Support Vector Machine based Prediction of Transcription Factor Binding

Smitha C S
Assistant Professor in Computer Science & Engineering
College of Engineering Muttathara
Thiruvananthapuram, Kerala

Dr. Afzal A L
Assistant Professor in Computer Science & Engineering
College of Engineering Muttathara

***Abstract*—Any living organism, both prokaryotic and eukaryotic is primarily based on the working of genes which comes under gene regulation. Protein synthesis is the primary concern in the gene regulation which is controlled by the transcription factor binding sites. The genetic behavior is primarily based on the interaction of synthesized proteins with the DNA regions. It is very difficult to identify exact binding sites and hence computational methods are used. The computationally identified binding sites are validated using experimental methods. Feature identification of motifs is made and remarkable features are ranked and selected based on Principal Component Analysis. Support Vector Machine based analysis and prediction increases the accuracy and prediction of the obtained results.**

***Keywords*— RBF kernel; Principal Component Analysis; Transcription Factor**

I. INTRODUCTION

The smallest unit of life is the cells that form the building block of living things. Cells contain several regulatory elements, and nucleus which houses the genes. Inside the nucleus there are chromosomes which contain genes as the genetic carrier in a coiled form [1]. The fundamental unit of hereditary information can be described as the genes which act as a medium through which basic traits are passed from an organism to its successor. The study of genome is a vital factor in identifying the genetic behavior which flags on to the detection and cure of many diseases. Innumerable data analysis with prediction is experimentally very difficult and hence Self organizing algorithms are used for data visualization in genes. DNA is the carrier of genes which in turn contain the nucleotides which of the prime focus. The genetic details are coded in the Deoxyribo Nucleic Acid (DNA). DNA comprises four nucleotides Adenine, Thymine, Cytosine and Guanine which are in short abbreviated as A, T, C and G. DNA is structured as a double stranded coil in the form of a helix. To form the basic structure of the helix Adenine pairs with Thymine using two hydrogen bonds whereas Cytosine pairs with Guanine with three hydrogen bonds. The Cytosine Guanine bonding is a comparatively stronger bond than the Adenine Thymine pair.

1.1. Protein Synthesis

Protein synthesis is marked by the uncoiling of DNA strands. The uncoiling of double stranded DNA is done upon the activity of enzymes like RNA Polymerase. This results in the formation of single stranded mRNA which is otherwise known as messenger RNA. It acts as a template for the formation of proteins. The process of formation of mRNA from DNA is known as transcription which in turn contains the relevant information for the generation of amino acids. Proteins [2] are formed as a result of the binding of amino acids. The amino acids are coiled, tangled and folded as two and three dimensional to generate the resulting protein molecules.

1.2. Transcription Factor

Many proteins are formed and some of the generated proteins may reenter the nucleus to interact with specific portions of DNA. These proteins are known as Transcription Factors (TF). The portions of DNA that interact with these Transcription Factors are known as Transcription Factor Binding Sites (TFBS) or in short binding sites [3]. The binding sites are specific for any organism and they play a major role in gene regulation.

This paper gives an insight into the prediction of DNA protein interaction. The computational prediction is primarily focused on feature extraction, analysis of extracted features and then further verification using the Support Vector Machine classifier. Section 2 lists some of the related work, Section 3 gives the proposed system and Performance assessment is shown in Section 4. Section 5 gives the future scope and applications and finally section 6 concludes the paper.

II. RELATED WORK

There are mainly 4 category of binding prediction algorithms which can be listed as enumerative, iterative, phylogenetic foot printing and content based algorithms. The enumerative algorithms use a background model for the binding prediction and the probability technique of Expectation Maximization is used in iterative approach for the binding prediction. The consensus of binding sites can be annotated by a set of frequently used subsequences. A set of known binding sites from evolutionarily related species are used in Phylogenetic foot printing. Ortholog and homolog modeling can be used for obtaining the relevant details including the regulatory elements. The content based algorithms use the divide and conquer approach; basis of which is to take a sequence of sufficient length, dividing it into sub sequences that is sufficient for calculations and the regularities of subsequence is used for prediction [4]. The chances of errors resulted from the analysis of a single algorithm is very large and hence a class of twelve algorithms are used for the accurate prediction. But this may lead to an increased additional load.

