ISSN No:-2456-2165

Comparison of Various Motif Discovery Algorithms

Prince Joseph
Lecturer in Computer Engineering, Govt Polytechnic College Pala

Abstract:- Finding patterns in DNA sequences, a challenging task in molecular biology and computer science, plays a crucial role in understanding gene expression and transcription regulation. This article explores various computational methods for identifying transcription factor binding sites TFBS and regulatory motifs in DNA sequences. We discuss the significance of these motifs in gene expression and disease susceptibility research. The article categorizes motif discovery algorithms into enumerative, probabilistic, combinatorial, and nature-inspired approaches, highlighting their strengths and limitations. It emphasizes the need for a comprehensive motif discovery tool that can handle diverse motif types, perform global searches efficiently, and reduce user-defined parameters.

Keywords:- *Motif Discovery, Transcription Factor Binding Sites TFBS, Gene Expression, DNA Sequences, Computational Methods.*

I. INTRODUCTION

One of the most difficult challenges in molecular biology as well as computer science is finding patterns in DNA sequences. Identification of regulatory motifs is critical for understanding the gene expression. The essential concept in gene expression is that each a gene encodes the instructions for making a protein. The process of expression begins with the binding of several recognised protein factors. As transcription factors, they bind to enhancer and promoter sequences (Li and Li, 2019).

Transcription is the first stage, which involves creating an RNA "copy" of a section of the DNA. This RNA sequence is read and interpreted to create a protein in the second stage of the process, known as translation. Gene expression is the combined result of these two actions. Numerous regulatory transcription factors (TFs), also known as Transcription Factor Binding Sites (TFBS), bind to certain DNA regions to control gene expression. In the past ten years, a significant new method for understanding transcription regulation networks has emerged: the computational identification of TFBS through the study of DNA sequence data (Ruzicka et al., 2017).

Finding sequence motifs can be challenging since intergenic regions are extremely long and highly varied, while sequence motifs are small (approximately 6–12 bp). Sequence motifs are frequently repeated and conserved, and they have a fixed size. These patterns are critical for identifying Transcription Factor Binding Sites (TF-BSs), which aids in understanding the mechanisms governing gene expression 3. Motifs can be classified as planted, structured, gapped, sequence, network, and motifs (Hashim et al., 2019).

An important issue in computational biology is the finding of weak motifs. It is challenging to solve because there are so many inconsistencies between the actual theme and its altered variants that false signals may mask the real ones. Further, it is challenging to identify and uncover regulatory elements using computer algorithms since they are typically brief and varied. The task of solving the theme finding problem is that of discovering overrepresented motifs as well as conserved motifs from the set of DNA sequences that are good candidates for becoming sites where transcription factors bind. Transcription factor is a protein that functions as a gene expression regulator, specifically regulating the start of the transcription process that produces mRNA using DNA as a template. The common sequence is called a motif. A "pattern" in a transcription factor's binding sites. Finding motifs will aid in the development of illness therapies and comprehension disease susceptibility (Mohanty and Mohanty, 2020).

Many techniques for analysing gene function start with the finding of a DNA motif. Finding Transcription Factor Binding Sites (TFBSs), which aid in understanding the mechanisms for controlling gene expression, is a crucial part of motif discovery. The development of quick and precise motif discovery technologies has utilised a variety of algorithms over the years. These algorithms are typically categorised as probabilistic or consensus techniques, and many of them take a lot of time to run and are prone to get stuck in local optimums. Recently, solutions to these issues have been offered using both nature-inspired algorithms and a variety of combinatorial algorithms (Hashim et al., 2019).

II. MOTIF DISCOVERY METHOD

Given a group of genes with strikingly comparable expression profiles, the mechanism behind the coordinated behaviour of genes can be investigated. Co-regulation of transcription is thought to be a common source of co-expression. Co-regulated genes are known to have some regulatory mechanisms in common, presumably at the transcriptional level. As a result, their promoter regions may share some patterns that serve as binding sites of transcription regulators. In order to find these regulatory elements, it makes sense to look for statistically significant overrepresented patterns in the promoter area of a group of co-expressed genes.

