Prediction on DNA Binding Sequence in Deep Learning Approach

Tanjina Akter[1], LubnaYasmin Pinky[2], Md. Mezbahul Islam[3] and Mohammad Motiur Rahman[4]

[1, 2, 3, 4] Department of Computer Science and Engineering (CSE), Mawlana Bhashani Science and Technology University (MBSTU), Santosh, Tangail-1902, Bangladesh

[1]tanjinashoshe007@gmail.com, [2]lubnaju@yahoo.com, [3]mailformezbah@gmail.com, [4]motiurcse@mbstu.ac.bd

ABSTRACT

Prediction of DNA-binding proteins from sequence information is the most challenging tasks in genome annotation. The principal study of our research is to determine DNA-binding proteins from primary protein sequences comparing the accuracy of the model in a deep learning based procedure. We also put the protein sequence in DNA binder tool to predict whether the sequence is DNA-Binding or non-DNA-Binding protein. We have encoded the datasets on hot encoded process and two stages of convolutional neutral network to detect the function domains of protein sequences. Our proposed method is being trained and tested with protein sequences of human and mouse, while it predicts 37% of training, 13% of testing and 50 % of the original sequences that finally count 74% of them for training and 26% of them for testing datasets of DNA-Binding proteins. The proposed method achieves a prediction accuracy of 70.50 %, sensitivity of 62.32%, specificity of 78.27% and the Matthew's correlation coefficient at 1. Support vector machine and logistic regression classifier are used upon our method to increase the accuracy. The accuracy has increased using both classifier and it raises 12% and 9% respectively. We have also count prediction analysis of precision, recall, f-measure and false discovery rate of the protein sequence.

Keywords: *DNA-binding and non DNA-binding, DNA sequence classification, Convolutional Neutral Network (CNN),Long Short Term Memory (LSTM), Support Vector Machine (SVM).*

1. INTRODUCTION

In the era of big data, transformation of biomedical big data into valuable knowledge has been one of the most important challenges in bioinformatics. Deep learning has advanced rapidly since the early 2000s and now demonstrates state-of-the-art performance in various fields. Bioinformatics of deep learning has to gain insight from data that has been emphasized in both academia and industry.

DNA-binding proteins are those proteins that bind and interact with DNA. DNA-binding proteins play important role in the structural composition of the DNA and in gene regulations. The other important role is to regulate and effect various cellular processes like transcription, DNA replication, DNA recombination, repair and modification of these proteins in their independently folded domains have at least one structural motif and have affinity to DNA [1]. Extraction of meaningful information from the DNA is a key elements in bioinformatics research and DNA sequence classification has a wide range of presentations such as genomic analysis, and biomedical data analysis [2]. The recent application of convolutional neural networks (LeCun ***et al.***, 2015; Salakhutdinov, 2015) to sequence-based problems in genomics task in computational biology. The adaptation of convolutional neural networks can be accomplished by considering a window of genome sequence as an image. Instead of processing 2-D images with three color channels (R,G,B), we consider a genome sequence as a fixed length 1-D sequence window with four channels (A,C,G,T) [3]. The two tasks we chose to explore are motif discovery and motif occupancy. In Deep Learning Architectures datasets for DNA Sequence Classification sequence alignment methods, such as BLAST and FASTA, are used. In this paper we have to propose the prediction of DNA-binding protein on deep learning by using Convolutional neural networks (CNN) and Long short term memory neural networks(LSTM). Deep learning is constituted multiple linear and non-linear transformation model. We also calculate the SVM prediction based accuracy, sensitivity, specificity and Matthew's correlation coefficient (MCC). We also compute prediction analysis measure precision, recall, f-measure and False Discovery Rate(FDR). In this research we also compare the result with DNA binder software. We also predict which sequence is DNA binding or which is non DNA binding protein with threshold.

2. BACKGROUND

The process of protein-DNA interaction has been an important subject of recent computational-biology research. DNA-binding proteins have a vital role in the biological processing of genetic information like DNA transcription, replication, maintenance and the regulation of gene expression [4]. On the prediction of DNA-binding proteins only from primary sequences utilizes two stages of convolutional neutral network to detect the function domains of protein sequences, and the long short-term memory neural network to identify their long term dependencies, a binary cross entropy to evaluate the quality of the neural networks [5]. Classification of DNA Sequences using Convolutional Neural Network approach extract of meaningful information from the DNA sequence is a key elements in bioinformatics research and DNA sequence classification has a wide range of presentations such as genomic analysis, and biomedical data analysis [6]. A new approach in classifying DNA sequences using the convolutional neural network while considering these sequences as text data. We used one-hot vectors to represent sequences as input to the model [7]. An integration of feature-based and deep learning approach for DNA motif discovery proposed the ability to identify transcription factor binding sites or motifs in the genome is one of the keys to decipher gene regulation mechanisms. Motifs are recurring sequence patterns in a genome and are the binding sites of transcription factors crucial for the regulation of protein production in cells [8].In Deeper Bind, a long short term recurrent convolutional network for prediction of protein binding specificities with respect to DNA probes [9]. We explore the use of species and lineage species models for prediction and comparing of DNA-binding protein of plant using SVM method [10]. An effective method for identifying microRNA guided cleavage sites from degradome of polyadenylated transcripts in plants are described in this paper [11]. DeepBind is perhaps the most notable success story for an application of deep learning to a challenging problem in biological domain which was proposed to address the shortcomings of classical tools in a scalable and efficient way through its deep convolutional architecture [12]. For computational methods we have used vast amount of protein sequences. The number of newly discovered protein sequences has been increasing extremely fast due to the advent of modern protein sequencing technologies. For example, in 1986 the Swiss-Prot [13] database contained only 3,939 protein sequence entries, but now the number has jumped to 88,032,926 according to the release 2017_07 of July, 5, 2017 by the UniProtKB/Swiss-Prot. Computational methods that have been used to predict the DNA-binding proteins can be broadly categorized into two groups: structure based methods [14] and sequence based methods [15]. Many machine learning algorithms are applied to solve this problem in the literature. Among them are: Logistic Regression[16], Hidden Markov Models [17], Random Forest [18], Artificial Neural Network [19], Support Vector Machines[15], Naive Bayes classifier [20] etc. Cai et al. first developed the SVM algorithm, SVM-Prot, in which the feature set came from three protein descriptors, composition (C), transition (T) and distribution (D) for extracting seven physiochemical characters of amino acids [21]. Manish Kumar et al identified the DNA-binding proteins using support vector machines and evolutionary profiles [22].A number of software, web-servers and prediction methods are available in the literature for DNA-binding protein prediction. Among them are: DNABinder [20], DNA-Prot [18], iDNA-Prot[23] etc.