The amino acids and nucleotides are represented using a standard IUPAC notation. The selected notation incorporates the entire DNA and RNA nucleotides as well as the twenty amino acids representation. The prediction is based on both nucleotide and amino acids which result in the analysis of both. The feature identification is based on regularity of patterns that produces remarkable number of distinguished features. Due to the invariant size of the data, the identified features will be of high dimensionality that may result in incorrect predictions in majority of the cases. This will require a well defined and thorough analysis of identified binding sites. To reduce this tedious work in high dimensionality techniques like wrappers, filters and embedded methods are commonly used. This employ additional parameters like F-score, correlation coefficient etc. for further analysis and evaluation.

Most of the binding predictions can be made based on the similarity matrix which is also called as Position Specific Scoring Matrix [5]. An enumerative approach is used for the feature extraction with the adequate data taken with the help of background subsequences. The similarity matrix can be prepared based on the DNA nucleotides or amino acids which result in varying dimensions. From the matrix, consensus can be prepared which can be used for finding the Information Content [6]. The similarity score can also be used to reveal the phylogenetic content.

The prediction of binding sites can be based on the protein DNA interaction residue [7]. The details about amino acid can be obtained from the Protein Data Bank. The approach use the data collected from a set of known binding sites which can be collected from some of the known databases like JASPAR and TRANSFAC. The technique is mainly based on the experimentally available binding sites that can be obtained by techniques like DNase Foot printing and ChIP techniques [8]. The approach has the disadvantage of tedious data collection, verification and validation.

2.1 Feature based approaches

The residue based binding prediction use Biological features. This includes the side chain pKa value, hydropathical index values and structural data. The hydropathical index gives whether an amino acid is more hydrophilic or more hydrophobic. The structural analysis is dependent on the distance between amino acid and nucleotides in the binding residue. If the distance between nucleotide and amino acid molecules are less than 3.5 A, then it is assumed to be a binding [9]. The confusing and complex structural analysis of the binding residue leads to a high overhead for the prediction.

2.2 Machine Learning Techniques

The Information Content obtained from the similarity matrix can be used for training and testing phase of machine learning. A Neural Network (NN) can be trained for the prediction results which can be tested using any testing set [10]. The data sets for the prediction can be taken as PDNA 62. The similarity can be accessed by BLAST values. BLAST hits can be collected from the protein databases like Uniprot. The approach require large amount of training data to extract the sufficient details for training. The number of hidden intermediate layers must be set up for better evaluation of the prediction.

Genetic algorithm based prediction is another technique for prediction which uses the principles of genetic algorithm [11]. The operations used are initialization, evaluation, selection, crossover and mutation. Initialization is used for selecting the sequences for analysis. The binding prediction can be made using crossover and mutation of selected subsequences. The genetic operations cannot be applied for all eukaryotic and prokaryotic binding predictions.

III. PROPOSED SYSTEM

The proposed method gives a computational approach for transcription factor binding identification using Support Vector Machine. A stepwise methodology was employed which is described in Figure 3.1. Preprocess the dataset obtained from NCBI database to create fixed length sequences. Obtain the Transcription Factor details from the UniProt database. Extract features like binding affinities and oligonucleotide properties. Perform Principal Component Analysis to filter the features that do not play a major role in prediction. Perform classification using Support Vector Machine (SVM) and assess performance measure of the classifier.

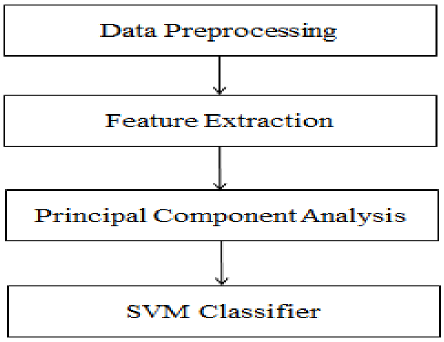

Figure 3.1: Proposed Method for binding prediction

3.1 Dataset and Preprocessing

The data used for processing is taken from the NCBI data repository with label NC_012920 with the name MT_NDI. The nucleotide sequence for processing is identified. The details can be obtained from the database National Centre for Biotechnology Information (NCBI) with the genbank accession number NC_012920. The main activities involved in preprocessing are: Human Mitochondrial DNA selected with 16569 base pairs, genetic data converted into frames, extract the noncoding sequences upstream of protein coding regions, DNA sequence restricted by start codon ATG, sliding window for DNA sequence with window size 7 and use this for feature extraction, Repeat this for positive as well as negative datasets.

