Sequence Based DNA-Binding Protein Prediction

Farisa T S, Elizabeth Isaac

Abstract**:** ***Protein and DNA have vital role in our biological processes. For accurately predicting DNA binding protein, develop a new sequence based prediction method from the protein sequence. Sequence based method only considers the protein sequence information as input. For accurately predicting DBP, first develop a reliable benchmark data set from the protein data bank. Second, using Amino Acid Composition (AAC), Position Specific Scoring Matrix (PSSM), Predicted Solvent Accessibility (PSA), and Predicted Probabilities of DNA-Binding Sites (PDBS) to produce four specific protein sequence baselines. Using a differential evolution algorithm, weights of the properties are taught. Based on those attained properties, merge the characteristics with weights to create an original super feature. And tensor-flow is used to paralyze the weights. A suitable feature selection algorithm of tensor flow's binary classifier is used to extract the excellent subset from weighted feature vector. The training sample set is obtained in the training process, after generating final features. The classification is learned through the support vector machine and the tensor flow. And the output is measured using a tensor surface. The choice is done on the basis of threshold of likelihood and protein with above-threshold chance is considered to be DBP and others are non-DBP.***

Keywords **:** ***AAC, DBP,PSA, PSSM.***

I. INTRODUCTION

Protein and DNA have vital role in our biological processes. The precise targeting of DNA-binding proteins (DBP) is therefore of important to protein function annotation. The intrinsic structure of protein-DNA interactions was discovered with enormous wet-lab efforts. It take high cost and time. Faced with the question of experimental identification of DBPs and the current sequences produced in the post genomic lifetime, it is desirable to produce full technology is strongly wanted automatic system for fast and accurate targeting of DBPs. Various categories like zinc finger can form the relation. These interactions have vital role in our biological processes. These all methods include the best featured. While on our experimentally defined composition are fixed in the fixed database of Protein Data Bank (PDB), only less tight structures of protein – DNA are identified, which is far lighter than the total. In recent years the advancement of next-generation sequencing technologies has completed many genome sequences of different species. Enormous numbers of DNA and protein sequences were made, many of them DNA-binding proteins. Examining how and when protein–DNA interactions will enhance genome comprehension. A complete picture of the interactions enables characterization of Gens transcribed in response to a dynamically changing environment at any given time. Traditionally, DNA-binding proteins or residues can be identified using various experimental techniques, such as conventional immuneprecipitation of chromatin, MicroChIP, Fast ChIP, RNA-binding protein peptide nucleic acid assisted identification, etc. Such methods however are time-consuming and expensive. It is also important to build computational tools that can quickly and accurately identify DNA-binding proteins or residues, given the huge amount of protein sequence data available. Many computational methods for targeting DBPs have emerged over the last decades. These approaches can be divided into two groups, roughly divided into two groups depending on the features they used: structure-based methods and methods based on sequences. The structure-based methods, e.g., DBD-Hunter and iDBPs, typically use both the structural and sequential details of target proteins, whereas the sequence-based methods only consider the protein sequence as data. For the structure-based methods, although they may show promising predictive performance, their application is limited as there is not always the structural protein information available. Conversely, the sequence-based methods may this shortcoming can be overcome only by using the sequence information as an input for the DBP prediction. Introduce a sequence predictor to further boost the accuracy of the DBP forecast. For make the proposed method a useful protein dependent sequence feature is taken. It involves (1) sequence based property, i.e., Amino Acid Composition (AAC), Position Specific Scoring Matrix (PSSM), Predicted Solvent Accessibility (PSA), and Predicted Probabilities of DNA Binding Sites (PDBS) are draw out to reflect basic characteristics; (2) differential evolution algorithm is used to find the mass of the properties; (3) From this, we are able to merge the property with their weight to create the main property; (4) an appropriate feature selection algorithm of tensor flow classifier and SVM, is utilized to pick a subgroup of the main property for further dividing the contrast between DBPs and non-DBPs. Thirdly, the prediction model is obtained by tensor flow classifier on the chosen subgroup.