3. PROPOSED METHODOLOGY

The research methodology consists of different procedure or methods that are used in deep learning approach for prediction DNA-binding proteins. The study contains two main phase which is pre-processing phase and post-processing phase. In pre-processing phase, the workflow focus on data pre-processing steps while in post-processing, the workflow can be broken down into model training and model evaluation given in below.

In Pre-processing phase:

- Raw data Collection
- Data Cleaning
- An encoding layer
- An embedding layer
- A CNN layer consist of
  - Two convolutional layers,
  - Max pooling operation.
- A LSTM layer
- Output

In Post Processing training Phase:

- Cross validation
- Model selection
- Hyper parameter optimization
- Evaluation(Performance matrices)

In deep learning approach we have proposed a model consists of several steps: an encoding layer, an embedding layer, a CNN layer, an activation layer and LSTM layer, shown in **Error! Reference source not found.**. The encoding layer maps a sequence to a fixed length digital vector. The embedding layer translates it into a continuous vector and extract the feature. The CNN layer consists of two convolutional layers, and followed by RELU and max pooling procedure. CNN layer is used to take non-linear features of protein sequences and Long Short-Term Memory (LSTM) of bidirectional networks capable of learning and vanishing gradient problem. After evaluation training datasets are cross validated and model selection phase are used for testing and then optimize the parameter and evaluate the performance matrices. For implementation we have been used Jupyter Notebook of Anaconda

Python in Keras framework, of neural networks library to predict DNA-binding proteins. For example a given protein sequence *P= MTGYTPD..*, after processing the bracing score *f*(*P*) to be a DNA-binding protein is calculated by the Eq 1. and the steps are used in this research shown in Fig. 2.

$$f\,(P)=LSTM(CNN(Embedding(encoding(P)))) \tag{1}$$

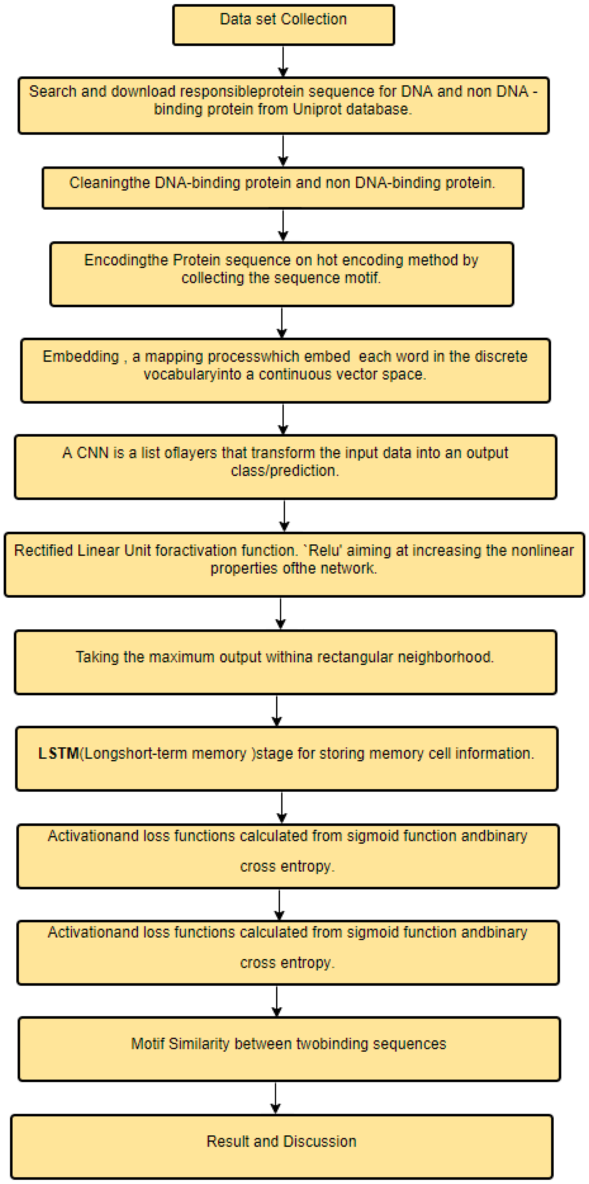

Data set Collection
Search and download responsibleprotein sequence for DNA and non DNA - binding protein from Uniprot database.
Cleaningthe DNA-binding protein and non DNA-binding protein.
Encodingthe Protein sequence on hot encoding method by collecting the sequence motif.
Embedding , a mapping processwhich embed each word in the discrete vocabularyinto a continuous vector space.
A CNN is a list oflayers that transform the input data into an output class/prediction.
Rectified Linear Unit foractivation function. `Relu' aiming at increasing the nonlinear properties ofthe network.
Taking the maximum output withina rectangular neighborhood.
LSTM(Longshort-term memory)stage for storing memory cell information.
Activationand loss functions calculated from sigmoid function andbinary cross entropy.
Activationand loss functions calculated from sigmoid function andbinary cross entropy.
Motif Similarity between twobinding sequences
Result and Discussion

Fig. 1. Steps of Proposed Methodology.