ISSN No:-2456-2165

The motif discovery method is divided into two main steps:

A. *Pre processing*
B. *Discovering of the method*

A. *Pre processing*

By using clean processes and assembly, it is getting the DNA sequences ready for precise motif finding. It is recommended that the researcher should choose as many target sequences as you can in the assembly process that might include motifs, work to keep sequences as brief as you can, and eliminate any sequences that are unlikely to do so. The input sequences are first clustered based on some data, and the desired sequences are then extracted and placed in the proper sequence database. Consequently, it is vital to clean the input sequences to conceal or eliminate confusing sequences (Sun et al., 2015).

B. *Discovering of the method*

The motif discovery strategy, which starts by modelling the sequences, is what happens in the intermediate step. Position-specific Weight Matrices (PWM) and consensus strings are the two methods for representing the motifs. Consensus strings allow degenerate symbols to be included in a string using the IUPAC code and have the same length as DNA sequence motifs, whereas PWM is a 4xm matrix, where m is the length of the motif. The likelihood of each nucleotide appearing at each motif index point is represented by each position in the matrix. Following motif representation, the proper goal function is chosen, and then the right search method is used (Sun et al., 2015).

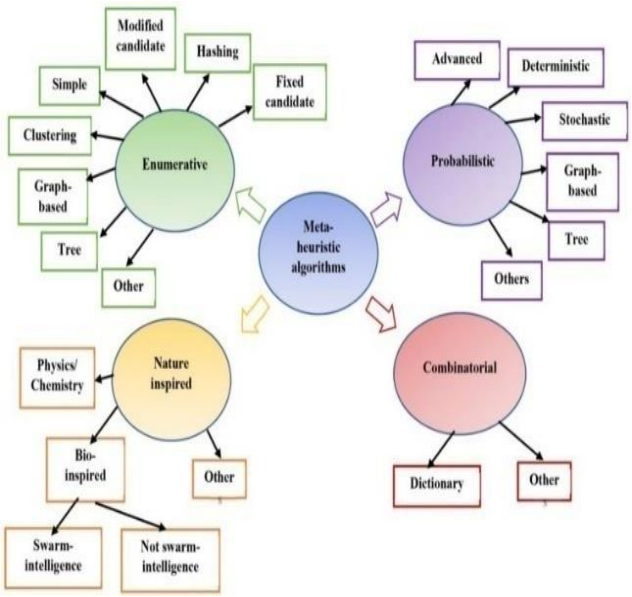

Fig 1: Different type of motif discovery algorithm (Sun et al., 2015)

➢ *Genetic algorithm for motif discovery*

We create a method for anticipating binding site motifs employing a genetic algorithm. The method investigates the search space of all potential starting places of the binding site motifs in various target sequences using a population that goes through evolution. It is based on the general framework of a genetic algorithm. Crossover participants compete for space in the population, and mutations happen with a particular probability. Initial tests showed that our method could produce good prediction accuracy in a short amount of calculation time. The fact that the calculation time of our approach does not explicitly depend upon the length of the target sequences and may not dramatically rise as the target sequences grow very long is a promising benefit (Mohanty and Mohanty, 2020).

➢ *Principles of motif discovering algorithm*

In order to distinguish biological signals from background noise, approximation algorithms often create a probability training model and score a statistical measure. Based on Gibbs sampling and MEME, an especially effective class of approximation algorithms is created. MEME discovers motifs by optimising the PWMs using Expectation Maximisation (EM), which continues to define three types of motif discovery sequence models: OOPS, ZOOPS, and TCM, which, respectively, correspond to one occurrence per sequence, one or two occurrence every sequence, as well as zero or more occurrences per sequence. Since they are straightforward and stable, probability training algorithms are frequently utilised (Zhang et al., 2016).

Enumeration strategy and probabilistic technique are the two main categories of motif finding algorithms. When solving the Panted (l, d) Motif Problem (PMP), which has a maximum number of mismatches (d) and a maximum length (l), the enumeration approach looks for consensus sequences; motifs can be predicted based upon the enumeration of words along with computing word similarities. The algorithms that utilise the word enumeration approach often discover the global optimum by exhaustively searching the whole search space to identify which ones emerge with potential substitutions (Hashim et al., 2019).