Manuscript received on February 03, 2021.
Revised Manuscript received on February 15, 2021.
Manuscript published on March 30, 2021.
* Correspondence Author

Farisa T S*, Department of Computer Science and Engineering, Mar Athanasius College of Engineering, Kothamangalam, Kerala, India.
Email: farisafari1996@gmail.com

Dr. Elizabeth Isaac , Department of Computer Science and Engineering, Mar Athanasius College of Engineering, Kothamangalam, Kerala, India.
Email: Elizabeth.issaac@gmail.com

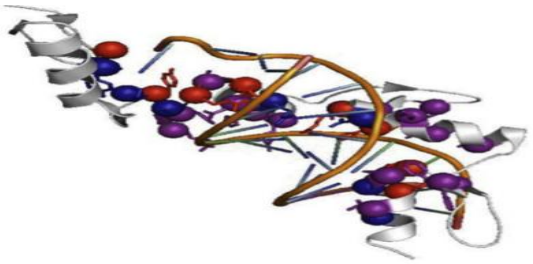


Fig 1. DNA Binding Residues

II. RELATED WORK

Protein and DNA have vital role in our biological processes. For accurately predicting DNA binding protein, develop a new sequence based prediction method from the protein sequence. Correctly finding DNA-binding proteins (DBP) is of supreme for the biological tasks.

A. Prediction of DNA-Binding Protein Sites

Various properties have been used to classify DBP and non-DBP to construct the DNA-binding residue predictor. The protein characteristics can generally be subdivided into sequence property, structural property and physical and chemical property. Residue properties like amino acid sequence, persistence, evolutionary retained residue, composition, structural motifs, structural neighborhood, global amino acid, secondary structure, electrostatic potential, and dipole and quadruple moment

- **Amino Acid Sequence:**

Of any sequence-based predictor, amino acid property is the basic character. There are mainly 20 amino acids with various properties, like polar residues, and hydrophobic residues.

- **Structural Features:**

Most recognized DNA-binding protein structures have unique composition pattern. All the patterns begin and end with helices based on the names, connected by a short region having different profile. These patterns used to make a distinction between DNA-binding proteins and other proteins.

- **Solvent Exposed Area:**

It is Waters at the surface of a protein can reach residues. Growing atom can theoretically be affected by water, and the region of an atom on the surface that can be affected by water is called the accessible molecular surface, or the region exposed to solvents. It is obvious that large polar residues have a wide, accessible surface on average and vice versa

- **Hydrophobicity**:

It is the physical property of a water repellent molecule. Hydrophobicity was commonly used in predictors which bind DNA.

- **Net charge, and quadruple moments**:

The net charge, the moment of the electric dipole and the quadruple moment measure how widely an electrical charge is spread over the protein. They differentiate fairly performing binding and non- binding proteins. The combinations of those features make DNA-binding protein prediction a relatively discriminatory tool.

B. DNA-Binding Protein Prediction using Methods of Mixed Feature Representation

DNA-binding proteins play vital roles in cellular processes such as the packaging of DNA, replication, transcription, regulation, and other activities associated with DNA. The accuracy depends mainly on the method of extraction of the features. Sequence-based methods are based only on the information about the protein sequence. The features are derived using sequence details, such as the structure of amino acids and the sum of amino acids, without taking into account any structural information. These methods are therefore highly effective and useful in the analysis of large-scale datasets of protein sequences. Classified DNA-binding proteins by using an SVM and coding characteristics from evolutionary evidence. The First to use PSSM profiles, which provide knowledge about evolution. It is important to use an efficient method of representation of features to improve the accuracy of the classification. Nevertheless, existing methods for rep-resenting features cannot differentiate DNA-binding proteins from non-DNA-binding proteins in an effective manner. In multi-functional representation method, which combines three methods of representation of features, namely K-Skip-N-Grams, Sequential and Structural properties, is used to represent the sequences of proteins and improve the representation of features. The classifier is also an SVM. The approach of mixed function representation is tested using a 10-fold cross-validation and a test set. The methods of mixed feature representation allow features to represent protein sequences from several aspects and enhance the impact of classification. The mixed vectors may however contain redundant or even contradictory vectors. Therefore the dimensions need to be that. Methods founded on structure are time consuming. Hence, these approaches refer only to small-scale datasets. And sequence-based approaches neglect the structural relation- ship of proteins and their physical and chemical properties. So such approaches cannot reach a high standard of accommodation.