4. MATERIALS AND METHODS

4.1 Datasets Collection

The raw protein sequences are collected from the Swiss-Prot dataset, subset of UniProt at a FASTA format which is text-based format. It is a widespread, freely accessible and downloadable database of protein sequences. For predicting accuracy we have taken two kind of datasets of DNA-Binding and non-DNA-Binding proteins. To obtain DNA-Binding proteins, we collect 158829 raw dataset by searching keyword "DNA-Binding" as positive sample of human and select 211003 datasets of non-DNA-Binding proteins as negative samples of mouse from Swiss-Prot database. In reality, the number of non-DNA-binding proteins is far greater than the DNA Binding proteins. To generalize our model, we have taken the sequences in testing sets of human and mouse. This sets may be included in the training dataset because of high similarity with some sequences in training dataset.

4.2 Datasets Cleaning

We extract sequences from raw dataset by searching keyword "DNA-Binding", then remove those sequences with length less than 40 or greater than 1,000 amino acids [5]. After cleaning, 27556 protein sequence are selected as positive samples of DNA-Binding and 38268 protein sequence are selected as negative sample of non DNA-Binding by using query selection condition molecule function and length [40 to 1,000] shown in Table 1. The positive and negative datasets are shown in Fig. 3 and Fig. 4

Table 1: Datasets of DNA-Binding and non-DNA-Binding proteins.

Data set	DNA-binding	non-DNA-binding	Total
Raw Datasets	158829	211003	369832
Original Datasets	27556	38268	65824

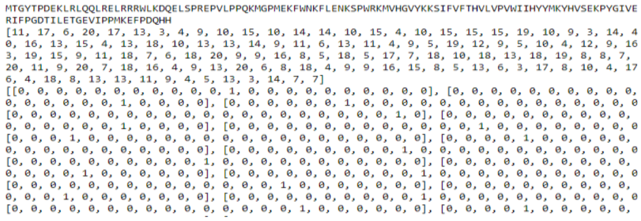
4.3 Encoding the protein sequence

Encoding the Protein sequence by collecting the sequence motif. Feature encoding is a critical task for making a statistical machine learning model and protein sequence classification. Various approaches are used for extraction feature. In this research we have proposed one-hot encoding method [5, 7]. The *auto encoder* idea is motivated by the concept of a good representation. In order to encode the protein sequence, we assign each amino acid encoder a nature number, see Table 2. It should be alarmed that the amino acids orders have no effects on the final evaluation .The encoding stage just generates a fixed length digital vector of a protein sequence [5].

$$P1 = encoding(P) = (0; 11; 17; 6; 20; 17; 13;3) \qquad (2)$$

4.3.1 One-hot encode method

Considering the amino acids value, we convert it into one hot encoded value. In One-hot encoding method [5,7] DNA nucleotides first need to be encoded numerically. They are typically represented as binary vectors with all but one entry set to zero, which indicates the category (one - hot coding). For example, DNA nucleotides (categories) are commonly encoded as A = (1 0 0 0), G = (0 1 0 0), C = (0 0 1 0) and T = (0 0 0 1) (Fig. 4).
For example we have already taken a DNA Binding sequence protein. In this sequence we have taken some part of the sequence S= MTGYTPD. But the Full part of the protein sequence is given below.
S=MTGYTPDEKLRLQQLRELRRRWLKDQELSPREP
VLPPQKMGPMEKFWNKFLENKSPWRKMVHGVYK
KSIFVFTHVLVPVWIIHYYMKYHVSEKPYGIVEKK
SRIFPGDTILETGEVIPPMKEFPDQHH.
Now the output value of the encoding sequence is given in Fig. 2.



```
MTGYTPDEKLRLQQLRELRRRWLKDQELSPREPVLPPQKMGPMEKFWNKFLENKSPWRKMVHGVYKKSIFVFTHVLVPVWIIHYYMKYHVSEKPYGIVE
RIFPGDTILETGEVIPPMKEFPDQHH
[11, 17, 6, 20, 17, 13, 3, 4, 9, 10, 15, 10, 14, 14, 10, 15, 4, 10, 15, 15, 15, 19, 10, 9, 3, 14, 4
0, 16, 13, 15, 4, 13, 18, 10, 13, 13, 14, 9, 11, 6, 13, 11, 4, 9, 5, 19, 12, 9, 5, 10, 4, 12, 9, 16
3, 19, 15, 9, 11, 18, 7, 6, 18, 20, 9, 9, 16, 8, 5, 18, 5, 17, 7, 18, 10, 18, 13, 18, 19, 8, 8, 7,
20, 11, 9, 20, 7, 18, 16, 4, 9, 13, 20, 6, 8, 18, 4, 9, 9, 16, 15, 8, 5, 13, 6, 3, 17, 8, 10, 4, 17
6, 4, 18, 8, 13, 13, 11, 9, 4, 5, 13, 3, 14, 7, 7]
[[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], [0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0], [0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0], [0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0], [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0
[0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], [0, 0, 0, 0, 1, 0, 0, 0, 0, 0,
0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], [0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0], [0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0], [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0], [0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0], [0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0
[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0], [0, 0, 0, 0, 1, 0, 0, 0, 0, 0,
```

Fig. 2. Output value of the encoding sequence.