The word enumeration method can be sped up by employing specialised data structures like suffix trees with parallel processing. Probabilistic methods make up the second group. It creates a position-Specific Weight Matrix (PSWM) or motif matrix, a probabilistic model that uses minimal search parameters and specifies a distribution of bases for each location in TFBS to distinguish between motifs and non-motifs (Zhang et al., 2016).

A probabilistic strategy belongs to the second group. It creates a position-Specific Weight Matrix (PSWM), also known as a motif matrix, which specifies a distribution of bases for each position in TFBS to discriminate between motifs and non-motifs and only needs a small number of search parameters.

ISSN No:-2456-2165

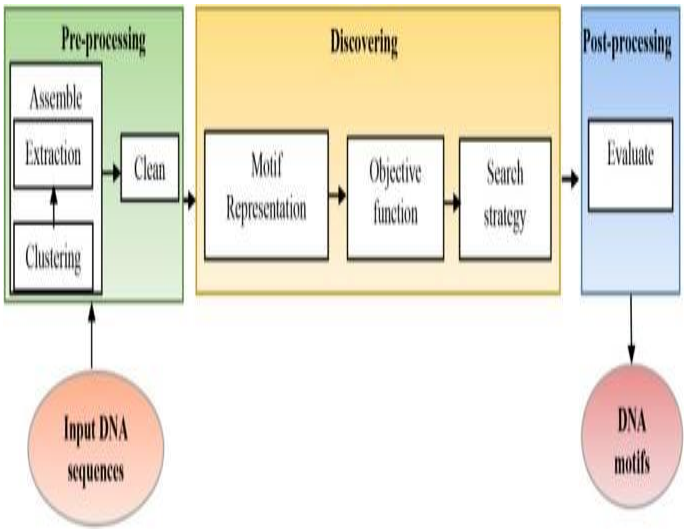

Fig 2: Steps involved in motif discovery process (Adapted from Hashim et al., 2019)

III. CLASSIFICATION OF ENUMERATIVE APPROACH

➢ *Word Count Method*

The first class is based on the simple act of counting words. YMF 134 and DREME 9, are two algorithms in this class that are currently in use. The YMF (Yeast Motif Finder) technique was created by Sinha et al. 29 and uses consensus representation to find short motifs in yeast genomes with a minimal number of degenerate sites. In the search space approach, YMF counts all motifs and calculates the z-score to produce the motifs with the highest z-scores. The Fisher's Exact test proposed DREME (Discriminative Regular Expression Motif Elicitation) algorithm to determine the relevance of motifs(Sun et al., 2010).

➢ *Clustering-based method*

The second class was presented as an alternative to employing two loops to find various motifs. In order to find small theme in big sequences (up to 50 Mb), it is further suggested a word clustering algorithm called CIS Finder. In order to calculate Position Frequency Matrices (PFMs) for n-mer word counts with and without gaps in both the test and control sets, one must first define nucleotide substitution matrices for each n-mer word. PFMs are stretched over flanking and gap areas, then clustered to produce non-redundant motifs (Bailey, 2010).

For exact terms (words without wildcards), our motif search is comprehensive; however, for words containing wildcards, it is heuristic in order to speed up the process. We use the Fisher's Exact Test to quantify the significance resulting from the relative enrichment for every motif in two sets of sequences in order to find statistically significant, discriminative motifs. The chance that the proportion of sequences in the first set that match the motif would be as high as what was observed (or higher) is calculated using the proportion of matching sequences in the second set (Bailey, 2010).

➢ *Tree-based method*

A tree-based search to quicken the word enumeration method is the third class. The Weeder algorithm, which is based on count matching patterns with the most severe mismatches. The motifs are initially represented using consensus sequence, and based on the difference between the k-mers of the input sequences and the consensus under a specified number of substitutions, k-mers are built. Each group is then assessed with a particular measure of significance. For discovering lengthy (l, d) motifs in big DNA sequences under ZOMOPS (Zero, one or multiple occurrence(s) of the motif instance(s) per sequence) constraints, the generated suffix tree was suggested in FMotif 11 algorithm (Sharov and Ko, 2009).

IV. CONCLUSION

There are several subclasses of each of the four groups of motif discovery algorithms, enumerative, probabilistic, combinatorial, and nature-inspired. The enumerative strategy is the only one that guarantees to locate all motifs (aside from weak motifs) and is an exhaustive search with a straightforward notion. But because it is so sluggish and needs so many parameters, dealing with extended motifs or large amounts of data becomes challenging. Additionally, the degenerative places are constrained by the motif representation restrictions.

