Transcription Factor Binding Site Prediction Using CnNet Approach

Mohamed Divan Masood[1], Dr Manjula[2], *Vijayan Sugumaran*[3,4]

[1]*Department of Computer Science and Engineering, B S Abdur Rahman Crescent Institute of Science and Technology, Chennai - 600048, India.*

[2]*Department of Computer Science and Engineering, Anna University, Chennai, India*

[3]*Department of Decision and Information Sciences, Oakland University, Rochester, MI, USA*

[4]*Center for Data Science and Big Data Analytics, Oakland University, Rochester, MI, USA*

Abstract — Controlling the gene expression is the most important development in a living organism, which makes it easier to find different kinds of diseases and their causes. It's very difficult to know what factors control the gene expression. Transcription Factor (TF) is a protein that plays an important role in gene expression. Discovering the transcription factor has immense biological significance, however, it is challenging to develop novel techniques and evaluation for regulatory developments in biological structures. In this research, we mainly focus on 'sequence specificities' that can be ascertained from experimental data with 'deep learning' techniques, which offer a scalable, flexible and unified computational approach for predicting transcription factor binding. Specifically, Multiple Expression motifs for Motif Elicitation (MEME) technique with Convolution Neural Network (CNN) named as CnNet, has been used for discovering the 'sequence specificities' of DNA gene sequences dataset. This process involves two steps: a) discovering the motifs that are capable of identifying useful TF binding site by using MEME technique, and b) computing a score indicating the likelihood of a given sequence being a useful binding site by using CNN technique. The proposed CnNet approach predicts the TF binding score with much better accuracy compared to existing approaches. The source code and datasets used in this work are available at https://github.com/masoodbai/CnNet-Approach-for-TFBS.git

Index Terms— Motif Discovery, Transcription Factor Binding Site, Convolution Neural Network, MEME, Sequence Specificity.

1. INTRODUCTION

The Next Generation Sequences (NGS) analysis has been one of the most challenging processes in computational biology. NGS technology has been used in several genetic processes and also to predict various genetic diseases with the help of the DNA (Deoxyribonucleic acid) double helix structure. Genetic sequencing comprises wide-ranging and appropriate tasks that include: a) identifying the similarity between two kinds of (homologous) sequences, b) developing proper gene feature selection method, based on computational methodology, c) identifying sequence dissimilarity and modifications such as mutations and particular nucleotide polymorphisms in the sequencing markers, and d) identification of molecular arrangement and assorted gene expression.

Presently, knowing the gene expression by using computational approaches is fairly difficult. Computational methods use a combination of statistical and functional analyses to understand gene expressions (Pearson 2013). This field is a subset of computational biology, which focuses more on understanding how DNA works at the molecular level to control a range of functions in living organisms. Moreover, excessive small size of factor helps to identify and control the rate of gene expression.

Transcription Factor (TF) is a protein that binds DNA and transcript of genetic information from DNA to Messenger-Ribonucleic Acid (mRNA). It controls the rate of gene expression, and binding to the specific gene sequences is named Transcription Factor Binding Sites (TFBSs). At present, identification of precise TF binding site is a challenging problem for any researcher in molecular biology (Quang, D, & Xie, X 2019). Also, uniqueness of genetic sequences can be found with TF and genetic diseases can be cured by the specificities of the gene sequences. Apart from this, TFs bind to regions such as the RNA polymerase and protein binding sites (Bulyk 2003).

TF binding site has mainly been used for identifying the disease variations, drug identification for specific protein, gene regulation as well as many applications in molecular biology (Morishita et al. 1998 & Mann et al. 2000). Many different methods are available for finding the TFBS. Generally, they are based on the principles of information theory or machine learning techniques, which are implemented on web servers (Banki et al. 2017). However, finding the sequences characteristics is a difficult process, hence, we need new approaches for finding the sequences specificity with good accuracy (Reddy et al. 2007). In this paper, we propose a novel approach called CnNet, which automatically learns motif scanners, along with rules for combining them to make good predictions, for sequence analysis tasks.

A motif refers to a common pattern in a given sequence, and a single motif is repeated in the same sequence. There are two regions, intron and exon, in gene sequences. The motif is only obtained from the exon region and proteins are subsequently formed. The different lengths of the motif are identifiable from the gene sequence dataset, with motif lengths varying from 8 to 24 (Fan et al. 2015). Identifying short sequences, where gene mutations happen, is a big challenge. The motif is a binding site, though common short sequences may be found, at the same time, in the middle of a motif (Bailey 2011).

The MEME method identifies the most accurate motif position because it's an example of the deterministic optimization method (Bailey et al. 2006). The Position Weight Matrices (PWM) is used to identify the potential TF binding site in the gene sequences. It can identify the characteristics of the sequence's specificity (Felicioli et al. 2012). The PWM is mainly used to discover the motif pattern and determines the differences in the sequences.

Currently, Deep Learning (DL) Technology is the most popular method for analyzing biological datasets. This technology attempts to model the relationships in data based on different approaches. There are several layers in a deep

learning network and hence, at every layer the incoming signal is modified and passed on to the subsequent layers. The multiple layers can perform both linear and nonlinear transformations. It differs from regular neural networks in terms of the direction of flow of neurons. Regular neural networks only allow neurons to flow in single direction and thus enabling only forward feed. Though feed forward networks are well suited for text and image recognition, sometimes the network demands full connectivity resulting in complex structures. Large datasets demand complex structure for efficient training and this has resulted in poor performance of traditional neural networks (Yaman et al 2023, Zhanget al 2023, Zhang et al. 2021, & Chen et al. 2021).

TF binding site is computed with different kinds of datasets that focus on Protein-Protein Interaction (PPI) (Luo et al. 2014), Microarrays (Annala et al. 2011) and DNA Sequences (Yu et al. 2023 & Alipanahi et al. 2015). The PPI network calculates the score using Dijkstra's algorithm and the TFBS is identified using the PWM. The proposed CnNet technique addresses some challenges: i) DNA Sequence datasets have been applied, ii) It has analyzed millions of DNA sequences by using Graphics Processing Unit (GPU) with parallel processing, iii) MEME technique has been used for motif scanning process, iv) our model has given an accurate value without keeping the bias constant, and v) Most importantly the training method has given very fast and accurate results without any data loss.

The remainder of this paper is organized as follows. Section 2 provides a brief review of the related literature. Section 3 discusses the design of the TFBS prediction System which implements our CnNet methodology. Section 4 describes the experimental results and evaluation, Finally, Section 5 provides the conclusion & future work.

2. Literature Survey

In computational biology, gene sequences analysis plays a key role and has major applications such as diagnosis of genetic diseases (Barany 1991), drug identification (Payne et al. 2007), structural variations (Li et 2022 & Feuk et al. 2006), and gene expression (Robinson et al. 2010), among others. Historically, Sanger & Tuppy (1951) first analyzed the DNA gene sequence. Then, Needleman & Wunsch (1970) discovered the difference between the two sequences through a computer algorithm. NGS, widely used in medical research, can easily identify disease and vital for diagnosis. In earlier time, pattern matching finds the exact occurrence of patterns in given sequences, when a specified pattern is present. Most techniques are based on a pattern matching algorithm, as in, for example, Brute-Force (Faheem 2010), Knuth-Morris-Pratt (KMP) (Rajesh et al. 2010), Boyer-Moore (Antonino & Villa 2010) and the Rabin-Karp (RK) Algorithm (Ondov et al. 2010).