Table 2: The Amino acids encoder section.

Amino acids	Letter	Code
Alanine	A	1
Cysteine	C	2
Aspartic	D	3
Glutamic	E	4
Phenylalanine	F	5
Glycine	G	6
Histidine	H	7
Isoleucine	I	8
Lysine	K	9
Leucine	L	10
Methionine	M	11
Asparagine	N	12
Proline	P	13
Glutamine	Q	14
Arginine	R	15
Serine	S	16
Threonine	T	17
Valine	V	18
Tryptophan	W	19
Tyrosine	Y	20
Illegal Amino acids	B,J,O,U,X,Z	0

4.4 Embedding the sequence

Embedding is a transformation from discrete values or scalars to dense real value vectors. The vector space model is used to represent words in natural language processing. Embedding is a mapping process that each word in the discrete vocabulary will be embed into a continuous vector space. In this way, similar words are mapped to similar regions. This embedding process is done by multiplying the one-hot vector from left with a weight matrix *W*, where e, is the real vectored value and xt, is the input. (3).

$$e = Wxt \qquad (3)$$

Supposed that the output length is 8, The embedding stage maps each number in *S*1 to a fixed length of vector. *S*1 becomes a 8 × 8 matrix (in 4) after the embedding stage [5].

$$P2 = embedding\ (P1) \qquad (4)$$

4.5 Convolution the sequence

CNNs are a special type of neural network whose hidden units are only connected to local receptive field. A CNN is a list of layers that transform the input data into an output class or prediction. The encoded amino acid sequence is converted into a fixed-size two-dimensional matrix as it passed through the embedding layer and can therefore be processed by convolutional neural networks like images. In this system we take the DNA sequence motifs at the window size. There are two convolution layer to convolute the encoded input sequence with kernel or filter. Support vector machine (SVM) algorithm is used for classification in convolution layer. After convolution pooling layer is executed and the max value is selected. After convolution layer each layer is fully connected to hidden layer shown in Fig. 3. As the example sequence, the convolution stage uses multiple 2-dimension filters and *W* present as matrixes, as in (5)

$$x^{l}_{j} = f(x_{i}^{l-1} \otimes \mathrm{Wj} + \mathrm{b}_{j}^{i}) \qquad (5)$$

Where *xj* is the *j th* feature map, *l* is the number of the layer, *Wj* is the *j-th* filter, $\otimes$ is convolution operator, *b* is the bias, and the activation function *f* uses `Relu'.

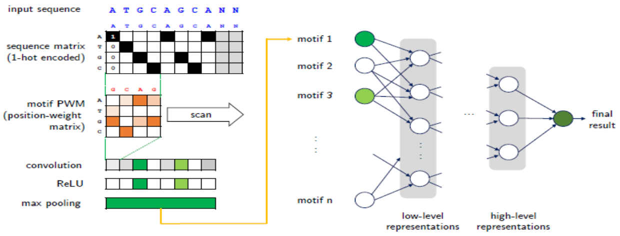


Fig. 3 Architecture of Deep learning Model.

4.6 RELU (Rectifier Linear Unit)

The activation function f, uses ‘RELU’ which aiming at increasing the nonlinear properties of the network, as shown in (6). The output of the activation layer is 0 or that positive value. If the threshold value is negative then we take zero otherwise at input value.

$$f(x) = max\,(0; x) \qquad (6)$$

The structure of convolution neural network is shown in Fig 6. Each filter is used to scan a feature in the sequences. To Understand the convolution neural network we take out a 2 × 8 filter (7) in the convolution layer and *W* is used to detect the *S*2, a 8-dimension vector, as shown below.

$$R = conv\,(P2) \qquad (7)$$

4.7 Max pooling

Taking the maximum output within a rectangular neighborhood. In convolution neutral network the model uses 2 filters to obtain 2 feature maps, then apply a max-pooling operation over the feature map and take the maximum value with the *pooling_length* = 2 is applied, *R* becomes in eqn.(8)

$$\mathrm{P3} = \mathrm{pool}\,(\mathrm{R}) \qquad (8)$$

The whole model of CNN layer is given in Fig. 6.

4.8 Long short-term memory (LSTM) Unit

A RNN composed of LSTM units is often called an LSTM network. LSTM is a special recurrent neural network architecture and provides a solution by incorporating memory units that allow the network to learn when to forget previous hidden states and when to update hidden states given new information. It uses purpose-built memory cells to store information. A common LSTM unit is composed of a cell, an input gate, an output gate and a forget gate. Each of the three *gates* can be thought of as a "conventional" artificial neuron, as in a multi-layer (or feed forward) neural network: that is, they compute an activation (using an activation function) of a weighted sum. LSTMs were developed to deal with the exploding and vanishing gradient problem when training traditional RNNs[5,9].

The LSTM generates fixed length feature representation of the output represented by *S*4 from the second convolution layer.

$$P4 = LSTM(P3) \qquad (9)$$

4.9 Activate functions of the system

A sigmoid function is called an activation function which demonstrates mathematics system such as real-valued, differentiable, positive, negative, maximum or minimum values [5]. It is defined as a strictly increasing function that exhibits a graceful balance between linear and non-linear function . So, in this work, we use it as the activation function of the network, see Eq 15.

$$O = sigmoid\,(P4) = 1 / (1+ e^{-S4}) \qquad (10)$$

We can calculate the system according to Logistic Regression[16],Random Forest [18], Artificial Neural Network [19], Support Vector Machines[20] algorithm.

