, vol. 15, no. 1, p. 298, 2014. 12

[15] C. Yan, M. Terribilini, F. Wu, R. L. Jernigan, D. Dobbs, and V. Honavar, "Predicting dna-binding sites of proteins from amino acid sequence," *BMC bioinformatics*, vol. 7, no. 1, p. 262, 2006. 12

[16] W.-Z. Lin, J.-A. Fang, X. Xiao, and K.-C. Chou, "idna-prot: identification of dna binding proteins using random forest with grey model," *PloS one*, vol. 6, no. 9, p. e24756, 2011. 12, 21, 23

[17] J. Zhou, Q. Lu, R. Xu, L. Gui, and H. Wang, "Cnnsite: Prediction of dna-binding residues in proteins using convolutional neural network with sequence features," in *Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on*. IEEE, 2016, pp. 78–85. 12

[18] A. Szilágyi and J. Skolnick, "Efficient prediction of nucleic acid binding function from low-resolution protein structures," *Journal of molecular biology*, vol. 358, no. 3, pp. 922–933, 2006. 12

[19] B. Liu, J. Xu, X. Lan, R. Xu, J. Zhou, X. Wang, and K.-C. Chou, "idna-prot—dis: identifying dna-binding proteins by incorporating amino acid distance-pairs and reduced alphabet profile into the general pseudo amino acid composition," *PloS one*, vol. 9, no. 9, p. e106691, 2014. 12, 13, 21, 23

[20] Y. Fang, Y. Guo, Y. Feng, and M. Li, "Predicting dna-binding proteins: approached from chous pseudo amino acid composition and other specific sequence features," *Amino acids*, vol. 34, no. 1, pp. 103–109, 2008. 12

[21] K. K. Kumar, G. Pugalenthi, and P. Suganthan, "Dna-prot: identification of dna binding proteins from protein sequence information using random forest," *Journal of Biomolecular Structure and Dynamics*, vol. 26, no. 6, pp. 679–686, 2009. 12, 23

[22] W. Lou, X. Wang, F. Chen, Y. Chen, B. Jiang, and H. Zhang, "Sequence based prediction of dna-binding proteins based on hybrid feature selection using random forest and gaussian naive bayes," *PLoS One*, vol. 9, no. 1, p. e86703, 2014. 12, 14, 21, 23, 27

[23] B. Liu, J. Xu, S. Fan, R. Xu, J. Zhou, and X. Wang, "Psedna-pro: Dna-binding protein identification by combining chous pseaac and physicochemical distance transformation," *Molecular Informatics*, vol. 34, no. 1, pp. 8–17, 2015. 12, 13, 23

[24] Q. Dong, S. Wang, K. Wang, X. Liu, and B. Liu, "Identification of dna-binding proteins by auto-cross covariance transformation," in *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on.* IEEE, 2015, pp. 470–475. 12, 23

[25] L. Wei, J. Tang, and Q. Zou, "Local-dpp: An improved dna-binding protein prediction method by exploring local evolutionary information," *Information Sciences*, vol. 384, pp. 135–144, 2017. 12, 13, 23

[26] R. Xu, J. Zhou, H. Wang, Y. He, X. Wang, and B. Liu, "Identifying dna-binding proteins by combining support vector machine and pssm distance transformation," *BMC systems biology*, vol. 9, no. 1, p. S10, 2015. 12

[27] J. Im, N. Tuvshinjargal, B. Park, W. Lee, D.-S. Huang, and K. Han, "Pnimodeler: web server for inferring protein-binding nucleotides from sequence data," *BMC genomics*, vol. 16, no. 3, p. S6, 2015. 12

[28] I. Paz, E. Kligun, B. Bengad, and Y. Mandel-Gutfreund, "Bindup: a web server for non-homology-based prediction of dna and rna binding proteins," *Nucleic acids research*, vol. 44, no. W1, pp. W568–W574, 2016. 12

[29] H. P. Shanahan, M. A. Garcia, S. Jones, and J. M. Thornton, "Identifying dna-binding proteins using structural motifs and the electrostatic potential," *Nucleic Acids Research*, vol. 32, no. 16, pp. 4732–4741, 2004. 12

[30] G. Nimrod, M. Schushan, A. Szilágyi, C. Leslie, and N. Ben-Tal, "idbps: a web server for the identification of dna binding proteins," *Bioinformatics*, vol. 26, no. 5, pp. 692–693, 2010. 12

[31] R. Xu, J. Zhou, B. Liu, Y. He, Q. Zou, X. Wang, and K.-C. Chou, "Identification of dna-binding proteins by incorporating evolutionary information into pseudo amino acid composition via the top-n-gram approach," *Journal of Biomolecular Structure and Dynamics*, vol. 33, no. 8, pp. 1720–1730, 2015. 12

[32] Y. Fang, Y. Guo, Y. Feng, and M. Li, "Predicting dna-binding proteins: approached from chous pseudo amino acid composition and other specific sequence features," *Amino acids*, vol. 34, no. 1, pp. 103–109, 2008.