The TF binding sites are located among motifs and can be identified from various datasets. However, new methods are needed to analyze these datasets and predict the TF binding sites. In computational biology, binding sites have used PWM to scan DNA sequences. Historically, computational methods have affected DNA binding site prediction, which has been elucidated in (Stormo 2000). Further, the representation of the TF binding site can be accomplished so that new sequences can be generated efficiently. Based on this representation, TF binding sites in each sequence can be located and a representation for sequence specificity can be provided.

Bailey (2011) introduced a MEME Technique and most of the papers reported in the literature have used the MEME algorithm for motif discovery. This algorithm extends the EM algorithm for scanning motifs. A big advantage of this algorithm is that it works without any prior knowledge of what motifs are present in the given sequence. A multi objective Genetic Algorithm (GA) was proposed by Boone et al. (2021), which is effective over a single objective.

Initially, the binding site was computed by sequence signals. The MAMOT technique obtains the signal rank and background rank of each probe by using Hidden Markov models (HMMs). Thereafter, the corrected signal rank of each probe is defined as the signal rank minus the background rank, and the average background rank as the mean of the background rank of the given probe among all the biological sequences (Schütz & Delorenzi 2008). Linhart et al. 2008 suggested a new method named Amadeus to analyze sequence signals. This system identified binding sites based on two methods: k-mer set memory algorithm with PWM. This technique obtains the signal rank and background rank of each probe by sorting their raw probe signal and background signal, respectively. Thereafter, the corrected signal rank of each probe is defined as the signal rank minus the background rank, and the average background rank as the mean of the background rank of the given probe among all the datasets.

The MEME, combined with the Hidden Markov Model (HMM) method, has given good results in the motif stage and the probability value is calculated using the HMM (Sharon et al. 2008). This method identifies short sequences using probability values. Their prediction method further takes a reproducible probe-specific but factor-independent bias into account. Their model is not completely automatic, as certain parameters were set intuitively. Machine Learning (ML) techniques predict TF binding sites with good accuracy. SVM play an important role in identifying binding sites, and give good classification accuracy (Sohn et al. 2009).

The Protein Binding Microarrays (PBM) dataset used in the RankMotif++ technique works by combining two algorithms, Random Forest (RF) and PWM (Chen et al. 2007). The PWM is used for aligning sequences and the RF for classifying common patterns of different lengths. It begins modeling by selecting a sample of sequences with evenly distributed binding intensities that is subsequently divided randomly into two equal sets – the training set and the validation set. Each sequence is then constructed, with several sets of descriptive variables that will be used by the ML technique. Next, the RF classifier is trained on the training set data using these variables. The motif-finding algorithm is also applied for a subset of sequences with high binding intensities.

Artificial Neural Network (ANN) is also used to find binding sites (Manioudaki & Poirazi 2013) and play a key role in aligning gene sequences. This model has the most neurons and thus progressively improves in terms of performance. Quan et al. (2020) have developed a technique using CNN (convolutional neural network) to predict a TFBS, named as SemanticsCS (Semantic ChIP-seq). SemanticCS technique is used in pinpointing substitutions leading to regulatory abnormalities and in assessing the impact of substitutions on the binding affinity for the RXR transcription factor.

Alipanahi et al. (2015) introduced a DeepBind method for analyzing DNA sequences using CNN, and identified the TFBS score based on the motif. This method is applicable to DNA microarray and sequencing data. However, it tolerates a moderate degree of noise and mislabeled training data and trains predictive models fully automatically, alleviating the need for careful and time-consuming hand-tuning. More importantly, a trained model can be applied and visualized in ways that are familiar to users of PWMs.

DeepSEA application on the other hand uses CNN to predict the effects of non-coding variants (Zhou & Troyanskaya 2015). It had found chromatin features from holdout genomic sequences with high accuracy. This surpassed the performance of the till then best method for TF prediction of sequences, which is gapped k-mer support vector machine. In discovering TFBS, Deep CNN (Zhang et al. 2021) provides a sample architecture which provides greater than 96 percent accuracy on a simulated dataset. However, overfitting is one of the major challenging problems in sequencing analysis. DNN suffers from the overfitting issue and dropout is a methodology for addressing the overfitting issue.

The utilization of deep learning techniques in the MachineTFBS model advances the identification of high-affinity TF binding sites from in vitro experiments. Yaman et al. (2023) have conducted experiments using Random Forest, eXtreme Gradient Boosting, and Deep Learning models with up to a 5-depth structure, as the choice of machine learning methods varies for different TF. However, the analysis yields less precise results due to the distinctive challenges associated with binding site identification (Yaman et al 2023). Neikes et al. (2023) introduce a method named Binding Affinities to Native Chromatin by Sequencing (BANC-seq), designed to ascertain the absolute apparent binding affinities of transcription factors to native DNA across the entire genome. BANC-Seq involves introducing a concentration range of a labelled TF into isolated nuclei. Binding dependence on concentration is subsequently assessed for each sample, allowing the quantification of apparent binding affinities throughout the genome. However, accurately measuring the impact of the chromatin environment on interactions between transcription factors and binding remains an open challenge (Neikes et al 2023).

3. Proposed Approach and System Design

3.1 Datasets

The Dialogue for Reverse Engineering Assessments and Methods (DREAM5) datasets used in our methodology has data tabulated column wise. The first column is the id and provides the name of the TF and the second column indicates the array type. Then, the third column is the probe sequences and subsequent columns provide background data and signal. The training.txt contains Protein Binding Microarray (PBM) data for 66 TF's. The TF site sequences are indicated by model, training and scoring. Moreover, testing.txt contains PBM data for 20 TFs. An important observation is that the data is not normalized with respect to DNA on each spot per slide.

3.2 System Architecture

Figure 1 shows the overall system design for TF binding score prediction from DNA gene sequences datasets using the proposed CnNet approach. The sequence specificity of DNA binding site is predicted using high throughput assay. CnNet method is used in two phases; 1) motif selection phase based on different lengths (4 to 24) using MEME algorithm, and 2) the discovered motif is passed on to the neural network for finding the sequence specificity.

In the first phase, the motifs are scored based on the model parameters, which has been elucidated in the appendices of the MEME algorithm. The highest motifs are based on selected scoring for prediction of binding sites. The final step in motif detection is to obtain the highest scoring motifs among all the overlapping subsequences. During this process, the user can specify the length of motif detection (4 to 24) and also specify the maximum Hamming Distance (D) that is permitted to also evaluate space dyed motifs. A negative factoring step is also incorporated to avoid detection of the same motif over several iterations of the algorithm. The output is in the form of PWM and is passed on to the neural network layers.

In the second phase, a CNN is used, which is a neural network that can process different length motifs. Initially, the convolutional layer is very useful for extracting specific subsequences. output is then pooled using Max-pooling function and finally the values are fed into a network layer. In convolutional layer, ReLU activation function has been used for intensification of the non-linearity results. Then, Max pooling function has been used for extracting the maximum values from the layer. The weights and biases of the neural network are fine-tuned using the backpropagation algorithm based on the predicted values. Stochastic Gradient Descent has been used for minimizing the loss function in the neural network. As executed in several deep learning tasks, hyper-parameter tuning is undertaken to avoid overfitting of the data specific to the training set as well as the testing set. Moreover, Random Sampling Algorithm has been used for hyper-parameter optimization (Antikainen et al 2022 & LeCun et al. 2015). Thus, the CnNet method provides TF binding sites using several processes.