III. PROPOSED WORK

Interactions of protein-DNA are common for all living species. Protein-to-DNA interactions are important for our biological processes. Precise targeting of DNA-binding proteins (DBP) is therefore of importantly necessary for protein role annotation. Develop a new sequence dependent predictor is used to accurately predict DBP. It uses a sequence-based approach, i.e. sequence-based method considers only the knowledge about the protein sequence as data. Next build a robust benchmark dataset from PDB for accurate prediction of DBP. Second, by using AAC, PSSM, PSA, and PDBS, generate four distinct protein sequence features. To find the value of the properties, a Differential Evolution Algorithm is used. Tensor flow is used for weight paralysis. With this learned weights, combine the characteristics with weights to create an original super power. To extract the excellent subgroup from the final property, an appropriate feature selection algorithm called binary classifier is employed. After generating final features, get the training sample set in training phase.

Prediction model is learned on selected feature subset by means of the tensor flow classifier and SVM. The decision is based on chance and threshold. Protein with an above-threshold probability is known to be DBP and others are non-DBP.

A. Dataset

The first important step in the development of statistical predictors is to construct a detailed, accurate, and stringent benchmark dataset. Benchmark dataset S can be formally denoted as follows for DBP prediction:

$$S = SposiUSnega \qquad (1)$$

Where Sposi means a positive subset containing only DBPs, Snega means containing only non-DBPs, and the sign in the set theory represents the ”union.” To build Sposi, we remove all DBP chains from PDB first. – DBP chain is marked in PDB as a DNA binding protein or contains at least one DNA binding protein.

B. Feature Representation

Prediction of the binding residues of protein-DNA are standard binary Problem with classification. Therefore, prepare a Machine-learning software for predicting how to encode Binding residues of protein-DNA with discriminative elements it’s one of the most critical moves. Various effectiveness the protein sequence extracts sequence-based features such as AAC, PSSM, PSA, and PPDBS.

- **Amino Acid Composition (AAC)**

AAC is a sequence-based function commonly used in many prediction tasks of protein attributes including prediction of DBP. Let AA1, AA2,…, AA19, and AA20 are the different amino acids, ni be the number of occurrence of particular amino acid in the sequence, and L be the protein length.

$$f_{AAC} = [\frac{n_1}{L}, \frac{n_2}{L}, ..., \frac{n_{20}}{L}]^T dots \qquad (2)$$

Where T means vector transposition.

- **Position Specific Scoring Matrix (PSSM)**

A position weight matrix (PWM), also named as a position-specific weight matrix or PSSM, is a widely used biological sequence representation of motifs (patterns). A simple position frequency matrix (PFM) is generated in the first step of constructing a PWM by counting each nucleotide’s occurrences at each location. A position probability matrix (PPM) that now generated from PFM by cutting the prior nucleotide value by the total number of series at each position, thereby normalizing the values. The elements in PWMs are most commonly measured as likelihoods for logging. That is, in a background model, the elements of a PPM are transformed.

- **Predicted Solvent Accessibility (PSA)**

Waters at the surface of a protein can reach residues. Growing atom can theoretically be affected by water, and the region of an atom on the surface that can be affected by water is called the accessible molecular surface, or the region exposed to solvents. It is obvious that large polar residues have a wide, accessible surface on average, and small hydrophobic residues have a small, accessible surface on average.