4.10 Post Processing Phase

After preprocessing, 20316 DNA-Binding and 26090 non-DNA-Binding datasets for training. After training, datasets is labeled to train machine learning algorithms and increase accuracy. The sample of data used for testing to provide an unbiased evaluation of a final model fit on the training dataset for performance metrics. For DNA-binding protein, I have used Logistic Regression (LR) and Support Vector Machine (SVM) for binary classification to fit the model of datasets. The logistic function that will transform the DNA-Binding protein as positive sample and non-DNA-Binding protein as negative sample. In SVM, hyper plane is a line selected to separate the points in the input variable space by their class, either class 0 for DNA-binding and class 1 for non DNA-binding proteins. Collection of training and testing datasets of Human and Mouse is given in Table 3.

Table 3: Collection of training and testing datasets.

Data set	**DNA-binding**	**non-DNA-binding**	**Total**
Original Datasets	27556	38268	65824
Training Datasets	20316	26090	46406
Testing	7240	12178	19418

For evaluation the parameter are optimized by the above algorithm and then the performance matrices are selected in result section for final model.

5 RESULTS

5.1 Experiment setups

For calculating prediction accuracy we have taken the datasets of DNA-Binding and non-DNA-Binding proteins of human and mouse. From those datasets 42% of them DNA-Binding and 58% non-DNA-Binding protein have been calculated. After collection of datasets we indicate that DNA-Binding proteins as positive sample and non-DNA-Binding proteins as negative samples. After completing the preprocessing phase, we predict 74% of them for training and 26% of them are testing datasets shown in Fig. 4.

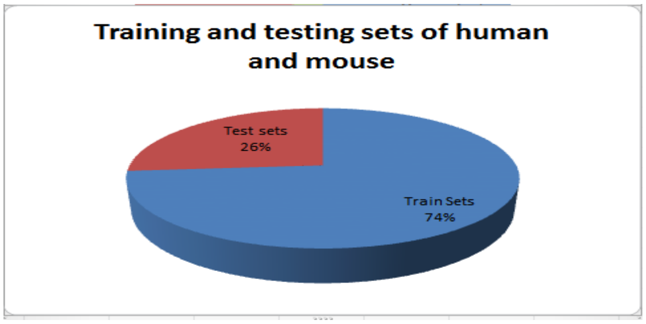


Fig. 4. Training and Testing datasets of of Human and Mouse

After preprocessing, 37% of them are chosen randomly for training, 13% of them for testing and 50% of them are total datasets of DNA-Binding and non-DNA-Binding proteins shown in Fig. 5.

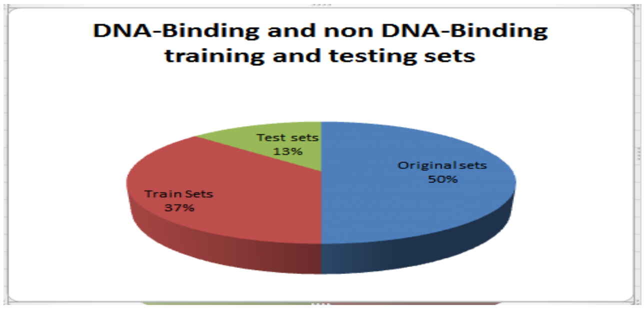


Fig 5. Training and Testing datasets of total DNA-Binding and non DNA-Binding protein sequence.

5.2 Evaluation measures

To evaluate the performance of the proposed method, the measures criteria includes accuracy, sensitivity, and specificity. There are defined in eqn. from 16 to 18.

$$Accuracy = \frac{\mathrm{TP + TN}}{\mathrm{TP + TN + FP + FN}} \times 100\% \qquad (11)$$

$$Sensitivity = \frac{TP}{TP + FN} \times 100\% \qquad (12)$$

$$Specificity = \frac{TN}{TN + FP} \times 100\% \qquad (13)$$

where TP, FP,TN and FN are counts of true positives (TP),false positives (FP),true negative (TN) and false negatives (FN) respectively. Among these measures, the accuracy is defined as the ratio of correctly predicted samples in test set, the sensitivity indicates the accuracy of true positive rate (TPR) of predicting positive samples, the specificity indicates the accuracy of true negative rate of predicting negative samples, and the accuracy is defined as the ratio of correctly predicted samples in test set [5],[8][29].

Lastly, the Matthew correlation coefficient (MCC) [8] is computed using the following formula:

$$MCC = \frac{\mathrm{TP * TN - FP * FN}}{\sqrt{\mathrm{TP + FP\,(FP + FN)\,(TN + FP)\,(TN + FN)}}} \qquad (14)$$

MCC produces a value in the range of [-1, 1], in which 1 indicates a perfect prediction, 0 means random prediction, and -1 represents a negative correlation.

Now the resultant performance measures of 70.50% accuracy, 62.32% sensitivity and 78.27% specificity of DNA-Binding protein and non DNA-Binding indicate that the values of Specificity is greater than the value of Sensitivity, then we can say that the negative of value non-DNA-Binding proteins is greater than positive value of DNA-Binding proteins and the accuracy is better for human and mouse.

Table 2: Performance measure of datasets

Datasets	Accuracy	Sensitivity	Specificity	MCC
Human and Mouse	0.705 or 70.50%	0.6252 or 62.32%	0.7827 or78.27%	1

Now the resultant performance measures of accuracy, sensitivity and specificity of DNA-Binding and non-DNA-Binding proteins of human and mouse given in Fig. 6.