3.3 Motif Scanning and Detection

Our proposed method starts with the detection of motifs using a modification of the MEME algorithm. Initially, the input DNA sequences for this CnNet method are given in a batch basis 'b' (No of sequences (Sn)/ batch size). This allows to control how many predictions to make at one time for given sequences, i.e., given 'n' strings (S1, S2, S3, S4 …., Sn) of fixed length, with each string varying over the alphabets {A, C, G, and T}. Given two integers 'l' and 'd', such that 'l' is the length of our motif and 'd' is the HM, find all strings 'x' such that |x| = l and every other input string has a variant of 'x' at a HM of at most 'd'.

The PWM is the first stage for identification of motif in DNA sequences. Once, the PWM is calculated, we make use of the MEME algorithm to find the motifs with a batchwise input scheme.

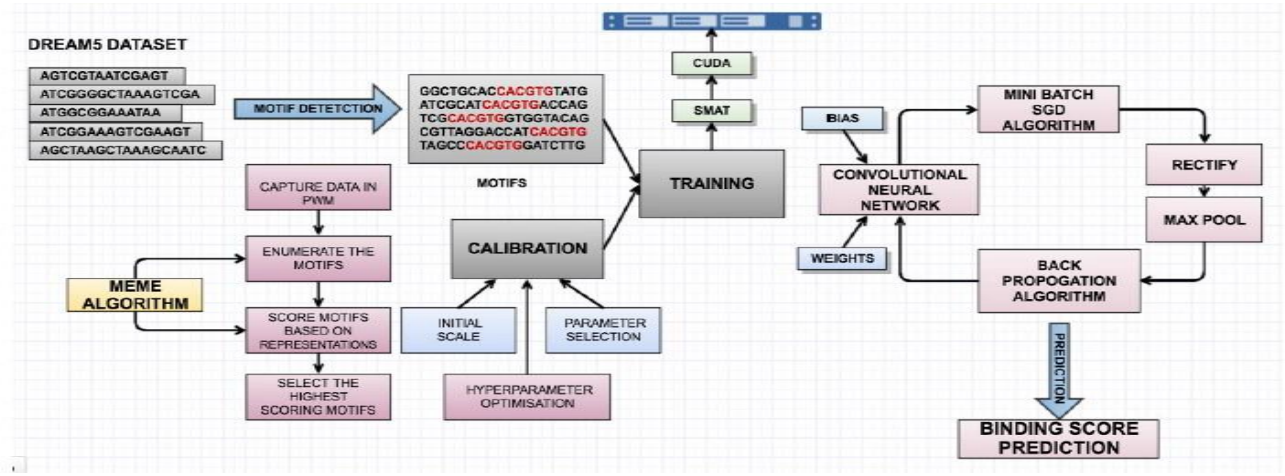

Fig. 1 System Architecture for TF Binding Score prediction using the CnNet Approach

MEME Procedure: *Input DNA sequence b (S_n/batch size)*
for b = to $pass_{max}$ do
for W = W_{min} to W_{max} by x $\sqrt{2}$ do
for $\lambda^{(0)} = \lambda_{min}$ to λ_{max} by x 2 do
Choose $\theta^{(0)}$ given W &$\lambda^{(0)}$
Run EM to convergence from the chosen
Value of $\emptyset^{(0)} = (\emptyset^{(0)}, \lambda^{(0)}, W$
Remove the outer columns of the consensus
Apply palindrome constraints to max G($\emptyset$)
Update the prior probability $U_{i,j}$ to the approximate consensus
End

Initially, the maximum and minimum motif lengths must be determined because the PWM is calculated, based on the length (λ_{min} =4 to λ_{max} = 24) (Fan et al. 2015). W_{min} to W_{max} are set to values depending on the LRT heuristic function. If Mi (i= 1…...n) is a discrete random variable with a parameter vector P_i (i=1……n) respectively, then M_i = (M_1, M_2… M_n) and the width is W. M can be considered to be a random variable whose instance sequences of length are l. An occurrence of M is a sample taken according to the distribution of M. In other words, an occurrence of Q_i is a sequence (q_1, q_2……q_i) where q_n is a sample from the discrete random variable M_i~ discrete (P_i). Thus, each discrete random variable defines the probability of seeing each possible letter of the sequence nucleotides at that position in an actual occurrence of the matrix.

Since the nucleotide b_i, at its position in the occurrence is an independent sample from the discrete random variable M_i, the fact that Q_i = a has no effect on the nucleotide at another position, j, in the motif. More precisely, for ($1 \le i \le$ W) and for all (a, bϵ δ)

$$P_r(b_j = b | b_i = a) = P_r(b_j = b), \quad j \neq i, \quad (1 \le i \le W) \tag{1}$$

As stated earlier, positions in a sequence that are not occurrences of a motif are termed background positions. Each position in a sequence which is not a motif occurrence is thus an independent sample from M_0~ discrete (P_0). The spacing factor √2 for the width is to be tried by MEME as well as large spacing between widths such as a factor of 2.

3.4 Finding DNA sequence specificities

In the second experiment, the MEME techniques have been used along with the CNN technique. Compared with earlier methods, the CNN technique has given good results. Moreover, the CNN technique comprises input and hidden and output layers. Convolutional, pooling, fully connected and normalization layers are the most important processes in the CNN. Each layer is comprising a number of neurons, and each neuron has a set of weights and biases to be learnt. Each neuron receives a number of inputs, sets a different weight for each and finally gives an output in single values. The neurons in each layer reflect the higher-level abstraction features of the neurons in the previous layers. The overall network gives different binding scores that are given to each sequence based on the output. Finally, the CNN model gives a predicted score.

3.4.1 Convolutional Layer

Convolutional layer function is based on mathematical processes which process the two variables f *and* g, input and produce the output of numerical values. Subsequently, the output values will pass to another set of neurons. This process is calculated based on weight and bias. Formally, the convolution operator * is defined as in the Equation given below,

$$(f * g)(t) = \int_{-\infty}^{\infty} f(x) g(t - x) dx \tag{2}$$

The convolutional layer is based on non-linear functions which means this process adopts the ReLU activation function. This CnNet approach has selected only one feature using multiple filters or kernel in the convolutional layer and find out the highest values. More concretely, let I be the input array of N dimensions and K the kernel, then the output at $i = (i_1, \ldots\ldots, i_N)$ is calculated by using the given equation,

$$(I * K)(i) = \sum_j I(j) K(i - j) \tag{3}$$

Where $i = (j_1, \ldots\ldots, j_N)$ is a coordinate of the input. It is to be noted that the convolutions at the borders present an edge

case, since they may need coordinates which are not part of the input.

3.4.2 Neural Network

The neural network system analyzes particular input features and gives the predicted output. This system provides approximate and possible values for consideration in terms of the previous value. An input, G→X, may denote features for specific data and Y is the predicted output for a particular X. Through this system, multiple prediction values can be obtained from the same input sample based on different features. This system comprises different layers, which means that successive layers will be followed by input layers, with each layer containing a large number of neurons. These multiple layers represent the problems of input data (LeCun et al. 2015).

