Automatically determine regulatory motifs in DNA sequence data

B.K.S. Ishanka
Computer Science Department
University Of Moratuwa
Katubedda, Sri Lanka
shehan.20@cse.mrt.ac.lk

E.G.D. Sandaruwan
Computer Science Department
University Of Moratuwa
Katubedda, Sri Lanka
dinesh.20@cse.mrt.ac.lk

Y.H.S. Amarasiri
Computer Science Department
University Of Moratuwa
Katubedda, Sri Lanka
amarasiriyhs.20@uom.lk

***Abstract*—The discovery of DNA motifs can be considered as an important victory in biological applications. Various algorithms have been used to create fast and accurate motif discovery tools over the years. Consensus and probabilistic methods are the two types of algorithms that are commonly used. In this paper, we present experiment results based on some motifs discovery algorithms.**

Keywords—Bioinformatics, Motifs, Algorithms, protein binding, DNA segments, DNA sequences, Motif discovery, Transcription factors

I. Introduction

Genes are essential functional units in molecular biology that contain genetic information. Also genes can be used as templates for protein transcription. Transcription is carried out by the RNA polymerase enzyme and a group of proteins known as transcription factors. The binding of transcription factors (group of proteins) to a binding site (a DNA segment) on the genomic sequence is the first step in protein transcription. These short DNA segments that act as binding sites are called motifs. The motifs finding problem is a fundamental problem in molecular biology and motifs finding is very important for applications such as locating regulatory sites and identifying drug targets.

There are different types of motifs such as planted motifs, structured motifs, sequence motifs, gapped motifs and network motifs. Below figure shows a simple block diagram of the motif discovery technique.

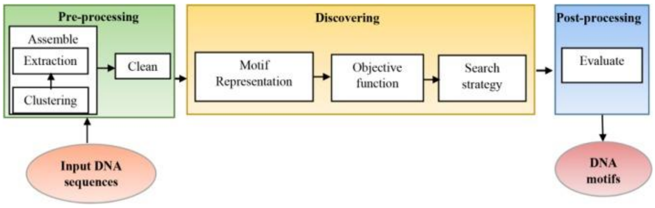

Figure 1: General block diagram of motif discovery techniques [5]

As seen in the diagram above, the technique for finding motifs is divided into three stages.

1. Pre-processing

   Under this step, do the assembly and clean of DNA sequences for accurate motif discovery. It is recommended to select as many target sequences that may contain motifs as possible during the assembling process. In the assembling stage, the input sequences are clustered based on some detail and desired sequences are extracted from the sequence database. After the assembling step, it is necessary to clean the input sequences to remove confounding sequences.

2. Discovering

   As a first step of the discovering stage, try to represent the sequences in different ways. Consensus string and Position-specific Weight Matrices are two ways to represent motifs. The appropriate objective function is defined after motif representation, and then the appropriate search algorithm is used.

3. Post-processing

   The resulting motifs are evaluated under this stage.

The remaining paper is structured as follows. In section 2, we present and discuss related literature. In section 3, we present our methodology. In section 4 we present and discuss the experimental results. In section 5, we conclude the paper with a summary of our results.

II. Literature Review

Identifying DNA motifs still remains as a difficult task for biologists and computer scientists. So the researchers have taken various approaches to develop motif discovery tools.

Motif finding algorithms can be divided into three categories based on the type of DNA sequence information used by the algorithm to deduce the motifs.

1. those that use promoter sequences from coregulated genes from a single genome
2. those that use orthologous promoter sequences of a single gene from multiple species (i.e., phylogenetic footprinting)
3. those that use promoter sequences of coregulated genes as well as phylogenetic footprinting

However, based on the combinatorial approach used in their design, most previous literature divided motif finding algorithms into two major groups.

1. word-based (string-based) methods

Word-based methods are mostly based on exhaustive enumeration such as counting and comparing oligonucleotide frequencies. Those enumerative methods ensure global optimality and they are suitable for short motifs. Also the word-based methods can also be very fast, when implemented with optimized data structures such as

[1] Code is available is at https://github.com/ShehanIshanka/motif-detector

suffix trees or parallel processing. However There are a number of drawbacks in this strategy.

- Since traditional transcription factor motifs also have many poor restricted locations, it needs to be post-processed with certain clustering systems.
- It has an issue with generating so many fictitious motifs.
- Long time processing since it checks all possible substrings

YMF, DREME, CisFinder, Weeder, FMotif, and MCES are examples of popular algorithms based on this approach.

YMF is optimized for yeast genomes, so it can't identify motifs that are too long. The DREME uses a simplified type of regular expression words to find multiple, short, non-redundant and statistically significant motifs in past manner.