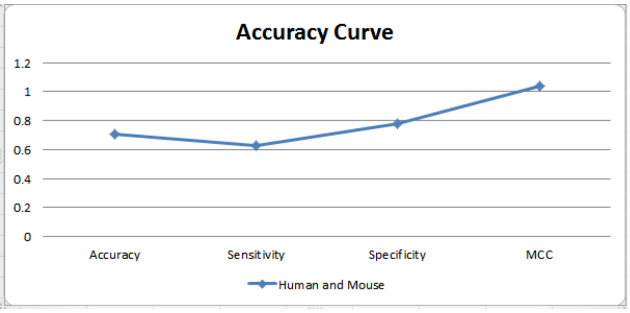


Fig. 6. Evaluation measures curve of DNA-Binding Proteins.

The comparison measure of accuracy, sensitivity and specificity of DNA-Binding and non-DNA-Binding proteins of human and mouse given in Fig. 7.

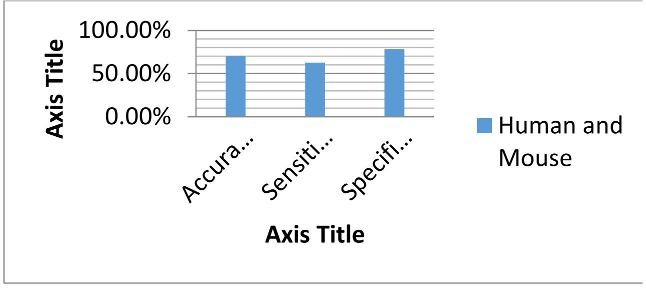


Fig. 7. Comparison measures curve of DNA and non DNA-Binding Proteins.

5.3 Performance Measures

Precision and Recall

The accuracy of the classifier is assessed by the f-measure and false discovery rate (FDR) [8,29]. The f-measure f is given by the following formula 22 to 25:

$$f\_measure = \frac{2pf}{p+r} \qquad (15)$$

Where p and r are the precision and recall rates, respectively [8, 29]. FDR is defined by 1 – precision [8,51]. Precision and recall rates are computed using the following formulas:

$$\text{Precision}, p = \frac{TP}{TP+FP}$$

(16)

$$\text{Recall}, r = \frac{TP}{TP+FN} \qquad (17)$$

Where TP, FP and FN are counts of true positives (TP), false positives (FP) and false negatives (FN), respectively, from the cross-validation experiment.
The result of the precision, recall ,f-measure and FDR(False Discovery Rate) is given in the Fig 8.

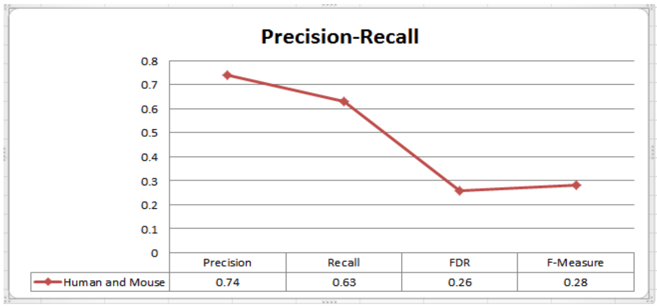


Fig. 8. Performance measure of precision, recall, FDR and f-measure rates obtained.

5.3 Performance comparisons

İn this section, we have been applied Support Vector Machine (SVM), Logistic Regression(LR), Random Forest (RF) algorithm on our proposed Neural Network(NN) model for increasing the accuracy of the model. After putting those method the accuracy of our model is increased and compare the performances with existing feature extraction methods. It is clearly shows that SVM has the highest value of accuracy with 82.94% followed by LR, RF and NN with 79.8%,75% and 70.50% respectively. Fig. 9 shows a comparison of accuracy of all compared methods including the proposed methods.

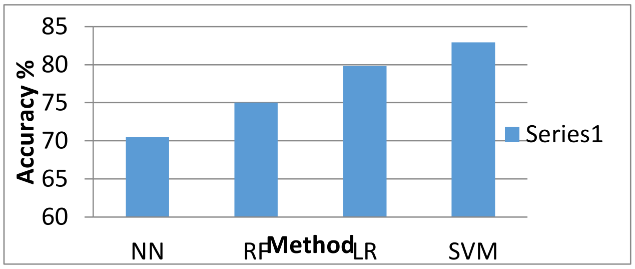
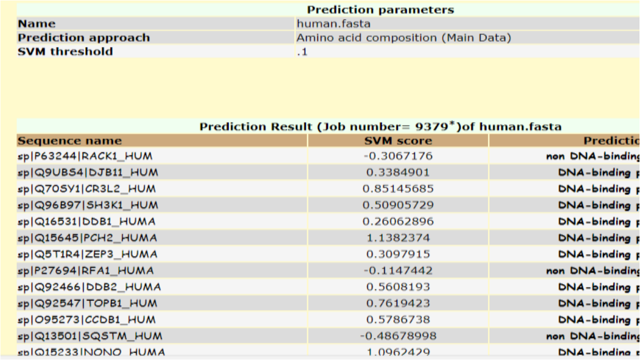


Fig. 9. Performance comparison accuracy.