The weight matrix, $W \in L^{m X n}$ is calculated for each layer with the total number of output (m) and input (n) nodes. For a particular input (X) sample, w_{ij}, the connection weight input nodes are i to j. The neural network system is vital for calculating the bias (b) between the nodes and the central property, and it is differentiable. The neural network system has bias values for each parameter, and the values vary, depending on how the output values are related to the previous values. Using this gradient information, the trainable parameters can be updated by taking small steps to minimize the distance function. The inputs for each node are multiplied by the weights in the incoming edge and these values are accompanied by the bias values. Here, the representation of the layer, L and the affine transform, $h_{\mathcal{L}}$, is calculated using the given equation,

$$h_{\mathcal{L}} = \sum W_{\mathcal{L}} h_{\mathcal{L}-1} + b_{\mathcal{L}} \quad (4)$$

Calibration is done using "hyper-parameter search". Calibration is very important in the neural network because it solves the under-fit and over-fit problems in aneural network. Accordingly, each calibration trial is tested and compared with its trained models to give accurate results. Moreover, millions of DNA sequences are used; hence the dataset will be large and take huge amount of time in the computation process to find calibration parameters. Subsequently, well-known Random Search techniques have been used for hyperparameter optimization. A randomized search just samples the parameters that fix the number of times in the search space rather than performing an exhaustive search.

Let f: Gn -> G be the cost function, which must be minimized. Let i € Gn, designate a position or candidate in the search-space then the algorithm is as follows,

Random Search Algorithm: -

Begin
For initialize i with random position
Sample position j from hypersphere
current position i, using log of uniform sampler
If f(j) < (i),
Then move to new position i = j
End

To train the CNN that uses sequences to make predictions from huge datasets, our neural network has four computational stages. Moreover, each stage has weights and biases which are continually updated based on the target score. Our neural network takes the output from the pooling stage and then tunes the weights and biases to match the target file and set the weights and biases for a given model.

A network has fixed threshold 'a' and weight 'w' respectively. The CnNet approach calculates the prediction score f(s) from the initial convolutional layer to the final output layer, as given by equation (5),

$$f(s) = netw\ [z[\ recta\ [conv\ (s)]]] \quad (5)$$

The prediction score f(s) is based on the dot-product scoring algorithm. In the batch gradient descent, this uses all the set of 'm'examples in each iteration. In Stochastic gradient descent only one example is used in the iteration. The mini-batch SGD algorithm uses both these ideas and takes in 'b' examples in one iteration, where 'b' is the mini-batch size.

The mini-batch gradient decent performs better than the batch gradient decent algorithm and allows us to make progress much faster. When compared with stochastic gradient decent method, the concept of vectorization helps the mini-batch SGD method. In particular, mini-batch gradient decent is likely to outperform stochastic gradient decent, only when a good vectorized implementation is available for that model.

Mini-Batch SGD Algorithm: -

L_i (w) = [Σ (Δ (f (i), (i)) + Δ (w 1, w 2)] /N,
 # Here w 1 and w 2 are the changes observed in the weights of the neural network layer.
Δ (f, t) = [(f− t) 2]/N.
MSGD Algorithm
Say b = 100, m = 1000
Begin
For i = 1, 101, 201….901
Θ: = Θ - η ▽(L i (w))
L i (w) = [Σ (Δ (f (i), (i)) + Δ (w 1, w 2)]/N
End

Rectification and pooling stages are used for intermediate processing and reducing the parameter dimensionality respectively. The ReLU activation function is defined by the following equation,

$$f(x) = \max(0, x) \quad (6)$$

where, x is the input to the neuron. The rectification stage takes the output of motif detection phase (M). Similarly, it calculates a value for similar size R = r(M), where a = (a_1, …, a_d) are tunable thresholds.

After convolution and rectification, pooling function will perform its operation on every layer independently and will reduce the dimensionality of the model. For this purpose, the max pooling method has been applied. The most common form of pooling layer is a filtered version or sub-array with size 3x3. The max pooling values have been identified by using given equations,

$$z = max\_pool(Y) \quad (7)$$

where, z_k = max ($Y_{1,k}$,⋯.$Y_{n,k}$), for each 1 ⩽ k ⩽ d. The backpropagation has been found to work faster than earlier approaches during the learning process. In CnNet, the gradients are efficiently computed during backpropagation. The neural network is 'N' (t+1) vector of weights 'n' and these two vectors are processed to generate the final score by using the following equation,

$$\text{Score (s)} = (w_d+1) + \Sigma w_k z_k, \quad 1 < k < d \tag{8}$$

where weight w_k is the weight of z_k contribution towards the score, and an additional bias term is also added. Finally, the highest score indicates a strong binding and the scores themselves are on an arbitrary scale.

4. Experimental Results and Evaluation

This section discusses the two different results offered by the MEME technique and Convolutional Neural Network. Keras, has been used to enable fast experimentation with the neural networks. A GPU is used in training the model as it has a number of cores which work in parallel processing. It reduces the time of training the data, compared to the CPU (Central Processing Unit). In our model, GPU (Nvidia GeForce GTX 1650) has been used with CUDA version 10.0.

4.1 MEME Technique

In the first category, the objective is to measure the biological relevance of the motifs discovered using the MEME algorithm. Subsequently, the agreement between the discovered motif and the known motif is measured by the predicted value in the training set. The parameters, ∅, of the sequence model discovered on a particular pass are converted by MEME into a log-odds scoring matrix, LO, and threshold, 't', where $LO_{x,j} = \log(p_{x,j}, p_{x,j})$ for $j = 1, \ldots \ldots, W \& x \in \delta, t = \log\left(\frac{1-\lambda}{\lambda}\right)^3$. The scoring matrix and threshold are used to score the sequences in a test set of sequences for which the positions of motif occurrences are known. The starting point is taken for each sequence based on the LO score matrix and the positions of the hits are compared to the position of the known occurrence.

The MEME compares favorably in terms of finding multiple motif widths, while the Gibbs sampler algorithm (Shida 2006) requires little prior knowledge. The Gibbs sampler algorithm can find multiple motifs in DNA or protein sequence sets but requires that the number of occurrences of each motif be specified individually for each sequence in the dataset. Initially, the motif Length L (4 to 24) was set for the MEME technique. Only one motif was searched from the MEME and the Gibbs sampler method. The ROC curve is plotted on the basis of the TPR & FPR. Both methods have been compared, based on ROC curves, and shown in Figure 2.