Preprocessing of the data is required to create nucleotide sequences of fixed length. Unnecessary data is removed from the input to make the processing easier and faster which is termed as noise removal. The data is converted to frames for easier processing. The processing includes finding triplets for the sequence termed as codons. There will be a start codon and end codon for demarcation. The sequences are divided into fixed length subsequences for processing so that the results can be obtained easily. The subsequences are formed based on the window size. An odd window size is selected so that there will be equal contribution for binding. A sliding window approach is used for the processing. The data consists of both positive and negative data sets. 5 data sets are prepared which was named as DS 1, DS 2, DS 3, DS 4 and DS 5. The data sets are made with varying window size 7, 9, 11, 13 and 15.

3.2 Feature Extraction

The main features related to DNA transcription are structural features, chemical features and letter features. The structural features include correlation between free and bound locations, based on the geometry of base pairs and based on hydroxyl radical cleavage of DNA. It also includes the physico chemical properties that can be obtained from the spatial configuration of amino acid with DNA. It depends on the distance between the nucleotide amino acid distances in the resulting residue. It can be extracted from the Protein Data Bank (PDB). One such chemical feature selected can be the distance of 3.5A^o between DNA and protein molecule in the interaction residue. The letter features include Position Weight Matrix based methods, solvent accessibility and evolutionary information.

The main features extracted are TATA Box, CAAT Box, Position Specific Scoring Matrix value, Strong Hydrogen Bonds, CpG Island, Melting Temperature, Enthalpy, Entropy and Free Energy. The structure of transcription factor can be taken to find the residue form. This is considered as an additional feature for the binding prediction.

TATA feature can be extracted based on the letter pattern in the DNA sequence. The windowed subsequence is searched for repeated patterns of T, TA and TAT. Regular expression based searching with lexeme identification can be done for obtaining the percentage of pattern TATA. The feature value is obtained as an integer corresponding to the percentage of tokens T, TA, TAT with combinations to form TATA. An example of a subsequence with 100% feature value is GATATAA.

CAAT is a regular expression based pattern search for the search of subsequence CAAT. The subsequence can be searched for the identification of repeated patterns of C, CA, CAA and CAAT. The integer feature value of 100% denote a successful identification of the pattern CAAT. The maximum value of search result can be returned as the feature value.

A modified approach for finding PSSM score is used as another feature. The technique is to take the windowed subsequences such that the consecutive subsequences differ by a gap of 6 segments. Each of the subsequences has a dimension of 7 nucleotides. A matrix is prepared with the rows corresponding to nucleotide bases in the order Adenine, Cytosine, Guanine and Thymine. The column values of the matrix are filled by taking the nucleotides in the corresponding positions with a matrix dimension of 4x7. A nucleotide of position ‘i’ stands in column i of the matrix. A block size of 6 to 7 is taken for the matrix preparation.

The frequency of each nucleotide is obtained from the matrix. This is used to find the weight of each nucleotide by considering the block size also.

$$\text{Weight} = \text{Frequency} / \text{Blocksize} \qquad (1)$$

This gives the weight for all the four DNA nucleotides. The score value of a particular subsequence is obtained by taking the sum of weights of all nucleotide bases in the sequence.

GC content can be obtained by taking the ratio of the sum of the number of Cytosine and Guanine base pairs to the total length of the subsequence. Higher the GC value more will be the probability of binding. High GC content gives a region for more probability of interaction.

Another feature content is CpG Islands. CpG Islands are regions with a high frequency of CpG sites. CpG Island is a region with at least 200 base pairs and a GC percentage that is greater than 50% and with an observed to expected ratio greater than 60%. Expected CpG value is found by considering the number of Cytosine and Guanine against the total number of nucleotides in the sequence. CpG stands for the bases Cytosine and Guanine separated by a single Phosphate molecule with Cytosine proximal to Guanine base. The calculated CpG value is compared with
those of the observed value.

Melting temperature is the temperature required for half of the DNA to be separated into single strands. It is based on the hybridization or melting process. A simple method to find the melting temperature is to assign 2^o C to each Adenine-Thymine pair and 4^o C to each Guanine-Cytosine pair. G-C pair is having three hydrogen bonds which lead to a high melting temperature. Melting temperature can be taken by considering all the individual bases. Salt adjustment of the medium is also considered to find the average melting temperature.