2. probabilistic sequence models

Model parameters are estimated using maximum likelihood principle or Bayesian inference in probabilistic models. Also a position weight matrix [4] is used to describe the motif model in this method. Here, position weight matrices are often represented as a pictogram, with each position represented by a stack of letters whose height is proportional to the information content of that position. Probabilistic methods have the benefit of having less search parameters, but since they rely on probabilistic regulatory region models, those models can be extremely sensitive to minor changes in the input data. Many of the probabilistic approach's algorithms are designed to find longer motifs. However, since these algorithms use some type of local search , such as Gibbs sampling, expectation maximization (EM), or greedy algorithms, they cannot guarantee that they can find globally optimal solutions. MEME, STEM, EXTREME, AlignACE and BioProspector are examples of popular algorithms based on this approach.

MEME (Multiple EM for Motif Elicitation) is a well-known motif recognition program that uses the EM algorithm to optimize PWMs (Position Frequency Matrices). The MEME algorithm works by finding an initial motif and then improving it with expectation and maximization steps until the PWM values do not change or the maximum number of iterations is reached.

Gibbs sampling is a well-known stochastic method that is similar to the EM algorithm. It converges to a local optimum and is less dependent on initial parameters. AlignACE and BioProspector are the two algorithms which are based on Gibbs sampling.

III. METHODOLOGY

As proposed in the literature, there are several kinds of traditional and deep learning techniques to mine motifs from DNA sequences. Here we attempt to predict whether a certain motif is a RBP or not using classification algorithms. We have used traditional classifiers as well as neural network approach.

1) Logistic Regression

2) SVM

3) Neural Network approach

The dataset for this task is from genome-wide RNA-protein CLIP-seq data. It includes different kinds of representations for different sequence motifs with binary value implicating RNA binding proteins or not. Although the dataset includes representations for several proteins, only Ago/EIF protein is selected. The data set includes separate datasets for training

[2]Dataset is available is at https://www.kaggle.com/c/tweet-sentiment-extraction/data

and tests. Table I shows the statistics in each set in the dataset.

TABLE I. Dataset

	Class 0	Class 1	Total
Train	3200	800	4000
Test	800	200	1000

Before describing the approaches, the dataset has been preprocessed. Preprocessing steps are listed down below.

1. Retrieve sequence motif strings from the data sets.
2. Replace “U” with “T”.
3. Generate one hot encode vectos for the sequence motif with encode vectors of each DNA bases ; “A”, “C”, “G”, “T”

The preprocessed data set is used for the approaches mentioned above. The approaches are described below. In the neural network approach, LSTM network is implemented with 128 LSTM cells. The final two layers of the network are dense layers with 64 units and 3 units (classes) respectively. Relu function is used as the activation function except in the last dense layer activation function is softmax. Categorical cross entropy is used as the loss function and Adam optimizer is used as the optimization function. One hot encoded motif sequences are used as features and classes are transformed to one hot encoded vector.

TABLE II. Accuracies of Models

Approaches	Accuracy
Logistic Regression	0.787
SVM	0.8
Neural Network approach	0.80

TABLE III. Precision, Recall and F1 Score of Models

Approaches	Precision	Recall	F1 score
Logistic Regression	0.57	0.52	0.5
SVM	0.4	0.5	0.44
Neural Network approach	0.72	0.70	0.71

IV. EXPERIMENT

The experiment is done on a HP laptop with i5 processor and 16GB RAM. The dataset2 used for training and testing is extracted from CLIP-seq dataset.

In the neural network approach, 1000 batches are used and the model is trained for only 3 epochs due to performance issues. Each of the approaches mentioned in the methodology section is tested and the results are provided in table II and table III.

From all the approaches attempted, the neural network approach and SVM approach have the highest accuracy and logistic regression has the highest F1 score.

V. CONCLUSION

VI. ACKNOWLEDGMENT

This research is completed on behalf of the semester 4 module, CS5614 – Bioinformatics, of MSc in Computer Science specialized in Data Science Course of batch 2020. We would like to thank Dr. Charith Chitraranjan for the guidance to complete the research and the lectures on Bioinformatics.

VII. REFERENCES

[1] *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3962653/*

[2] https://www.nature.com/scitable/definition/transcription-dna-transcription-87/

[3] https://bmcbioinformatics.biomedcentral.com/articles/10.1186/1471-2105-8-S7-S21

[4] https://www.sciencedirect.com/science/article/abs/pii/0022283690902239?via%3Dihub

[5] https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6490410/