IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

A Survey on Regulatory Motifs in DNA Sequence

Rasika Malgi
Dept. AI&DS
SIES, Nerul
Navi Mumbai, India

Kartik Batchu
Dept. AI&DS
SIES, Nerul
Navi Mumbai, India

Siddhi Bachala
Dept. AI&DS
SIES, Nerul
Navi Mumbai, India

Nidhi Nandikol
Dept. AI&DS
SIES, Nerul
Navi Mumbai

Abstract- To untangle the mechanisms that regulate gene expressions is still a vital challenge in biology that biologists face today. A significant task in this challenge is to recognize regulatory motifs or elements. The Regulatory motifs in DNA sequence discovery are a primary step in many systems for studying gene function. Regulatory Motifs are interpreted as short (6-8 bases long), repetitive patterns that have well-defined biological functions and features.

As per early discussion, genetic activity is monitored in response to environmental changes. Regulatory Motifs are in charge of recruiting Transcription Factors, or regulatory proteins, to the required gene. Motifs are also recognized by micro RNAs, which bind the motifs given through nucleotide, that recognize motifs based on their GC content; and other RNAs, which use a combination of DNA sequence and structure. Once bound, they can activate or subdue the expression of the associated gene.

With the current mathematical abilities, regulatory motif discovery and analysis have proceeded in the field of genomic studies. This paper presents a generalization of motif discovery algorithms with new sub-categories that facilitates creating a successful motif discovery algorithm. It also presents a summary of the comparison between them.

Keywords: Motifs, Nucleotide Transcription factors, RNA, Splicing signals

I.INTRODUCTION

There are 2 types of DNA in the DNA sequence, there are genotypes and phenotypes. [5] Genotypes are the genetic materials in an organism. The phenotypes are the characteristic traits we see in an organism, that is the outcome of a genetic pattern or code. Genome is also referred to as the sequence of the genes that is present in the DNA. If we extract a particular DNA strand and examine it, we could find that there are nucleotides in them known as thymine, cytosine, etc. When we observe the sequence, we can see that, over some distance, a pattern is repeated in the sequence and this repeated sequence pattern is called a regulatory motif. These regulatory motifs control the gene expressions, through which we can synthesize the DNA information and use it further for creating the protein and RNA. The gene is the basic unit of genetic information in deoxyribonucleic acid (DNA), defined as part of a base sequence that serves as a template for a copying process called transcription. The main idea of gene expression is that each gene contains the information needed to produce a protein. Gene expression begins with the binding of various protein factors, called transcription factors, to enhancer and promoter sequences. Transcription factors regulate gene expression by activating or repressing transcriptional machinery.

Finding patterns in DNA sequences is one of the most difficult problems in molecular biology and computer science. In its simplest form, the problem can be formulated as follows: Given a set of sequences, find unknown patterns that occur frequently. If an m-letter pattern occurs in exactly every row, a simple sum over all m-letter patterns that occur in those rows gives the solution. However, when it comes to DNA sequences, it is not so simple, as the patterns include mutations, insertions, or deletions of nucleotides.

DNA motifs are defined as nucleic acid sequence patterns that have some biological significance, such as DNA binding sites for regulatory proteins (i.e. transcription factors). Typically, the motif is very short (5–20 base pairs (b-p) long) and is known to occur in different genes or occur multiple times in a single gene [1]. DNA motifs are often related to structural motifs found in proteins.

Patterns can appear on both strands of DNA. Transcription factors bind directly to double-stranded DNA. Sequences can have zero, one, or more copies of a pattern. In inclusion to the regular forms of DNA motifs, two specific classes of DNA motifs are recognized: palindrome motifs and gapped binary motifs (gap). A palindrome pattern is a subsequence identical to its reverse complement, e.g.

For example CACGTG. The spaced crown pattern consists of two smaller conserved sites separated by a spacer (gap). A spacer occurs in the middle of the motif because the transcription factors bind as monomers. This means that transcription factors are composed of two subunits that have two distinct contact points with the DNA sequence.

The DNA-binding part of transcription factors is conserved but usually very small (3–5 b-p). These two contacts are separated by non-retained spacer. Spacers are mostly fixed in length but can vary slightly.

Given a set of DNA sequences (promoter regions), the problem of motif discovery is to find overrepresented and conserved motifs from platonism sequences that are good candidates for transcription factor binding sites. A large number of algorithms have been developed to find DNA patterns.

Most of these algorithms focus to infer motifs by observing the regulatory regions (promoters) of several co-existing genes of the same genome. It is assumed that gene co-expression mainly arises from transcriptional co-regulation. Since co-regulated genes are known to share some similarities in their regulatory mechanisms, possibly at the transcriptional level, their promoter regions may contain some common motifs, which are the transcription factor binding sites. A practical approach to locating these regulatory elements is to explore for statistically disproportionate motifs in the promoter regions of such a group of co-expressed genes. A statistically overrepresented topic means that a topic appears more often than one would expect by chance.

Therefore, these algorithms search for over-expressed motifs in this collection of promoter sequences. However, most of these pattern recognition algorithms appear to perform successfully in yeast and other lower organisms but their execution is very less in higher organisms.

II.METHODOLOGY

The first stage of the methodology begins with the initialization of DNA sequencing. It prepares DNA sequences for precise pattern discovery through assembly and cleanup steps. In the assembly step, it is recommended to select as many target sequences as possible that may contain patterns, try to keep sequences as short as possible and remove sequences that probably do not contain any patterns. The assembly step is performed by grouping the input sequences based on specific information and then extracting the desired sequences into the appropriate sequence database. The input sequence must then be cleaned to hide or remove the hidden sequences.