Thermodynamic calculations include Enthalpy (H), Entropy (S) and Free Energy (G). Free Energy is calculated based on Enthalpy and Entropy by considering the Temperature (T).

$$G = H - T * S \qquad (2)$$

The free energy for protein DNA interaction complex is obtained by taking the individual elements, Transcription Factor and DNA segment.

$$G_{\text{protein-DNA}} = G_{\text{protein}} + G_{\text{DNA}} \qquad (3)$$

Enthalpy and Entropy is calculated by taking the dimers of bases. The bases are taken as a triplet. From the triplet, dimers are taken at a time for processing. The enthalpy and entropy for binding nucleotides and amino acids are considered for free energy calculation.

3.3 Feature Selection

After extracting the DNA features and structural analysis of proteins, there is a need to analyze whether all these features play a major role in binding prediction. Principal Component Analysis is used for feature selection to filter the unimportant features in binding prediction. It transfers a data matrix of m objects by k variables which are correlated into a new set of uncorrelated axes or Principal Components. These Principal Components are linear combinations of the original variables. New axes are orthogonal and represent the directions with maximum variability [12]. In Principal Component Analysis Eigen vectors are formed from the covariance matrix. Then order them by Eigen values from highest to smallest. This gives principal components in the order priority which helps to ignore components of lesser importance. The features mostly used in the prediction are Free Energy, Entropy, PSSM, CAAT, CpG Island and Strong Hydrogen bonds.

3.4 Classification

Support Vector Machine (SVM) is used for the classification. This is a supervised machine learning approach used for the classification problems. Here the classification is between the binding set and non-binding set. SVM uses statistical learning methodology by using a hypothesis space of a linear space of functions in a high dimensional feature space [13]. Given a set of labeled training data, training an SVM classifier involves finding a hyper plane that maximizes the geometric margin between positive and negative training samples. The hyper plane is given as f(x). A test instance x is assigned a positive label if f(x)>0 and negative label otherwise. The RBF kernel is used for implementing SVM as a kernel function.

IV. PERFORMANCE ASSESSMENT

A classifier is having a binary mode of prediction. It means the mapping from input instances to predicted classes. The prediction is Yes or No which indicate the positive and negative classes simultaneously. Classification is based on mainly two types of outcomes. The outcome can be labeled as Actual class and Predicted class. There are four possible outcomes for a binary classifier. They are True Positive (TP), True Negative (TN), False Positive (FP) and False Negative (FN). If the predicted value and actual value are positive, then it is called TP, but if the predicted value is positive for an actual negative value then it is FP. Similarly if the predicted outcome is negative for an actual positive value then it is called FN and if the predicted negative outcome is obtained for an actual negative then it is TN.

Accuracy is the proportion of the total number of predictions that were correct.

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \qquad (4)$$

Sensitivity is the percentage of binding sites correctly predicted as binding sites.

$$\text{Sensitivity} = TP / (TP + FN) \qquad (5)$$

Specificity is the percentage of nonbinding sites correctly predicted as nonbinding sites.

$$\text{Specificity} = TN / (TN + FP) \qquad (6)$$

ROC is a technique for visualizing and analyzing the classifiers performance. ROC graph is widely used to analyze the machine learning classifiers as a performance graphing method. It is a graphical plotting with True Positive rates on the Y-axis and False Positive rates on the X-axis. So it is a plot of Sensitivity on Y-axis against 1-Specificity on X-axis.

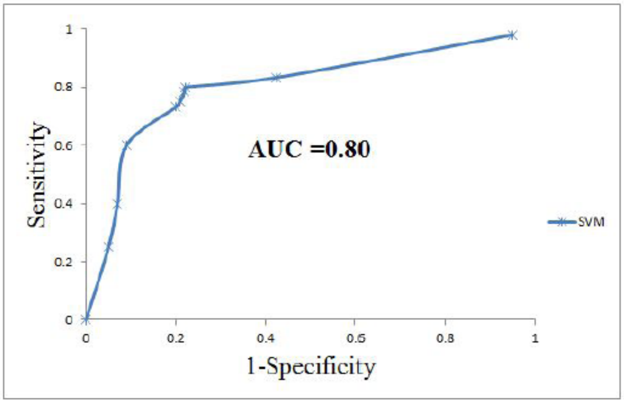

Fig. 4.1: ROC curve of SVM classifier

Fig. 4.1 shows the ROC graph. The Area Under Curve value is greater than 0.5 which indicates a good classifier.