- **Predicted Probabilities of DNA-Binding Sites (PPDBS)**

In theory, all DBPs are targeted correctly when the binding sites of DNA (DBSs) can be predicted with 100% perfection. However, most state-of-the-art DBS predictors can achieve an accuracy of only 80%. Direct use of DBS predictor results to find the DBPs is not a good method help to refine DBP prediction rightness. Here, DBS predictor’s predictive probability results function as a new feature and draw out the applicable feature from it to refine the DBP prediction rightness.

C. Learning Weights to Multi-view Functions

Most important steps in creating a DBP prediction model is how to integrate the four base features. The simplest and easy method is to merge basic properties serially and directly in order to get a main character (i.e., AAC+PSSM+PSA+PPDBS). Differential evolution (DE) algorithm, most efficient types of evolution algorithms is to find the perfect weights of these base characteristics.

D. Prediction using Tensor flow

TensoFlow is a low-level math toolkit that targets researchers they know that what they are done to build learning architectures, play with them and turn them into produce a software. Tensorflow is an open source library of software for solving large numerical computations.

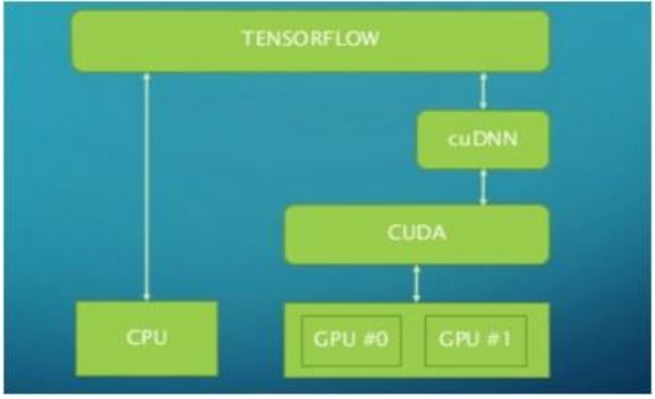


Fig 2. Architecture of Tensorflow

We can use the TensorFlow library do to numerical computations, which in itself doesn’t seem all too special, however these computations are finished with main data drift graphs. Based on the graphs, vertex constitute operations, and the edge constitute facts, they are typically arrays or tensors, that are linked between those edges. The term” TensorFlow” is derived from the operations on multidimensional data arrays or tensors carried out by neural networks. And binary classifier tensor flow is used for DBP prediction. And here we use the binary classifier of tensor flow along with SVM. And a performance comparison of both are done.

IV. RESULT AND DISCUSSION

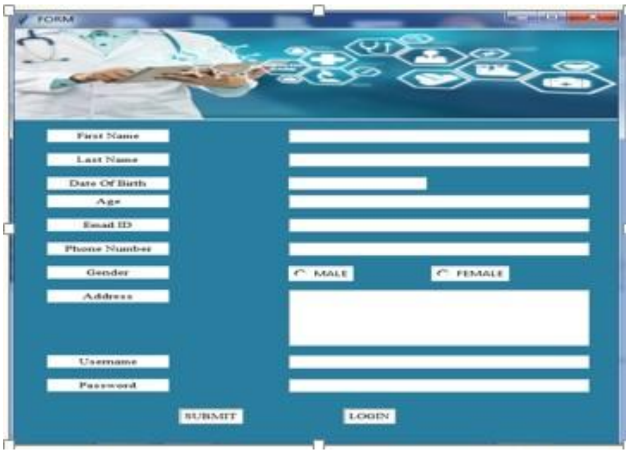


Fig 3. Registration Page

Enter valid credentials to register in to the system for DBP prediction. And login to the system using valid username and password.