TABLE 1

A comparison of the ROC, recall and precision values for the MEME & Gibbs sampler technique

Model	*ROC*	*Recall*	*Precision*
MEME	0.93	0.95	0.92
Gibbs Sampler	0.81	0.91	0.89

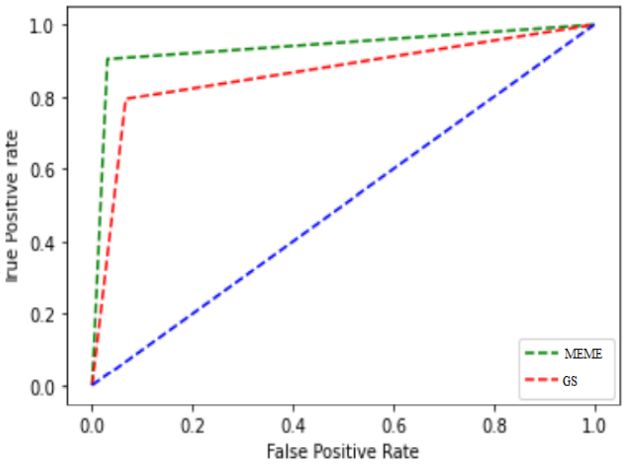

Fig. 2. Performance graph for MEME & Gibbs sampler technique based on ROC.

Table 1 shows the ROC, recall and precision values for single motif discovery. The MEME algorithm considered a similar motif from gene sequences dataset. The results of the Gibbs sampling algorithm are quite similar to those of the MEME, based on the ROC, recall and precision values. The CnNet model has used the MEME technique for motif selection phase because the MEME technique has given good results and also finds a motif in the best time possible, compared to the Gibbs sampler algorithm. Subsequently, the MEME algorithm has identified motifs of different lengths. Finally, the motif of the different length obtained through MEME algorithm, is given to the neural network system.

4.2 Convolutional Neural Network

In the second category, different length of motif has been given to the convolutional neural network to find the TFBS score. The model has used separate paths for different features and merged them at a higher level, where the combined features are processed using a fully connected layer. Similar to the sequence processing path, the shape can also be processed using convolutional/pooling layers. Instead of building a model for each TF, the process of predicting binding sites for the TF can be combined. In particular, for each (TF) task, there are three nodes present. The first node represents the fact that the region is an unknown binding site. The second node represents non-binding, and, lastly, the third node represents binding. The error is then calculated using a weighted cross entropy loss function, where the weight of unknown sites is set at 0, non-binding at 1 and binding at the ratio of non-binding to binding.

4.2.1 Calibration Parameter

The calibration phase evaluates the quality of each parameter set by a 3fold cross-validation on the training set. Each model is trained on a different 2/3 of the data and its performance evaluated on the other 1/3 data held-out. The calibration parameters are scored by averaging the three values (Alipanahi et al. 2015). Once the best calibration parameters have been identified and the top six models picked, the model with the best training performance is chosen as the final model returned by the entire pipeline. The six models are trained to ensure that the final training

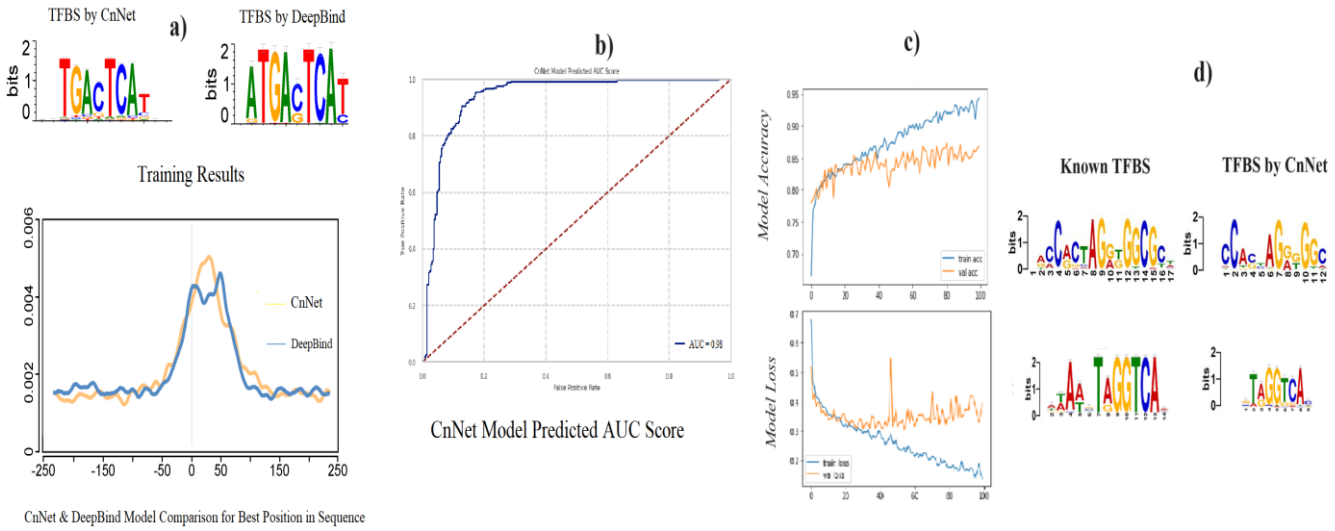

Fig 3 Visualization of the models learned on the ChIP-seq Sequences Dataset a) CnNet and DeepBind model best Transcription Factor Binding Sequences with Position b) AUC curve for CnNet approach c) Model Accuracy and Loss d) Some of the binding site acquired through CnNet Approache with actual known TFBS.

Fig 4. Predicted all the TF binding score using CnNet model from gene sequences dataset, a) Area Under Curve Score, b) Pearson Correlation Coefficients Score.

4.2.2 *Learning Momentum*

The simple method to speed up learning by increasing the rate of learning for every parameter in training is called the momentum method or accelerated gradient. If a particular parameter keeps increasing at every step, it is more likely to increase in future steps and hence the step size of that particular parameter should be scaled up. In practice, a traditional momentum rate (0.9) will speed up training, and the prediction performance of the final network. Several momentum methods have been proposed in mathematical optimization. This model has used the Nesterov Momentum with the rate coefficient sampled from the range [.95, .99].

4.2.3 *Number of learning steps*

A total of 20000 parameter update steps have taken place. The current performance of the trained model is noted at intervals of every 4000. The problem of overfitting can be identified if the result in the 5000th interval is better than the one achieved at the 20000th interval. This is termed as early stopping.

4.2.4 *Dropout expected value.*

The dropout expected value takes one of three values: 0.5 (strong), 0.75 (weak), or 1.0 (no dropout).

The starting point for activation maximization is chosen to be the matrix with 0.025 for all the entries. An L2-regularized gradient ascent with a learning rate of 0.01, using 1000 iterations, is applied to find an input which maximizes a particular class, i.e., a sharp or broad peak. It is to be noted that the activation before applying softmax is maximized, since the softmax output can be maximized by minimizing the probability for the other classes. For the DeepBind method, the reference is chosen to be the background frequencies of the nucleotides in the region of 2000 bp upstream and 200 bp downstream of the transcription start position (Swindell et al. 2012).

The representation learned for a particular TF can thus be leveraged to improve the representation for other TFs. To optimize the network parameters, the ADAM algorithm with a learning rate of 0.001 and a batch size of 256 has been used. Single-task models (models for an individual TF) were trained for a maximum of 500 epochs with a patience of 10. The multi-task model was trained for 50 epochs. Network

parameters were initialized using the Xavier initialization (Glorot & Bengio 2010). Before training, the training set was first partitioned into a training set (80%) and a validation set (20%). The network was trained for a maximum of 500 epochs with an early stopping patience of 10, using the validation loss as the benchmark data.