Many of the enumerative approach's weaknesses, such as speed, handling lengthy motifs and large amounts of data, the quantity of necessary parameters, degraded positions, and the inability to detect weak motifs, are overcome by the probability approach. But because probability is a complicated idea, not all motifs can be found using this method.The key components of the first two categories are combined in the third category, which is called a nature-inspired approach.This strategy uses a straightforward idea and a global search, yet it can also handle large amounts of data and lengthy motifs. Its flexible motif encoding allows for an infinite number of degraded places.The last group is the combinatorial method, which depends on the hybrid algorithms that come together to generate the necessary algorithm for its effectiveness.

A useful tool for motif discovery can be created from the many suggested solutions to the challenge. These elements need to be present in the tool: All models, including OOPS, ZOOPS, TCM, should be identified. It ought to be capable of doing global searches, optimise scoring, It should have optimised data structures, parallel processing capability, and It must be able to distinguish between long and short motifs. It must be able to locate numerous motifs simultaneously, without having to discard the first one in order to find the next, and it must be able to find multiple motifs of different lengths. It must feature an automatic mechanism that reduces the amount of user-determined necessary parameters.

The primary objective of the motif finding challenge is to find unique, highly prevalent, unidentified signals in a collection of sequences. The most popular algorithms for discovering motifs attempt to create profiles that maximise the information content score by generating a generative probabilistic representation of these overrepresented signals. Although these profiles provide a very effective way to represent the signals, the optimal motif corresponds to the global maximum of a non-convex continuous function, which poses a significant challenge.The employment of EM in conjunction with various random starts or any of the other potent stochastic global techniques that could produce fruitful initial hypotheses (such as projection algorithms) aims to enhance the quality of the outcomes. Global techniques instead suggest that a viable solution is in the nearby region rather than making first predictions in the convergence zone of the most advantageous local maximum.

REFERENCES

[1]. Bailey, T. L. (2011). DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics*, *27*(12), 1653-1659.https://doi.org/10.1093/bioinformatics/btr261

[2]. Hashim, F. A., Mabrouk, M. S., & Al-Atabany, W. (2019). Review of Different Sequence Motif Finding Algorithms. *Avicenna journal of medical biotechnology*, *11*(2), 130–148.

[3]. Li, X., & Li, P. (2019). Random projections with asymmetric quantization. *Advances in Neural Information Processing Systems*, *32*.

[4]. Mohanty, S., & Mohanty, S. (2020). Genetic algorithm-based motif search problem: a review. In *Smart Intelligent Computing and Applications: Proceedings of the Third International Conference on Smart Computing and Informatics, Volume 1* (pp. 719-731). Springer Singapore.https://doi.org/10.1007/978-981-13-9282-5_69

[5]. Ruzicka, M., Kulhanek, P., Radova, L., Cechova, A., Spackova, N., Fajkusova, L., & Reblova, K. (2017). DNA mutation motifs in the genes associated with inherited diseases. *PLoS One*, *12*(8), e0182377.

[6]. Sharov, A. A., & Ko, M. S. (2009). Exhaustive search for over-represented DNA sequence motifs with CisFinder. *DNA research*, *16*(5), 261-273.https://doi.org/10.1093/dnares/dsp014

[7]. Sun, H. Q., Low, M. Y. H., Hsu, W. J., & Rajapakse, J. C. (2010, November). ListMotif: A time and memory efficient algorithm for weak motif discovery. In *2010 IEEE international conference on intelligent systems and knowledge engineering* (pp. 254-260). IEEE.**DOI:** 10.1109/ISKE.2010.5680875

[8]. Sun, C., Huo, H., Yu, Q., Guo, H., & Sun, Z. (2015). An affinity propagation-based DNA motif discovery algorithm. *BioMed research international*, *2015*.https://doi.org/10.1155/2015/853461

[9]. Zhang, Y., Wang, P., & Yan, M. (2016). An entropy-based position projection algorithm for motif discovery. *BioMed research international*, *2016*.https://doi.org/10.1155/2016/9127474