5.4 Prediction Analysis

We predict the DNA-Binding protein by using DNAbinder tools, which is a webserver for predicting DNA-Binding protein. We put the sequence of binding proteins in the DNAbinder and set the the SVM threshold at 0.1 then it predict which sequence is DNA-Binding or non DNA-Binding protein. The prediction approach is amino acids composition. Then it provides result with SVM score and predict the sequence is shown in Fig. 10



Prediction parameters	
Name	human.fasta
Prediction approach	Amino acid composition (Main Data)
SVM threshold	.1

Prediction Result (Job number= 9379*)of human.fasta

Sequence name	SVM score	Predictic
sp|P63244|RACK1_HUM	-0.3067176	non DNA-binding
sp|Q9UBS4|DJB11_HUM	0.3384901	DNA-binding p
sp|Q70SY1|CR3L2_HUM	0.85145685	DNA-binding p
sp|Q96B97|SH3K1_HUM	0.50905729	DNA-binding p
sp|Q16531|DDB1_HUMA	0.26062896	DNA-binding p
sp|Q15645|PCH2_HUMA	1.1382374	DNA-binding p
sp|Q5T1R4|ZEP3_HUMA	0.3097915	DNA-binding p
sp|P27694|RFA1_HUMA	-0.1147442	non DNA-binding
sp|Q92466|DDB2_HUMA	0.5608193	DNA-binding p
sp|Q92547|TOPB1_HUM	0.7619423	DNA-binding p
sp|O95273|CCDB1_HUM	0.5786738	DNA-binding p
sp|Q13501|SQSTM_HUM	-0.48678998	non DNA-binding
sp|Q15233|NONO_HUMA	1.0962429	DNA-binding p

Fig. 10 . Prediction of DNA-binding protein in DNA binder

6. DISCUSSION

Existing methods often devote themselves for extracting features from sequences but ignoring motif information and affected large noise with lower accuracy of prediction. For this , we propose a deep learning based computational method to predict DNA-binding proteins from primary sequences. In this research, we have calculated 42% DNA-Binding and 58% non DNA-Binding protein sequence of human and mouse. After collecting and filtering, we have selected the datasets for training and testing. For each dataset, 37%of them are chosen randomly for training, 13%of them for testing and 50% of them are total datasets. Finally we have to predict 74% training and 26% testing datasets for human and mouse. In this paper we have to count the accuracy, sensitivity, specificity and MCC of two kinds of species by SVM (support vector machine) and RF (random forest) classification model. We have to calculate 70.50 % accuracy,62.32% sensitivity and 78.27% specificity of human and mouse. In this section we observe that the specificity is higher than sensitivity which indicate that non-DNA-Binding proteins are mostly responsible in protein sequences. We also count prediction analysis of precision, recall, f-measure and FDR of the protein sequence. Compared the proposed method with existing method the accuracy becomes 82.94% in SVM and 79.8 in LR method and raises 12% and 9% from previous. The DNAbinder tools predict which sequence is DNA-Binding or non DNA-Binding protein.

7. CONCLUSION

In this paper, we present a deep learning based approach for predicting DNA binding functions of proteins only using primary sequences. The deep learning framework capable of extracting feature automatically and predicting structure by using machine learning algorithm Support Vector Machine (SVM),logistic regression and training models in a back propagation way. The two layers of CNN plus LSTM networks allow for an increase in learning power and contain more potential for motif refinements in both of local connectivity and long-term dependence. We have calculated accuracy, sensitivity, specificity, MCC, precision, recall for prediction analysis. Compared the other model we observe that the accuracy of proposed model raises by putting SVM classifier. We have bought the primary protein sequence in DNA binder software for predicting which sequence is DNA binding or which is non DNA binding protein. It also demonstrates substantial generality across multi-species testing. Moreover, the method outperforms most of the existing feature extraction methods plus a successful machine learning algorithm. The investigation of the deep learning model in predicting DNA binding functions of proteins might yield a competitive tool for future. The proposed deep learning approach would have many other potential applications, such as identifying DNA binding proteins, investigate to drug for diseases, protein detection, RNA prediction etc. In future, we wish to update the prediction method by incorporating an enhanced dataset.

REFERENCES

[1] Shahana Yasmin Chowdhury, Swakkhar Shatabda & Abdollah Dehzangi . iDNAProt-ES: Identification of DNA-binding Proteins Using Evolutionary and Structural Features. Scientific Reports. 2017; 7,14938

[2] Kumar M, Gromiha MM, Raghava GPS. Identification of DNA-binding proteins using support vector machines and evolutionary profiles. BMC bioinformatics. 2007;8:463. https://doi.org/10.1186/1471-2015-8-463

[3] Haoyang Zeng, Matthew D. Edwards, Ge Liu, David K. Gifford; Convolutional neural network architectures for predicting DNA–protein binding. Bioinformatics.2016;Volume 32, Issue 12, https://doi.org/10.1093/bioinformatics/btw255

[4] Andrea Szabóová, Ondřej, Kuželka, Filip Železný and,Jakub Tolar. Prediction of DNA-binding proteins from relational features. Proteome Science.2012;10:66

[5] Qu Y-H, Yu H, Gong X-J, Xu J-H, Lee H-S .On the prediction of DNA-binding proteins only from primary sequences: A deep learning approach. PLoS ONE.2017; 12(12): e0188129. https://doi.org/10.1371/journal.pone.0188129

[6] Nurul Amerah Kassim1, and Dr Afnizanfaizal Abdullah2. Classification of DNA Sequences Using Convolutional Neural Network Approach.Innovations in Computing Technology and Applications.2017; Volume 2 ISBN: 978-967-0194-95-0