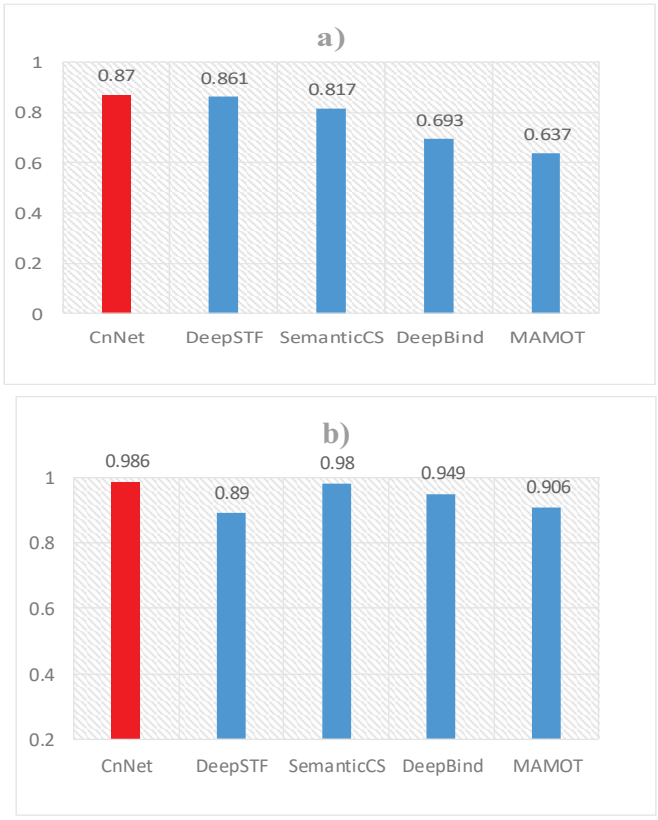

Fig. 5. Comparative performance analysis of CnNet Model with existing combined model a) Pearson Correlation Coefficient (PCC) Score, b) Area Under Curve (AUC) Score

4.3 Performance Evaluation

The model estimates the probability of the TFBS event occurring for a particular motif. The results of models have been evaluated by PCC and AUC.

4.3.1 Pearson Correlation Coefficient

The PCC measures the linear dependence among two variables, X & Y. A Pearson value 'r' is a score inclusive between +1 & -1, where +1 indicates a positive correlation, 0 indicates a no linear correlation and -1 indicates a negative correlation. The correlation is measured between the predicted probe intensity *p* and the actual intensity *a* using the (centered) Pearson correlation r, as given by the equation below,

$$r(p,a) = \frac{\sum_{i=1}^{N}(pi - \bar{p})\,(ai - \bar{a})}{\sqrt{\sum_{i=1}^{N}(pi - \bar{p})^2 \sum_{i=1}^{N}(ai - \bar{a})^2}} \quad (9)$$

where N is the total number of probe sequences in the array. '$\bar{p}$' indicates the mean probe intensity across all predicted probe intensities. '$\bar{a}$' indicates the mean across all actual probe intensities. Figure 4. (b) shows the Pearson correlation values for finding the TF binding site as the testing target by using the CnNet method.

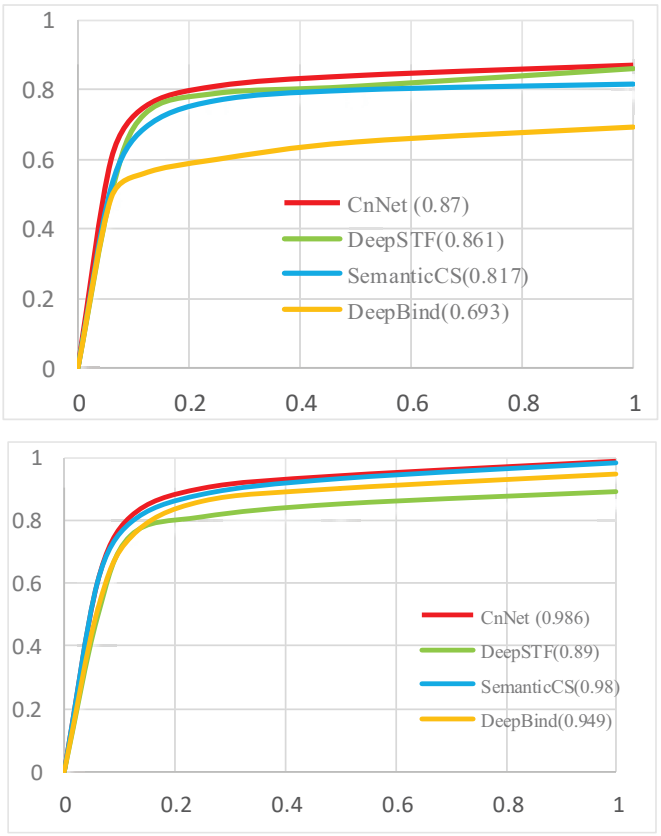

Fig. 6. Comparison between CnNet model with SemanticCS, DeepSTF and DeepBind based on PCC and AUC value.

4.3.2 Area Under the Curve

The AUC is used in classification analysis in order to determine which of the used models best predicts the classes when True Positive Rates (TPR) are plotted against False Positive Rates (FPR). In the present model, assuming there are 'n' probes under consideration, each probe is labeled 1, and the TPR and FPR are calculated as shown in the equation below,

$$TPR = TP / (TP+FN) \quad (10)$$

$$FPR = FP / (FP+TN) \quad (11)$$

where, TP = True Positive; TN = True Negative; FP = False Positive; FN = False Negative. Figure 4. (b) shows the AUC values for finding the TF binding site as the testing target by using CnNet method.

4.3.3 Experimental Comparison

The CnNet approach integrates the MEME technique with the CNN technique. Earlier approaches identified the TFBS using different combined algorithms. Table 2 shows some of the combined methods with AUC and PCC values.

TABLE 2

Combined models with PCC and AUC values

Models	***Methods used***	***PCC***	***AUC***	***References***
DeepSTF	CNN + Bi-LSTM	0.861	0.890	Ding et al. 2023
Semantic CS	Word embedding + CNN	0.817	0.98	Quan et al. 2020
DeepBind	Rankmotif + Neural Network	0.693	0.949	Alipanahi et al. 2015
FeatureReduce	PWM + dinucleotides and/or k-mers	0.693	0.693	Weirauch et al. 2013
MAMOT	PWM + HMMs	0.637	0.906	Schütz & Delorenzi 2008
RankMotif++	PWM + k-mers + Random Forest	0.518	0.975	Chen et al. 2007

This study utilizes two different algorithms, namely, CNN and the MEME technique. Initially, MEME algorithms were used to scan the motif in different length and thereafter give it to the neural network system. Figure 4 shows the CnNet method with existing model based on PCC and AUC score.

The MEME with neural network method helps to identify the TFBS from gene sequences dataset. The CnNet model is trained to create the best model with the least error. Moreover, the proposed model is created with little loss and saved in an HDF5 (Hierarchical Data Format 5) format so that it can be used to make predictions by loading weights onto the model. The proposed model has provided good results when compared to two well-known models, namely, DeepBind and DeepSEA.

The sequence logo is a graphical representation of the TFBS and it has been represented using WebLogo tools. The CnNet approach has given precise TF binding sites by using shallow models with large filters. The focus of this technique is to automatically set the features in the sequences, resulting in an improved performance. Identifying its sequence specificity in DNA sequences is a very challenging process and takes considerable time to find a single TF. Moreover, it's also better to find multiple TF than to simultaneously identify a single TF using computational methods.