[33] X.-W. Zhao, X.-T. Li, Z.-Q. Ma, and M.-H. Yin, "Identify dna-binding proteins with optimal chou's amino acid composition," *Protein and peptide letters*, vol. 19, no. 4, pp. 398–405, 2012. 12

[34] J. Lyons, A. Dehzangi, R. Heffernan, Y. Yang, Y. Zhou, A. Sharma, and K. Paliwal, "Advancing the accuracy of protein fold recognition by utilizing profiles from hidden markov models," *IEEE transactions on nanobioscience*, vol. 14, no. 7, pp. 761–772, 2015. 12, 19

[35] B. Liu, L. Fang, F. Liu, X. Wang, J. Chen, and K.-C. Chou, "Identification of real microrna precursors with a pseudo structure status composition approach," *PloS one*, vol. 10, no. 3, p. e0121501, 2015. 13

[36] H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, "The protein data bank, 1999–," in *International Tables for Crystallography Volume F: Crystallography of biological macromolecules*. Springer, 2006, pp. 675–684. 13, 14

[37] I. Dondoshansky and Y. Wolf, "Blastclust (ncbi software development toolkit)," *NCBI, Bethesda, Md*, 2002. 14

[38] M. Remmert, A. Biegert, A. Hauser, and J. Söding, "Hhblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment," *Nature methods*, vol. 9, no. 2, pp. 173–175, 2012. 14, 16

[39] D. B. Kuchibhatla, W. A. Sherman, B. Y. Chung, S. Cook, G. Schneider, B. Eisenhaber, and D. G. Karlin, "Powerful sequence similarity search methods and in-depth manual analyses can identify remote homologs in many apparently orphan viral proteins," *Journal of virology*, vol. 88, no. 1, pp. 10–20, 2014. 16

[40] Y. Taguchi and M. M. Gromiha, "Application of amino acid occurrence for discriminating different folding types of globular proteins," *BMC bioinformatics*, vol. 8, no. 1, p. 404, 2007. 17, 18

[41] A. Sharma, J. Lyons, A. Dehzangi, and K. K. Paliwal, "A feature extraction technique using bi-gram probabilities of position specific scoring matrix for protein fold recognition," *Journal of theoretical biology*, vol. 320, pp. 41–46, 2013. 17, 18

[42] S. F. Altschul, T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. J. Lipman, "Gapped blast and psi-blast: a new generation of protein database search programs," *Nucleic acids research*, vol. 25, no. 17, pp. 3389–3402, 1997. 18

[43] R. Sharma, A. Dehzangi, J. Lyons, K. Paliwal, T. Tsunoda, and A. Sharma, "Predict gram-positive and gram-negative subcellular localization via incorporating evolutionary information and physicochemical features into chou's general pseaac," *IEEE Transactions on NanoBioscience*, vol. 14, no. 8, pp. 915–926, 2015. 18, 19

[44] A. Sharma, K. K. Paliwal, A. Dehzangi, J. Lyons, S. Imoto, and S. Miyano, "A strategy to select suitable physicochemical attributes of amino acids for protein fold recognition," *BMC bioinformatics*, vol. 14, no. 1, p. 233, 2013.

[45] A. Dehzangi, S. Sohrabi, R. Heffernan, A. Sharma, J. Lyons, K. Paliwal, and A. Sattar, "Gram-positive and gram-negative subcellular localization using rotation forest and physicochemical-based features," *BMC bioinformatics*, vol. 16, no. 4, p. S1, 2015.

[46] A. Dehzangi, A. Sharma, J. Lyons, K. K. Paliwal, and A. Sattar, "A mixture of physicochemical and evolutionary–based feature extraction approaches for protein fold recognition," *International journal of data mining and bioinformatics*, vol. 11, no. 1, pp. 115–138, 2014. 18

[47] A. Dehzangi, K. Paliwal, J. Lyons, A. Sharma, and A. Sattar, "Enhancing protein fold prediction accuracy using evolutionary and structural features," in *IAPR International Conference on Pattern Recognition in Bioinformatics*. Springer, 2013, pp. 196–207. 19

[48] L. Breiman, "Random forests," *Machine learning*, vol. 45, no. 1, pp. 5–32, 2001. 20

[49] Y. Freund and R. E. Schapire, "A desicion-theoretic generalization of on-line learning and an application to boosting," in *European conference on computational learning theory*. Springer, 1995, pp. 23–37. 20

[50] Y. Freund, R. Schapire, and N. Abe, "A short introduction to boosting," *Journal-Japanese Society For Artificial Intelligence*, vol. 14, no. 771-780, p. 1612, 1999. 20

[51] D. M. Powers, "Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation," 2011. 20

[52] K.-C. Chou, "Some remarks on protein attribute prediction and pseudo amino acid composition," *Journal of theoretical biology*, vol. 273, no. 1, pp. 236–247, 2011. 20, 21

[53] B. Efron and G. Gong, "A leisurely look at the bootstrap, the jackknife, and cross-validation," *The American Statistician*, vol. 37, no. 1, pp. 36–48, 1983. 21

[54] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2825–2830, 2011. 22

[55] J. D. Thompson, T. Gibson, D. G. Higgins *et al.*, "Multiple sequence alignment using clustalw and clustalx," *Current protocols in bioinformatics*, pp. 2–3, 2002. 27

[56] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 27, no. 8, pp. 1226–1238, 2005. 27

[57] Q. Zou, J. Zeng, L. Cao, and R. Ji, "A novel features ranking metric with application to scalable visual and bioinformatics data classification," *Neurocomputing*, vol. 173, pp. 346–354, 2016. 27