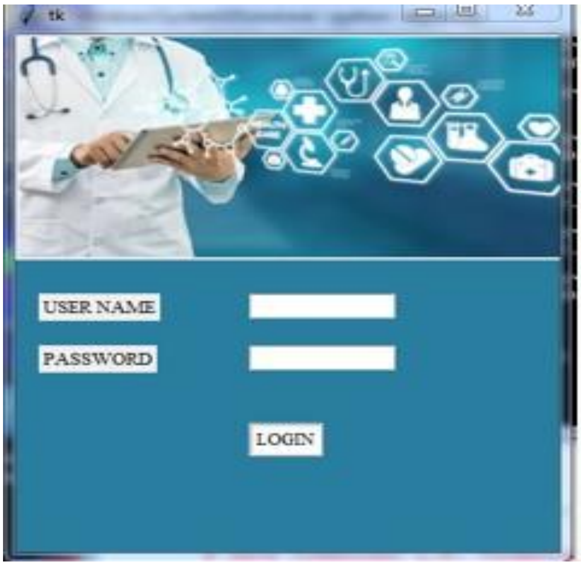
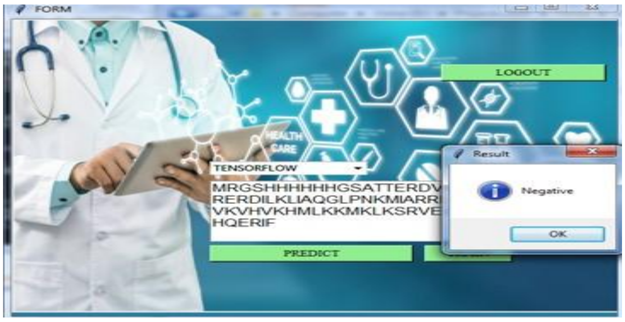


Fig 4. Login Page

DBP can be predicted using tensor flow binary classifier by putting the protein sequence as input. The model is created and accuracy is calculated. Also compare it with SVM prediction and tensor flow produce high result. Training phase is also done by these two methods. The protein sequence features are extracted. Mainly focused on AAC, PSSM, and PSA.

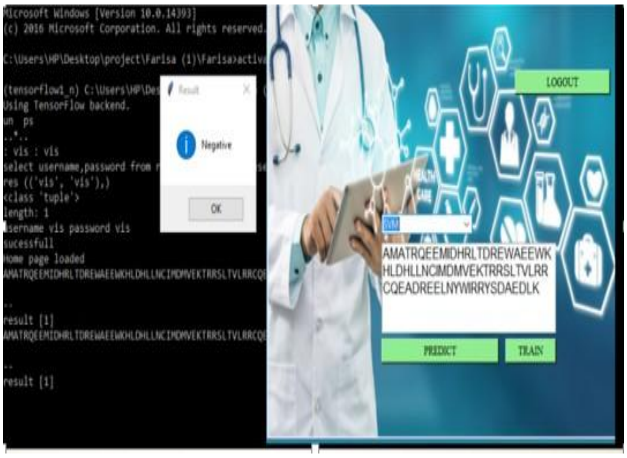


TENSORFLOW

MRGSHHHHHHGSATTERDV
RERDILKLIAQGLPNKMIARR
VKVHVKHMLKKMKLKSRVE
HQERIF

Negative

Fig 5. Prediction Using Tensorflow Binary Classifier



```
Microsoft Windows [Version 10.0.14393]
(c) 2016 Microsoft Corporation. All rights reserved.

C:\Users\HP\Desktop\project\Farisa (1)\Farisa>activa

(tensorflow1_n) C:\Users\HP\Des
Using TensorFlow backend.
un  ps
..*..
: vis : vis
select username,password from r
res (('vis', 'vis'),)
<class 'tuple'>
length: 1
sername vis password vis
sucessfull
Home page loaded
AMATRQEEMIDHRLTDREWAEEWKHLDHLLNCIMDMVEKTRRSLTVLRRCQ

..
result [1]
AMATRQEEMIDHRLTDREWAEEWKHLDHLLNCIMDMVEKTRRSLTVLRRCQ

..
result [1]
```