5. Conclusion And Future Works

The proposed CnNet approach has predicted accurate TF binding score for any given sequence using the MEME with CNN technique. MEME has been used in the motif selection phase and CNN techniques for predicting the sequences specificity from gene sequences dataset. The discovery of TFBSs in gene sequences plays a key role in protein formation. The properties of a protein can be defined by the TFBS. From a biomedical point of view, generally speaking, finding drugs that target a specific protein pose a huge challenge. Our intention, going forward, is to identify drugs for particular diseases in line with computational methods using the TFBS. Furthermore, the receptors and TF are essential components in the response to different viruses and diseases. Recently, SARS-CoV-2 is a new virus which affected the human respiratory system. The corona virus disease can be controlled by targeting the immune system and by improving the dysfunctional immune system. The infection can be easily targeted by identifying the TF binding site of cells in the respiratory system. Similarly, the same approach can be used to easily find new drugs for all different types of diseases.

References

[1] Alipanahi, B, Delong, A, Weirauch, MT & Frey, BJ. Predicting the sequence specificities of DNA-and RNA-binding proteins by deep learning. Nature biotechnology, vol. 33, no. 8, pp. 831-836, 2015.

[2] Annala, M, Laurila, K, Lähdesmäki, H & Nykter, M. A linear model for transcription factor binding affinity prediction in protein binding microarrays. PloS one, vol. 6, no. 5, pp. e20059, 2011.

[3] Antonino, T & Villa, O. Accelerating DNA analysis applications on GPU clusters. Application Specific Processors (SASP). IEEE 8th Symposium on. IEEE, vol. 13, pp. 71-76, 2010.

[4] Antikainen AA, Heinonen M, Lahdesmaki H. Modeling binding specificities of transcription factor pairs with random forests. BMC Bioinform, 23(1):212, 2022.

[5] Badis, G, Berger, MF, Philippakis, AA, Talukder, S, Gehrke, AR, Jaeger, SA, Chan ET, Metzler, G, Vedenko, A, Chen, X, Kuznetsov, H, Wang, CF, Coburn, D, Newburger, DE, Morris, Q, Hughes, TR & Bulyk ML. Diversity and complexity in DNA recognition by transcription factors. Science, vol. 324, no. 5935, pp. 1720-1723, 2009.

[6] Bailey TL and Elkan C. Unsupervised learning of multiple motifs in biopolymers using expectation maximization. Mach. Learn., vol. 21, pp. 51–80, 1995.

[7] Bailey, TL. DREME: motif discovery in transcription factor ChIP-seq data. Bioinformatics, vol. 27, no. 12, pp. 1653-1659, 2011.

[8] Bailey, TL, Williams, N, Misleh, C & Li, WW. MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic acids research, vol. 34, no. 2, pp. W369-W373, 2006.

[9] Barany, F. Genetic disease detection and DNA amplification using cloned thermostable ligase. Proceedings of the National Academy of Sciences, vol. 88, no. 1, pp. 189-193, 1991.

[10] Berger, MF, Philippakis, AA, Qureshi, AM, He, FS, Estep, PW & Bulyk, ML. Compact, universal DNA microarrays to comprehensively determine transcription-factor binding site specificities. Nature biotechnology, vol. 24, no. 11, pp. 1429-1435, 2006.

[11] Boone, K, Wisdom, C, Camarda, K, Spencer, P & Tamerler, C. Combining genetic algorithm with machine learning strategies for designing potent

antimicrobial peptides. BMC bioinformatics, vol. 22, no. 1, pp. 1-17, 2021.

[12] Bulyk, ML. Computational prediction of transcription-factor binding site locations. Genome biology, vol. 5, no. 1, pp. 201-211, 2003.

[13] Chen, X, Hughes, TR & Morris, Q. RankMotif++: a motif-search algorithm that accounts for relative ranks of K-mers in binding transcription factors. Bioinformatics, vol. 23, no. 13, pp. i72-i79, 2007.

[14] Chen, C, Hou, J, Shi, X, Yang, H, Birchler, JA, & Cheng, J. DeepGRN: prediction of transcription factor binding site across cell-types using attention-based deep neural networks. BMC bioinformatics, vol. 22, no. 1, pp. 1-18, 2021.

[15] Ding, Pengju, Yifei Wang, Xinyu Zhang, Xin Gao, Guozhu Liu, and Bin Yu. "DeepSTF: predicting transcription factor binding sites by interpretable deep neural networks combining sequence and shape." Briefings in Bioinformatics (2023): bbad231.

[16] Faheem, HM. Accelerating motif finding problem using grid computing with enhanced brute force. In Advanced Communication Technology (ICACT), The 12th International Conference on IEEE, vol. 1, pp. 197-202, 2010.

[17] Fan, Y, Wu, W, Yang, J, Yang, W & Liu, R. An algorithm for motif discovery with iteration on lengths of motifs. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), vol. 12, no. 1, pp. 136-141, 2015.

[18] Felicioli C and R. Marangoni. BpMatch: An Efficient Algorithm for a Segmental Analysis of Genomic Sequences. IEEE/ACM Transactions on Computational Biology and Bioinformatics, Vol. 9, No. 4, 2012

[19] Feuk, L, Carson, AR & Scherer, SW. Structural variation in the human genome. Nature Reviews Genetics, vol. 7, no. 2, pp. 85-97, 2006.

[20] Laurila, K, Yli-Harja, O & Lahdesmaki, H. A protein–protein interaction guided method for competitive transcription factor binding improves target predictions. Nucleic acids research, vol. 37, no. 22, pp. e146-e146, 2009.

[21] Lawrence, CE, Altschul, SF, Boguski, M, Liu, JS, Neuwald, AF & Wootton, JC. detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. Science, vol. 262, no. 5131, pp. 208-214, 1993.

[22] LeCun Y, Bengio Y, and Hinton G. Deep learning. Nature 521.7553, PP 436-444, 2015.

[23] Linhart, C, Halperin, Y & Shamir, R. Transcription factor and microRNA motif discovery: the Amadeus platform and a compendium of metazoan target sets. Genome research, vol. 18, no. 7, pp. 1180-1189, 2008.

[24] Li Y, Quan LJ, Zhou YT, et al. Identifying modifications on DNA-bound histones with joint deep learning of multiple binding sites in DNA sequence. Bioinformatics, 38(17):4070–7, 2022.

[25] Liu, LA & Bradley, P. Atomistic modeling of protein–DNA interaction specificity: progress and applications. Current opinion in structural biology, 22(4), 397-405, 2012.

[26] Luo, J, Li, G, Song, D & Liang, C. CombiMotif: A new algorithm for network motifs discovery in protein–protein interaction networks. Physica A: Statistical Mechanics and its Applications, vol. 416, pp. 309-320, 2014.

[27] Mann, MJ & Dzau, VJ. Therapeutic applications of transcription factor decoy oligonucleotides. The Journal of clinical investigation, 106(9), 1071-1075, 2000.

[28] Manioudaki, ME & Poirazi, P, 'modeling regulatory cascades using Artificial Neural Networks: the case of transcriptional regulatory networks shaped during the yeast stress response', Frontiers in genetics, vol. 4, pp. 110-125, 2013.