An intermediate step is the topic discovery process which begins with a suggested sequence. There are two ways to represent models: a consensus string and a position-specific weight matrix (PWM). There are a total of strands with DNA sequence patterns of the same length; it allows symbols to degenerate into strings using IUPAC code and PWM is a 4xm matrix, where m is the length of the pattern. Each position in the matrix represents probabilities for each nucleotide at each indexed position in the pattern. After the pattern representation, an appropriate objective function is determined and finally, an appropriate search algorithm is applied.

There are two main types of pattern discovery algorithms: i.e. Enumeration Methods and Probabilistic Techniques. [2]The enumeration method looks for consensus sequences; patterns are predicted based on word sum and computer word similarity, so This method is sometimes called the word sum method for solving Panted (l, d) pattern problem(PMP), where the pattern length is (l) and the maximum number of mismatches is(d). In addition, these algorithms require several user-defined parameters, such as pattern length, number of mismatches allowed, and the number of mismatches the pattern must appear in 7 minimum sequence numbers.

The enumeration approach can be accelerated using specialized data structures such as suffix trees or parallel processing methods, these methods are convenient for large-length motifs but other algorithms fail to do so. Parallel processing can help in comparing two separate DNA sequences simultaneously in an efficient manner thereby resulting in an increase in the speed of the process. When these methods are used in the algorithm then even a large-length motif can be divided into two equal small-length motifs using parallel processing. To implement this method in a simpler manner we can make use of the PYTHON language on anyone of its libraries like Pycharm, Pandas, etc.

The second group is probabilistic methods. It builds a probabilistic imitation called the position-specific weight matrix (PSWM) or pattern matrix,specifies the distribution of bases at each position in the TFBS to categorize patterns from non-patterns, and it requires sheer limited Search parameters.[2] This algorithm can be used efficiently for motifs with lengths greater than 10bp (1bp= 3.4A= 340pm).

Recently, new nature-inspire algorithms are expected to solve complex and dynamic problems with equitable time and the best cost.[2] These algorithms simulate the behavior of insects or other animals to clarify complications The expandable algorithm can overcome the drawback of local search and synthesize regional Search and global search.

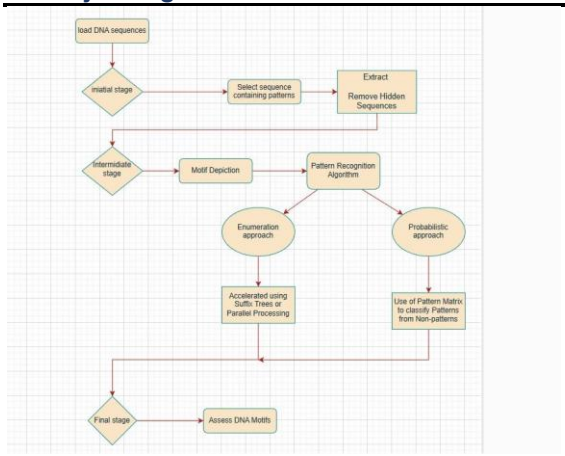

Fig.1 Factors affecting regulation

III.RESULT

In molecular biology and bioinformatics, analyzing regulatory motifs in DNA sequences is a critical task. Regulatory motifs are short DNA sequences that transcription factors recognize and use to regulate gene expression. There are several methods for identifying regulatory motifs in DNA sequences. One common approach is to use computational algorithms to look for sequence patterns that are statistically more common in a set of DNA sequences than would be expected by chance.

MEME, DREME, and HOMER are some popular tools for identifying regulatory motifs. To find motifs, these tools typically employ a variety of algorithms, such as position weight matrices (PWMs), Gibbs sampling, and hidden Markov models (HMMs).[3]Experimental validation is typically required to prove the function of a putative regulatory motif once it has been found. This may entail techniques like reporter gene assays or electrophoretic mobility shift assays (EMSAs).Overall, identifying and characterizing regulatory motifs in DNA sequences is a crucial step in understanding how genes are regulated and may have repercussions in a number of industries, including biotechnology, agriculture, and medicine .Experimental validation is typically required to prove the function of a putative regulatory motif once it has been found. This may entail techniques like reporter gene assays or electrophoretic mobility shift assays (EMSAs).

IV.CONCLUSION

Certain DNA regions called regulatory motifs are essential for controlling how genes are expressed. The promoter regions, enhancers, silencers, and other regulatory components of the genome can all include these patterns.

To find regulatory patterns in DNA sequences, several computational and experimental techniques have been developed. These techniques include chromatin immune precipitation (Ch IP) experiments, which locate DNA sequences bound by certain transcription factors, and motif discovery algorithms, which search DNA sequences for overrepresented patterns. Gene expression depends on regulatory motifs, and modifications to these motifs can modify how genes are expressed. Hence, it is critical for many fields of biology, such as developmental biology, genetics, and medicine, to understand the processes of gene regulation mediated by regulatory motifs. To summarize, regulatory motifs are critical components of gene expression regulation, and identifying and characterizing them is critical for understanding the complex regulatory networks that control gene expression.

V. REFERENCES

[1] “Book: Computational Biology - Genomes, Networks, and Evolution”

[2] Mohan K das & Ho-Kwok Dai, "A survey of DNA motif finding algorithms" 01 November 2007

[3] Gert Thijs & Kathleen Marchal & Magali Lescot & Stephane Rombauts & Bart De Moor & Pierre Rouze & Yves Moreau ”A Gibbs Sampling Method to Detect Overrepresented Motifs in the Upstream Regions of Coexpressed Genes”

[4] Fatma A. Hashim & Mai S. Mabrouk & Walid Al-Atabany “Review of Different Sequence Motif Finding Algorithms”

[5]”Genotype-Phenotype Distinction”