[7] Ngoc Giang Nguyen1, Vu Anh Tran1, Duc Luu Ngo1, Dau Phan1, Favorisen Rosyking Lumbanraja1, Mohammad Reza Faisal1, Bahriddin Abapihi1, Mamoru Kubo2, Kenji Satou2, .DNA Sequence Classification by Convolutional Neural Network . Biomedical Science

and Engineering.2016; Vol.09 No.05(2016), Article ID:65923,7 pages 10.4236/jbise.2016.9502

[8] Nung Kion Lee, Farah Liyana Azizan, Yu Shiong Wong & Norshafarina Omar. DeepFinder: An integration of feature-based and deep learning approach for DNA motif discovery. Biotechnology & Biotechnological Equipment (2018), 32:3, 759-768, DOI: 10.1080/13102818.2018.1438209

[9] Hamid Reza Hassanzadeh, May D. Wang. DeeperBind: Enhancing Prediction of Sequence Specificities of DNA Binding Proteins.IEEE International Conference on Bioinformatics and Biomedicine (BIBM) (2016)

[10] Motion G B, Howden A J M, Huitema E, et al. DNA-binding protein prediction using plant specific support vector machines: validation and application of a new genome annotation tool. Nucleic Acids Research, 2015,43(22):e158±e158. https://doi.org/10.1093/nar/gkv805 PMID: 26304539.

[11] Yun Zheng Yong-Fang Li Ramanjulu Sunkar Weixiong Zhang . SeqTar: an effective method for identifying microRNA guided cleavage sites from degradome of polyadenylated transcripts in plants.Nucleic Acids Research.2012;volume 40, Issue 4, 1, https://doi.org/10.1093/nar/gkr1092.

[12] Babak Alipanahi, Andrew Delong, Matthew T Weirauch, and Brendan J Frey. Predicting the sequence specificities of dna-and rna-binding proteins by deep learning. Nature biotechnology, 2015.

[13] Consortium, U. et al. Uniprot: the universal protein knowledgebase. Nucleic acids research.2017; 45, D158–D169.

[14] Zhao, H., Yang, Y. & Zhou, Y. Structure-based prediction of dna-binding proteins by structural alignment and a volume-fraction corrected dfire-based energy function. Bioinforma.2010;26, 1857–1863.

[15] Lou, W. et al. Sequence based prediction of dna-binding proteins based on hybrid feature selection using random forest and gaussian naive bayes. PLoS One 9, e86703 (2014).

[16] Szilágyi, A. & Skolnick, J. Efficient prediction of nucleic acid binding function from low-resolution protein structures. J. molecular biology.2006; 358, 922–933.

[17] Shanahan, H. P., Garcia, M. A., Jones, S. & Thornton, J. M. Identifying dna-binding proteins using structural motifs and the electrostatic potential. Nucleic Acids Research..2004; 32, 4732–4741.

[18] Kumar, K. K., Pugalenthi, G. & Suganthan, P. Dna-prot: identification of dna binding proteins from protein sequence information using random forest. J. Biomol. Struct. Dyn..2009; 26, 679–686

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

[20] Kumar, M., Gromiha, M. M. & Raghava, G. P. Identification of dna-binding proteins using support vector machines and evolutionary profiles. BMC bioinformatics 8, 463 (2007)

[21] Cai CZ, Han LY, Ji ZL, Chen X, Chen YZ. SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. Nucleic Acids Research. 2003; 31 (13):3692±3697. https://doi.org/10.1093/nar/gkg600 PMID: 12824396

[22] Manish Kumar, Michael M Gromiha and Gajendra PS Raghava. Identification of DNA-binding proteins using support vector machines and evolutionary profiles.BMC Bioinformatics.2007;8:463 https://doi.org/10.1186/1471-2105-8-463

[23] Lin, W.-Z., Fang, J.-A., Xiao, X. & Chou, K.-C. Idna-prot: identification of dna binding proteins using random forest with grey model. PloS one 6, e24756 (2011).

[24] Daniel Quang, Xiaohui Xie; DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences, Nucleic Acids Research. 2016; volume 44, https://doi.org/10.1093/nar/gkw226

[25] Cheng, X., Zhao, S.-G., Xiao, X. & Chou, K.-C. Iatc-misf: a multi-label classifier for predicting the classes of anatomical therapeutic chemicals. Bioinforma. 33, 341–346 (2016).

[26] Liu, B., Wang, S., Long, R. & Chou, K.-C. Irspot-el: identify recombination spots with an ensemble learning approach. Bioinforma. 33, 35–41 (2016).

[27] Liu, L.-M., Xu, Y. & Chou, K.-C. Ipgk-pseaac: identify lysine phosphoglycerylation sites in proteins by incorporating four different tiers of amino acid pairwise coupling information into the general pseaac. Medicinal Chem. 13, 552–559 (2017).

[28] Qiu, W. et al. Irna-2methyl: identify rna 2 ′ -o-methylation sites by incorporating sequence-coupled effects into general pseknc and ensemble classifier. Medicinal chemistry (Shariqah (United Arab. Emir., https://doi.org/10.2174/1573406413666170623082245 (2017).

[29] Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc Ser B. 1995; 57:289–300.

[30] Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta (BBA)-Protein Struct. 1975;405:442–451.

[31] Asgari E, Mofrad MRK. Continuous distributed representation of biological sequences for deep proteomics and genomics. PLoS ONE. 2015; 10(11):1±15. https://doi.org/10.1371/journal.pone.0141287.

[32] Jo T, Hou J, Eickholt J, Cheng J. Improving Protein Fold Recognition by Deep Learning Networks. Scientific reports. 2015; 5:17573. https://doi.org/10.1038/srep17573 PMID: 26634993.