[29] Morishita, R., Higaki, J., Tomita, N., & Ogihara, T. Application of transcription factor "decoy" strategy as means of gene therapy and study of gene expression in cardiovascular disease. Circulation research, 82(10), 1023-1028, 1998.

[30] Needleman, SB, Wunsch, CD. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol., vol. 48, no. 3, pp. 443-453. doi:10.1016/0022-2836(70) 90057-4. PMID 5420325, 1970.

[31] Neikes, H.K., Kliza, K.W., Gräwe, C., Wester, R.A., Jansen, P.W., Lamers, L.A., Baltissen, M.P., van Heeringen, S.J., Logie, C., Teichmann, S.A. and Lindeboom, R.G.,. Quantification of absolute transcription factor binding affinities in the native chromatin context using BANC-seq. Nature Biotechnology, 1-9, 2023.

[32] Ondov, BD, Cochran, C, Landers, M, Meredith, GD, Dudas, M & Bergman, NH. An alignment algorithm for bisulfite sequencing using the Applied Biosystems SOLiD System. Bioinformatics, vol. 26, no. 15, pp. 1901-1902, 2010.

[33] Overington JP, Al-Lazikani B, Hopkins AL. How many drug targets are there?" Nature Reviews. Drug Discovery. 5 (12): 993–6. doi:10.1038/nrd2199. PMID 17139284, 2006.

[34] Pavesi G, P. Mereghetti, G. Mauri, and Pesole G. Weeder web: Discovery of transcription factor binding sites in a set of sequences from co-regulated gene. Nucleic Acids Res., vol. 32, pp. 199–203, 2004.

[35] Payne, DJ, Gwynn, MN, Holmes, DJ & Pompliano, DL. Drugs for bad bugs: confronting the challenges of antibacterial discovery. Nature Reviews Drug discovery, vol. 6, no. 1, pp. 29-40, 2007.

[36] Pearson, WR. An introduction to sequence similarity ("homology") searching. Current protocols in bioinformatics, 42(1), 3-1, 2013.

[37] Quan, Lijun, Xiaoyu Sun, Jian Wu, Jie Mei, Liqun Huang, Ruji He, Liangpeng Nie, Yu Chen, and Qiang Lyu. "Learning useful representations of DNA sequences from ChIP-seq datasets for exploring transcription factor binding specificities." IEEE/ACM Transactions on Computational Biology and Bioinformatics 19, no. 2: 998-1008, 2020.

[38] Quang, D, & Xie, X. FactorNet: a deep learning framework for predicting cell type specific transcription factor binding from nucleotide-resolution sequential data. Methods, 166, 40-47, 2019.

[39] Rajesh, S, Prathima, S & Reddy, LSS. Unusual Pattern Detection in DNA Database Using KMP Algorithm.

International Journal of Computer Applications, vol. 12, no. 2, pp. 197-202, 2010.

[40] Reddy T, DeLisi C, and Shakhnovich B, "Binding site graphs: A new graph theoretical framework for prediction of transcription factor binding sites," PLoSComput. Biol., vol. 3, no. 5, pp. 0844– 0854, 2007.

[41] Robinson, MD, McCarthy, DJ & Smyth, GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics, vol. 26, no. 1, pp. 139-140, 2010.

[42] Sanger, F & Tuppy, H. The amino-acid sequence in the phenylalanyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. Biochem. J, vol. 49, no. 4, pp. 463-481. doi:10.1042/bj0490463, 1951.

[43] Sanger, Frederick, Nicklen S, and Coulson AR. DNA sequencing with chain-terminating inhibitors. Proceedings of the national academy of sciences 74.12, 5463-5467, 1977.

[44] Schütz, F & Delorenzi, M. MAMOT: hidden Markov modeling tool. Bioinformatics, vol. 24, no. 11, pp. 1399-1400, 2008.

[45] Sharon, E, Lubliner, S & Segal, E. A feature-based approach to modeling protein–DNA interactions. PLoS Computational Biology, vol. 4, no. 8, pp. e1000154, 2008.

[46] Sohn, I, Shim, J, Hwang, C, Kim, S & Lee, JW. Informative transcription factor selection using support vector machine-based generalized approximate cross validation criteria. Computational Statistics & Data Analysis, vol. 53, no. 5, pp. 1727-1735, 2009.

[47] Stormo, GD. DNA binding sites: representation and discovery. Bioinformatics, vol. 16, no. 1, pp. 16-23, 2000.

[48] Weirauch, MT, Cote, A, Norel, R, Annala, M, Zhao, Y, Riley, TR, Rodriguez, JS, Cokelaer, T, Vedenko, A, Talukder, S, Consortium, D, Bussemaker, HJ, Morris, QD, Bulyk, ML, Stolovitzky, G & Hughes, TR. Evaluation of methods for modeling transcription factor sequence specificity. Nature biotechnology, vol. 31, no. 2, pp. 126-134, 2013.

[49] Yaman, Oğuz Ulaş, and Pınar Çalık. "MachineTFBS: Motif-based method to predict transcription factor binding sites with first-best models from machine learning library." Biochemical Engineering Journal 198,108990, 2023.

[50] Yu YT, Ding PJ, Gao HL, et al. Cooperation of local features and global representations by a dual-branch network for transcription factor binding sites prediction. Brief Bioinform, 24(2):bbab036, 2023.

[51] Zhang, Y., Wang, Z., Zeng, Y., Zhou, J., & Zou, Q. High-resolution transcription factor binding sites prediction improved performance and interpretability by deep learning method. Briefings in Bioinformatics, 2021.

[52] Zhang YQ, Wang ZX, Zeng YQ, et al. A novel convolution attention model for predicting transcription factor binding sites by combination of sequence and shape. Brief Bioinform , 23(1):bbab525, 2022.

[53] Zhou, J & Troyanskaya, OG. Predicting effects of noncoding variants with deep learning–based sequence model. Nature methods, vol. 12, no. 10, pp. 931-934, 2015.

Mohamed Divan Masood is Assistant Professor in the School of Computer Science in B S Abdur Rahman Crescent Institute of Science and Technology at India. He received his PhD degree in Computer Science and Engineering from Anna University, India in 2018. Currently his research group to developing the model for automation system and intelligent drug design.

Manjula is currently a professor in the department of Computer Science and Engineering in Vellore Institute of Technology, India. She has completed PhD degree in Computer Science and Engineering at Anna University in 2004. Dr Manjula has published more than 150 papers. She presided over 4 scientific research government projects, India. Currently, she has working on computational biology and drug discovery for covid.

Vijayan Sugumaran is currently a Distinguished University Professor of Management Information System, Chair of the Department of Decision and Information Sciences, and Co-Director of the Center for Data Science and Big Data Analytics, Oakland University, Rochester, MI, USA. He has authored and co-authored more than 300 peer-reviewed articles in Journals, Conferences, and Books such as Information Systems Research, ACM Transactions on Database Systems, IEEE Transactions on Big Data, and IEEE Transactions on Education. He has edited 20 books and serves on the Editorial Board of eight journals. His research interests include big data management and analytics, ontologies, and semantic web, intelligent agent, and multiagent systems. Contact him at sugumara@oakland.edu.