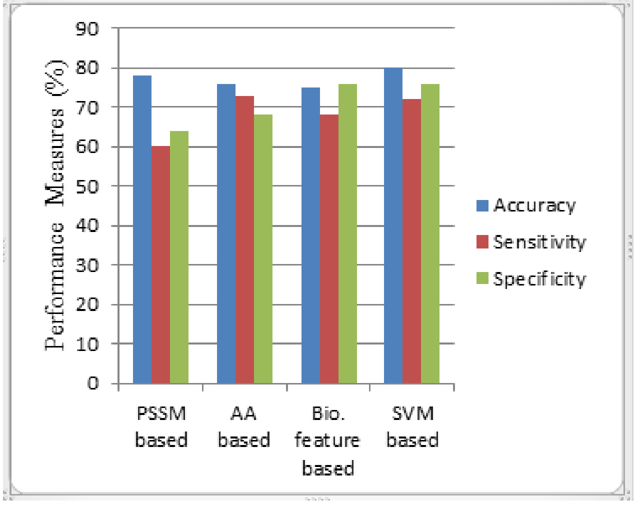

Fig. 4.2: Comparison with existing techniques

Fig. 4.2 shows the comparison of the proposed prediction with some of the existing techniques. The existing techniques used for comparison are PSSM based, Amino Acid (AA) based and biological (biochemical) feature based. Our model outperforms the existing ones which is a clear indication of better prediction.

V. FUTURE SCOPE AND APPLICATIONS

The interaction of multiple Transcription Factors to form a complex Transcription Factor binding model can be extended as a future work. The model acts as a regulatory network for determining the binding locations. The regulatory network acts as a model for determining the hereditary traits of an organism. It can also act as a means for detection of mutations in the genetic data which leads to the early diagnosis of tumors.

CONCLUSION

Transcription Factor binding prediction is one of the most important activities in gene regulation. The prediction is highly complicated because of the large size and confusing data. Computational techniques help in the prediction of binding sites which can be experimentally validated. A group of 5 data sets were selected for the study. The proposed method using feature extraction and feature selection was analyzed using the classifier Support Vector Machines. The performance comparison was made with accuracy, sensitivity, specificity and ROC graph. The proposed system was analyzed and found to be better than existing systems.

REFERENCES

[1] Stephane. M. Nelson, Lynnette. R. Furguson and William. A. Denny, “DNA and the Chromosome – Varied Targets for Chemotherapy,” BioMedCentral, Cell & Chromosome, 2004,doi:10.1186/1475-9268-3-2.

[2] Sankar. Subramanian and Sudhir Kumar, “Neutral Substitutions occur at a Faster Rate in Exons than in Noncoding DNA in Primate Genomes,” Genome Research, 2003, 13, pp.834-844.

[3] N. M. Luscombe and J. M. Thornton, “Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity,” Journal of Molecular Biology, July 2002.

[4] T. T. Nguyen and I. P. Androulakis, “Recent advances in the computational discovery of transcription factor binding sites,” Algorithms, pp. 582–602, September 2009.

[5] D. T. Jones, “Protein secondary structure prediction based on position specific scoring matrices.,” Molecular Biology, no. 292, pp. 195–202, 1999.

[6] G. D. Stromo, “DNA binding sites: Representation and discovery,” Bioinformatics, vol. 16, no. 1, pp. 16–23, 2001.

[7] S. Ahmad and A. Sarai, “PSSM based prediction of DNA binding sites in proteins,” BMC Bioinformatics, February 2005.

[8] G. D. Stromo, “DNA binding sites: Representation and discovery,” Bioinformatics, vol. 16, no. 1, pp. 16–23, 2001.

[9] Nanjiang Shu, Tuping Zhou and Sven Hovmoller, “Prediction of Zinc-Binding Sites in Proteins from Sequence,” Bioinformatics, Vol. 24, No.6, 2008, pp.775-782.

[10] Brusic V, Rudy G, Honeyman G, Hammer J and Harrison L, “Prediction of MHC ClassII Binding Peptides using an Evolutionary Algorithm and Artificial Neural Network,” Bioinformatics, 1998, pp.121-130.

[11] X. Chang, W. Zhou, C. Zhou, and Y. Liang, “Prediction of transcription factor binding sites using genetic algorithm.,” in IEEE Computational Systems Bioinformatics Conference, 2006.

[12] J. E. Jackson, “A user’s guide to principal components,” John Wiley and Sons, 1991.

[13] C. Cortes and V. Vapnik, “Support vector networks,” Machine Learning, vol. 20, pp. 273–297, September 1995.