Negative

SVM

AMATRQEEMIDHRLTDREWAEEWK
HLDHLLNCIMDMVEKTRRSLTVLRR
CQEADREELNYWIRRYSDAEDLK

Fig 6. Prediction Using SVM

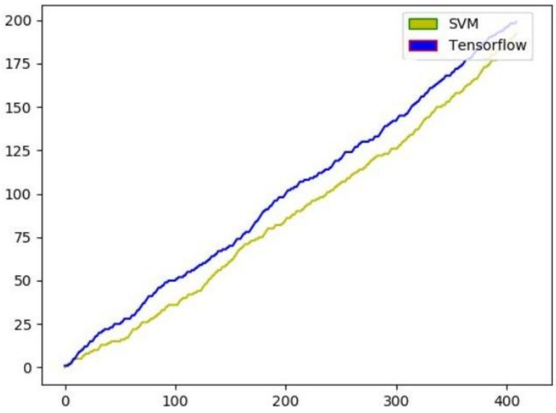


Fig 7. Comparison Graph

To evaluate the performance evaluation of the proposed model, to identify the Accuracy (ACC), Specificity (SP), Sensitivity (SE). And also determine the scoring matrix, TP (True positive), FP (False Positive), TN (True Negative), and FN (False Negative). The figure 7 shows the comparison of SVM method with proposed methods on the dataset. From the evaluation, the proposed method classifier has higher accuracy rate than SVM classifier.

V. CONCLUSION

Accurate recognition of protein functions is one of the most critical tasks of protein annotation. To improve DBP prediction performance, a new sequence based predictor has been developed and implemented. We can derive the four separate single-view functions for a given protein, i.e., AAC, PSSM, PSA and PPDBS. Based at the weights, which is calculated on the basis of DE set of rules on the education set, the process of function merging can be accomplished to generate a final feature vector. A suitable feature selection algorithm of tensor flow's binary classifier is used to extract the excellent subset from weighted feature vector. The training sample set is obtained in the training process, after generating final features. The classification is learned through the support vector machine and the tensor flow. And the output is measured using a tensor surface.

The choice is done on the basis of threshold of likelihood and protein with above-threshold chance is considered to be DBP and others are non-DBP.

REFERENCES

1. K. C. Wong, Y. Li, C. Peng et al, *A Comparison Study for DNA Motif Modeling on Protein Binding Microarray*, in:IEEE/ACM Transactions on Computational Biology & Bioinformatics, vol. 13, no. 2, pp. 1-1, 2016.
2. J. N. Si, R. Zhao, and R. L. Wu, *An Overview of the Prediction of Protein DNA-Binding Sites*, in:International Journal of Molecular Sciences, vol. 16, no. 3, pp. 5194-5215, 2015.
3. P. W. Rose, A. Prlić, C. Bi et al., *The RCSB Protein Data Bank: views of structural biology for basic and applied research and education*, Nucleic acids research, vol. 43, no. D1, pp. D345-D356, 2015.
4. X. P. Schmidtke, and X. Barril, *Understanding and predicting drug-gability. A high-throughput method for detection of drug binding sites*, in:Journal of medicinal chemistry, vol. 53, no. 15, pp. 5858-5867, 2010.
5. Pradnya P. Mandlik, Samruddhi S. Mhatre, Hiding Data into Reserve Space before Image Encryption using Blowfish Algorithm, Volume 140– No.10, April 2016.

AUTHORS PROFILE

Farisa T S received Bachelor of Technology in Computer Science and Engineering from KMEA Engineering College Edathala in 2018 and currently pursuing Master of Technology in Computer Science and Engineering from Mar Athanasius College of Engineering, Kothamangalam affiliated to APJ Abdul Kalam Technological University. Her research interest is in Computer Security.

Dr. Elizabeth Isaac is currently working of Computer Science and Engineering, Mar Athanasius College of Engineering, Kothamangalam, Kerala, India. She received her B-Tech Degree in Computer Science and Engineering in 2008 from Mar Athanasius College of Engineering, Kothamangalam, and M-Tech in Computer Science and Engineering from Vellore Institute of Technology, Chennai in 2010. She received PhD on computer architecture in 2018.She has around 9 years of teaching and research experience in various institutions in India. Her research interests include Computer Architecture.

Retrieval Number: 100.1/ijrte.B3665079220
DOI:10.35940/ijrte.B3665.039621
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication